U.S. ARMY RESEARCH OFFICE

Report No. 92-1

March 1992

TRANSACTIONS OF THE NINTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Sponsored by the Army Mathematics Steering Committee

HOST

University of Minnesota
Minneapolis, Minnesota
18-21 June 1991

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 Ninth Army Conference on Applied Mathematics and Computing was held at the Army High Performance Computing Research Center (AHPCRC), at the University of Minnesota, on 18 - 21 June 1991. The Sponsor of these annual meetings is the Army Mathematics Steering Committee (AMSC). Its members would like to thank Professor George Sell, Director of APHCRC, for serving as chairperson on local arrangements. He along with his staff personnel are to be commended for coordinating the many details needed to conduct this successful scientific meeting.

The participants of the conference were treated to an Open House on 17 June, with tours of the facilities and a demonstration of the visualization and graphic facilities. The conference was very well attended with more than one hundred participants, including about forty scientists from the army laboratories. The technical program consisted of five special sessions scheduled on topics such as Smart Materials, Design of Real-time Control, Probabilistic Algorithms, and Large-scale Optimization. The conference featured more that forty contributed papers presented in nine technical sessions. In addition there were seven invited speakers, whose names are listed below together with the titles of their talks.

SPEAKER AND AFFILIATION	TITLE OF ADDRESS
Professor Roger Brockett **Harvard University** **Cambridge, Massachetts**	**Continuous Computations**
Professor Rudolf E. Kalman **University of Florida** **Gainesville, Florida**	**Identification of Systems from Noisy Data - A New Look at Statistics from the Real World**
Dr. Oliver Pironneau **Institut National de Recherche** **Le Chesmay, France**	**Implementation of the k-Epsilon Turbulence Model in Finite Element Compressible Navier-Stokes Solvers**
Professor G. Kallianpur **University of North Carolina** **Chapel Hill, North Carolina**	**Stochastic Analysis**

Dr. Linda R. Petzold **Lawrence Livermore National Laboratory** **Livermore, California**	**On the Numerical Solution of Constrained Dynamical Systems**
Professor A. Arvind **Massachusetts Institute of Technology,** **Cambridge, Massachussetts**	**Implicit Parallel Programming and Dataflow Architecture**
Dr. Gunter Stein **Honeywell Corporation** **Minneapolis, Minnesota**	**Robust Control**

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

TABLE OF CONTENTS

Title	Page
Foreword	iii
Table of Contents	v
Program	xi
A Note on the Aspect Angle formed between the Convex Hull and its Interior Points, in the Context of the Euclidean Traveling Salesman Problem T.M. Cronin	1
Manifold Method of Material Analysis Gen-hua Shi	51
Analytical Solution of Elastic-Plastic Thick-Walled Cylinders with General Hardening Peter C.T. Chen	77
Analysis of Shear Banding in ARMCO if Iron, Tungsten Alloy and Depleted Uranium R.C. Batra and C.H. Kim	91
Analysis and Computation of Solutions to an Evolution Problem in Nonlinear Viscoelasticity Donald A. French	107
Numerical Modelling of Mode I Linear Viscoelastic Fracture M.K. Warby, J.R. Walton and J.R. Whiteman	115
Nonlinear Static and Dynamic Analyses of a Generic Enclosure Subjected to an Internal Pressure Aaron Das Gupta	125

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

Title	Page
Calculation of Elastic-Plastic Wave Propagation on the Connection Machine Mark A. Olson and Kent D. Kimsey	139
Finite Element Solution of Transient In-Bore Response Problems Kenneth A. Bannister, Stephen A. Wilkerson and Donald A. Rabern	151
Computing the PSVD of Two-by-Two Triangular Matrices Gary E. Adams, Adam W. Bojanczyk and Franklin T. Luk	165
An Asynchronous Array Design for MVDR Beamformers Moon S. Jun and Shietung Peng	183
General Algorithm Based Error Correction and Orthogonal Polynomials Daniel Boley	199
Accurate Frequency Analysis of Measured Time-Dependent Signals Over Short Time Intervals Reo Olson and Daniel H. Cress	213
The Arithmetic Fourier Transform (AFT): Iterative Computation and Image Processing Applications Donald W. Tufts and Haiguang Chen	227
Combinatorial Aspects of the Hilbert Scheme Alyson A. Reeves	251
Using Groebner Bases to Determine the Nature of Field Extensions Moss E. Sweedler	255
Analytic Solution of the Period Four Quadratic Recursion Polynomial Harry J. Auvermann	259
Beyond Rolle's Theorem Bruce Anderson	271

Title	Page
Iterative Methods and Finite Difference Schemes for Incompressible Flow John C. Strikwerda and Dongho Shin	279
Numerical Simulation of Sabot Discard Aerodynamics Using Computational Fluid Dynamics Michael J. Nusca	297
Various Finite Difference Schemes for Transient Three Dimensional Heat Conduction Rao Yalamanchili and Surya R. Yalamanchili	309
High Performance Simplification-Based Automated Deduction Maria Paola Bonacina and Jieh Hsiang	321
Constructive Relational Programming: A Declarative Approach to Program Correctness and High Level Optimization Paul Broome and James Lipton	337
Real-Time Reasoning in Deadline Situations Madhura Nirkhe, Sarit Kraus and Donald Perlis	347
An Overview of the Modular UNIX -Based Vulnerability Estimation Suite Jill H. Smith, Wendy A. Winner and Phillip J. Hanes	355
A logical Framework for Operations on Distributed Data P. Broome and B.D. Broome	363
An Object-Oriented Approach to Large-Scale Battlefield Simulation Michael Brewer and Pat Burns	379
Evolving Phase Boundaries in Deformable Continua Morton E. Gurtin	391
A Central Limit Theorem for Extreme Sojourns of Diffusion Processes Simeon M. Berman	395

Title	Page
3-D Shape from a Shaded and Textural Surface Image Yoonsik Choe and R.L. Kashyap	399
Recurrence Relations, Continued Fractions and Time Evolution in Many-Particle Systems M. Howard Lee	403
Image Singularities of Green's Functions for Anisotropic Elastic Half-Spaces and Bimaterials T.C.T. Ting	415
The Computation of Crystalline Microstructure Mitchell Luskin and Charles Collins	419
On Dynamical Aspects of a Phase Transition Problem Hiroaki Fujimoto and Harumi Hattori	427
Energy Dissipation in an Elastic Material Containing a Mobile Phase Boundary Subjected to Concurrent Dynamic Pulses Jiehliang Lin and Thomas J. Pence	437
A Unified Representation for Some Combinatorial Optimization Problems Wing Shing Wong	451
Classification of Finite Dimensional Filters from Lie Algebraic Point of View Stephen S.-T. Yau	459
An Accurate Algorithm for Minimal Partial Realizations Adam W. Bojanczyk, Tong J. Lee, and Franklin T. Luk	467
The Hyperbolic Transformations in Signal Processing and Control Adam Bojanczyk and Allan O. Steinhardt	479
Iterative Algorithms for Integral Equations of the First Kind Mark G. Vangel	489

Title	Page
On the Analysis of Superharmonic Oscillations J.J. Wu	507
Constitutive Coefficients for Viscohyperelastic Materials A.R. Johnson and C.J. Quigley	517
High-T Superconductivity and the Photoelectric Effect Richard A. Weiss	529
Quantum Theory of Time and Thermodynamics Richard A. Weiss	565
Ultrafast Coherent Heat Engines Richard A. Weiss	623
Thermodynamics and Gravity Richard A. Weiss	671
Robust Stabilization, Robust Performance, and Disturbance Attenuation for Uncertain Linear systems Yeih J. Wang, Leang S. Shieh and John W. Sunkel	699
Minimax Linear Splines Royce W. Soanes	721

NINTH ARMY CONFERENCE ON APPLIED MATHEMATICS AND COMPUTING

Host

Army High Performance Computing Research Center
Minnesota Technology Center Building
University of Minnesota
Minneapolis, Minnesota

18-21 June 1991

AGENDA

Tuesday, 18 June 1991

0745 - 1600 **Registration - Seminar Room**

0815 - 0830 **Opening Remarks - Seminar Room**

0830 - 0930 **General Session I - Seminar Room**

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

CONTINUOUS COMPUTATIONS
Roger Brockett, Harvard University, Cambridge, Massachusetts

0930 - 1000 **Break**

1000 - 1200 **Special Session 1 - Large-Scale Optimization - Seminar Room**

Chairperson: Kenneth D. Clark, U.S. Army Research Office, Research Triangle Park, North Carolina

TENSOR METHODS FOR LARGE SPARSE SYSTEMS OF NONLINEAR EQUATIONS
Robert B. Schnabel, University of Colorado at Boulder, Colorado

A MATRIX ANALYSIS OF CONJUGATE GRADIENT ALGORITHMS
Steven F. Ashby, Lawrence Livermore National Laboratory, Livermore, California and Martin H. Gutknecht, Eidgenössische Technische Hochschule, Zürich, Switzerland

A NOTE ON THE ASPECT ANGLE FORMED BETWEEN THE CONVEX HULL AND ITS INTERIOR POINTS, IN THE CONTEXT OF THE EUCLIDEAN TRAVELING SALESMAN PROBLEM
Terence M. Cronin, CECOM Center for Signals Warfare, Warrenton, Virginia

A NEW ALGORITHM FOR LARGE OPTIMIZATION PROBLEMS WITH NONLINEAR CONSTRAINTS
Jorge Nocedal, Northwestern University, Evanston, Illinois

******

1000 - 1200 **Technical Session 1 - Analytical and Numerical Methods for Material Models, I - Commons Room**

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

MOLECULAR MODELING OF ENERGETIC MATERIALS
George F. Adams, Betsy M. Rice, Cary F. Chabalowski, and Pamela J. Kaste, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

MANIFOLD METHOD OF MATERIAL ANALYSIS
Gen-hua Shi, U.S. Army Engineer Waterways Experiment Station, Vicksburg, Mississippi

ANALYTICAL SOLUTION OF ELASTIC-PLASTIC THICK-WALLED CYLINDERS WITH GENERAL HARDENING
Peter C.T. Chen, Benet Laboratories, Watervliet, New York

ANALYSIS OF SHEAR BANDING IN TWELVE MATERIALS
Romesh C. Batra and C. H. Kim, University of Missouri-Rolla, Rolla, Missouri

ANALYSIS AND COMPUTATION OF SOLUTIONS TO AN EVOLUTION PROBLEM IN NONLINEAR VISCOELASTICITY
Donald A. French, University of Cincinnati, Cincinnati, Ohio

APPLICATION OF THE MTS FLOW LAW TO THE SIMULATION OF ADIABATIC SHEAR BANDS
John W. Walter, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

1200 - 1330 **Lunch**

1330 - 1530 **Technical Session 2 - Analytical and Numerical Methods for Material Models, II - Seminar Room**

Chairperson: Peter C.T. Chen, Benet Laboratories, Watervliet, New York

CONSTITUTIVE COEFFICIENTS FOR VISCOHYPERELASTIC MATERIALS
Arthur Johnson, C. J. Quigley, D. L. Cox, L. C. Bissonnette, and W. C. Maciejewski, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

FINITE ELEMENT MODELLING OF CRACK GROWTH IN A FINITE BODY IN THE CONTEXT OF MODE I LINEAR VISCOELASTIC FRACTURE
J. R. Whiteman and M. K. Warby, Brunel University, Uxbridge, England and J. R. Walton, Texas A&M University, College Station, Texas

NONLINEAR STATIC AND DYNAMIC ANALYSES OF A GENERIC ENCLOSURE SUBJECTED TO AN INTERNAL PRESSURE
Aaron Das Gupta, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

CALCULATION OF ELASTIC-PLASTIC WAVE PROPAGATION ON THE CONNECTION MACHINE
Mark A. Olson and Kent D. Kimsey, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

FINITE ELEMENT SOLUTION OF TRANSIENT IN-BORE RESPONSE PROBLEMS
Kenneth A. Bannister and Stephen A. Wilkerson, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

NONLINEAR STRESS ANALYSIS OF IN-BORE PROJECTILES
Shih C. Chu, U.S. Army Armament R&D Center, Dover, New Jersey

******

1330 - 1530 **Technical Session 3 - Signal Processing: Algorithms, Architectures, and Applications - Seminar Room**

Chairperson: John Strikwerda, University of Wisconsin-Madison, Madison, Wisconsin

LINEAR PREDICTION AND SVD OF A MATRIX PRODUCT
Adam W. Bojanczyk and Franklin T. Luk, Cornell University, Ithaca, New York

AN ASYNCHRONOUS ARRAY DESIGN FOR MVDR BEAMFORMERS
Moon S. Jun, New Mexico State University, Las Cruces, New Mexico and Shietung Peng, University of Maryland, Baltimore County, Catonsville, Maryland

GENERAL ERROR CORRECTION PROBLEM AND ORTHOGONAL POLYNOMIALS
Daniel Boley, University of Minnesota, Minneapolis, Minnesota

ACCURATE FREQUENCY ANALYSIS OF REAL-TIME SIGNALS OVER SHORT TIME INTERVALS
Robert E. Olson and Daniel H. Cress, U.S. Army Waterways Experiment Station, Vicksburg, Mississippi

THE MATHEMATICS OF THE ARITHMETIC FOURIER TRANSFORM AND APPLICATIONS TO IMAGE PROCESSING
Donald W. Tufts, University of Rhode Island, Kingston, Rhode Island

SIGNATURE PREDICTION METHODS AND COMPUTER RESOURCES FOR ITEM LEVEL ANALYSES
Paul Stay, Ed Davisson, and Susan Coates, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

Tuesday (Continued)

1530 - 1600 **Break**

1600 - 1700 **General Session II - Seminar Room**

Chairperson: Terence M. Cronin, CECOM Center for Signal Warfare, Warrenton, Virginia

IDENTIFICATION OF SYSTEMS FROM NOISY DATA--A NEW LOOK AT STATISTICS FROM THE REAL WORLD
Rudolf E. Kalman, University of Florida, Gainesville, Florida

------

Wednesday, 19 June 1991

0800 - 1600 **Registration - Seminar Room**

0830 - 1030 **Special Session 2 - Probabilistic Algorithms - Seminar Room**

Chairperson: J. Michael Steele, University of Pennsylvania, Philadelphia, Pennsylvania

SURVEY OF PROBABILITY APPLICATIONS IN THE THEORY OF ALGORITHMS
J. Michael Steele, University of Pennsylvania, Philadelphia, Pennsylvania

ON THE EXACT VALUES AND CENTRAL LIMIT THEOREMS OF CLASSICAL PROBLEMS IN COMBINATORIAL OPTIMIZATION AND COMPUTATIONAL GEOMETRY
Florin Avraam, Northeastern University and Dimitris Bertsimas, Massachusetts Institute of Technology, Cambridge, Massachusetts

TRAVELING SALESMAN PROBLEMS WITH A SELF-SIMILAR ITINERARY
Steven P. Lalley, Purdue University, West Lafayette, Indiana

******

0830 - 1030 **Technical Session 4 - Algebra and Symbolic Computation - *Commons Room***

Chairperson: Ram P. Srivastav, State University of New York at Stony Brook, New York

THE NUMBER OF GROUPS OF ORDER N
Keith Dennis, Cornell University, Ithaca, New York

CELL DECOMPOSITION FOR THE P-ADICS AND COMPUTATIONS WITH P-ADIC ALGEBRAIC NUMBERS
Devdatt P. Dubhashi, Cornell University, Ithaca, New York

COMBINATORIAL ASPECTS OF THE HILBERT SCHEME
Alyson Reeves, Cornell University, Ithaca, New York

GROEBNER BASES AND FIELD EXTENSIONS
Moss Sweedler, Cornell University, Ithaca, New York

ANALYTIC SOLUTION OF THE PERIOD FOUR QUADRATIC RECURSION POLYNOMIAL
Harry J. Auvermann, U.S. Army Atmospheric Sciences Laboratory, White Sands Missile Range, New Mexico

HIGHER ORDER ROLLE'S THEOREMS
Bruce Anderson, Cornell University, Ithaca, New York

1030 - 1100 **Break**

1100 - 1200 **General Session III - Seminar Room**

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

IMPLEMENTATION OF THE k-EPSILON TURBULENCE MODEL IN FINITE ELEMENT COMPRESSIBLE NAVIER-STOKES SOLVERS
Olivier Pironneau, Institut National De Recherche, Le Chesnay, France

1200 - 1330 **Lunch**

1330 - 1530 **Technical Session 5 - Large Scale Computation and Fluid Dynamics - Seminar Room**

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

ITERATIVE METHODS AND FINITE DIFFERENCE SCHEMES FOR INCOMPRESSIBLE FLOW
John C. Strikwerda, University of Wisconsin-Madison, Madison, Wisconsin

NUMERICAL SIMULATION OF SABOT DISCARD AERODYNAMICS USING COMPUTATIONAL FLUID DYNAMICS
Michael J. Nusca, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

A COMPUTATIONAL STUDY OF CYLINDRICAL SEGMENTS IN THE WAKE OF A PROJECTILE
Jubaraj Sahu and Charles J. Nietubicz, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

SCHEDULING NUMERICAL SIMULATION COMPUTATIONS ON MIMD MACHINES
N. P. Chrisochoides, E. N. Houstis, and J. R. Rice, Purdue University, West Lafayette, Indiana

VARIOUS FINITE DIFFERENCE SCHEMES FOR TRANSIENT THREE DIMENSIONAL HEAT CONDUCTION
Rao Yalamanchili, U.S. Army Armament R&D Center, Dover, New Jersey

BALLISTIC SIMULATIONS OF TANK CHARGES WITH THE PRESENCE OF PROPELLANT DUST AND ULLAGE
Lang-Mann Chang, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

******

1330 - 1530 **Technical Session 6 - Foundations of Software Systems and Battle Management - Commons Room**

Chairperson: Kenneth A. Bannister, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

HIGH PERFORMANCE SIMPLIFICATION BASED AUTOMATED DEDUCTION
Maria P. Bonacina and Jieh Hsiang, State University of New York at Stony Brook, New York

CONSTRUCTIVE RELATIONAL PROGRAMMING: A DECLARATIVE APPROACH TO PROGRAM CORRECTNESS AND HIGH LEVEL OPTIMIZATION
Paul Broome, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground and James Lipton, University of Pennsylvania, Philadelphia, Pennsylvania

REAL-TIME REASONING IN DEADLINE SITUATIONS
Donald Perlis, University of Maryland, College Park, Maryland

AN OVERVIEW OF THE MODULAR UNIX-BASED VULNERABILITY ESTIMATION SUITE
Wendy A. Winner, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

A LOGICAL FRAMEWORK FOR OPERATIONS ON DISTRIBUTED DATA
Paul Broome and B. D. Broome, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

AN OBJECT-ORIENTED APPROACH TO LARGE-SCALE BATTLEFIELD SIMULATION
Michael Brewer and Patrick J. Burns, Colorado State University, Fort Collins, Colorado

1530 - 1600 **Break**

Wednesday (Continued)

1600 - 1700 **General Session IV - Seminar Room**

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

STOCHASTIC ANALYSIS
G. Kallianpur, University of North Carolina at Chapel Hill, North Carolina

------

Thursday, 20 June 1991

0800 - 1600 **Registration - Seminar Room**

0830 - 1030 **Special Session 3A - Mathematics of Smart Materials - Seminar Room**

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

CONCEPTS OF SURFACE FORCES FOR EVOLVING PHASE INTERFACES
Morton E. Gurtin, Carnegie Mellon University, Pittsburgh,

MICROSTRUCTURE AND MACROSCOPIC PROPERTIES OF MAGNETOSTRICTIVE MATERIALS
Richard D. James, University of Minnesota, Minneapolis, Minnesota

SURFACE ENERGY AND MICROSTRUCTURE IN COHERENT PHASE TRANSITIONS
Robert V. Kohn, Courant Institute of Mathematical Sciences, New New York, New York

DIELECTRIC PROPERTIES OF PIEZOELECTRIC COMPOSITES
Marco Avellaneda and Tamara Olson, Courant Institute of Mathematical Sciences, New York, New York

******

Thursday (Continued)

0830 - 1030 **Technical Session 7 - Stochastic Methods and Applications - Commons Room**

Chairperson: Mark Vangel, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

A CENTRAL LIMIT THEOREM FOR EXTREME SOJOURNS OF DIFFUSION PROCESSES
Simeon M. Berman, Courant Institute of Mathematical Sciences, New York, New York

QUANTITATIVE THEORIES FOR INTERFACIAL CHAOTIC MIXING
James Glimm and Qiang Zhang, State University of New York at Stony Brook, New York

NUMERICAL TREATMENT OF ITO-TYPE STOCHASTIC DIFFERENTIAL SYSTEMS
G. S. Ladde, University of Texas at Arlington, Texas

3-D SHAPE FROM A SHADED AND TEXTURAL SURFACE IMAGE
Yoonsik Choe and R. L. Kashyap, Purdue University, West Lafayette, Indiana

RECURRENCE RELATIONS, CONTINUED FRACTIONS AND TIME EVOLUTION IN MANY-PARTICLE SYSTEMS
M. Howard Lee, University of Georgia, Athens, Georgia

STABILITY OF LOTKA-VOLTERRA MODEL
G. S. Ladde, University of Texas at Arlington, Texas, and
S. Sathananthan, Jarvis Christian College, Hawkins, Texas

1030 - 1100 **Break**

1100 - 1200 **General Session V - Seminar Room**

Chairperson: Paul Broome, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, Maryland

ON THE NUMERICAL SOLUTION OF CONSTRAINED DYNAMICAL SYSTEMS
Linda R. Petzold, Lawrence Livermore National Laboratory, Livermore, California

1200 - 1330 **Lunch**

1330 - 1530 **Special Session 3B - Mathematics of Smart Materials - Seminar Room**

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

IMAGE SINGULARITIES OF GREEN'S FUNCTIONS FOR ANISOTROPIC ELASTIC HALF-SPACES AND BIMATERIALS
T. C. T. Ting, University of Illinois, Chicago, Illinois

THE COMPUTATION OF CRYSTALLINE MICROSTRUCTURE
Mitchell Luskin, University of Minnesota, Minneapolis, Minnesota

DYNAMICS OF A PHASE TRANSITION PROBLEM
Harumi Hattori, West Virginia University, Morgantown, West Virginia

ASYMPTOTIC ENERGY DISSIPATION DUE TO ACOUSTIC REVERBERATIONS IN AN ELASTIC MATERIAL CONTAINING A MOBILE PHASE BOUNDARY
Thomas J. Pence, Michigan State University, East Lansing, Michigan

******

1330 - 1530 **Technical Session 8 - Symbolic Methods and Discrete Mathematics - Commons Room**

Chairperson: Royce Soanes, Benet Laboratories, Watervliet, New York

SYMBOLIC UNCOUPLING OF MULTIBODY EQUATIONS OF MOTION, PART I-THEORY, PART II-IMPLEMENTATION
Roger A. Wehage and Michael J. Belczynski, U.S. Army Tank-Automotive Command, Warren, Michigan

SYMBOLIC ALGEBRA METHODS USED IN THE CONTROL OF A STEWART PLATFORM
James L. Overholt, U.S. Tank-Automotive Command, Warren, Michigan, and Ashraf Zeid, Computer Sciences Corporation, Warren, Michigan

REPRESENTATION AND MODELING OF MULTIBODY KINEMATICS AND DYNAMICS USING BOND GRAPHS
Ashraf Zeid, Computer Sciences Corporation, Warren, Michigan, and Roger Wehage, U.S. Army Tank-Automotive Command, Warren, Michigan

CHARACTERIZATION OF GRAY CODES IN CUBE-BASED NETWORKS WITH APPLICATIONS
Shahram Latifi, University of Nevada, Las Vegas, Nevada

A UNIFIED GRADIENT APPROACH TO SOLVE SOME NP-HARD PROBLEMS
Wing S. Wong, AT&T Bell Laboratories, Holmdel, New Jersey

1530 - 1600 **Break**

1600 - 1700 **General Session VI - Seminar Room**

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

IMPLICIT PARALLEL PROGRAMMING AND DATAFLOW ARCHITECTURE
A. Arvind, Massachusetts Institute of Technology, Cambridge, Massachusetts

Friday, 21 June 1991

0800 - 1200 **Registration - Seminar Room**

0830 - 1030 **Special Session 4 - Computational Issues in Real-Time Control - Seminar Room**

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

FINITE DIMENSIONAL FILTERS WITH NONLINEAR DRIFT
Stephen Yau, University of Chicago, Chicago, Illinois

AN ACCURATE ALGORITHM FOR MINIMAL PARTIAL REALIZATIONS
Adam W. Bojanczyk, Tong J. Lee, and Franklin T. Luk, Cornell University, Ithaca, New York

HYPERBOLIC FACTORIZATIONS IN CONTROL AND SIGNAL PROCESSING
Adam W. Bojanczyk and Allan Steinhardt, Cornell University, Ithaca, New York

ROBUST STABILIZATION, ROBUST PERFORMANCE, AND DISTURBANCE ATTENUATION FOR UNCERTAIN LINEAR SYSTEMS
Yeih J. Wang and Leang S. Shieh, University of Houston, Houston, Texas, and John W. Sunkel, NASA-Johnson Space Center, Houston, Texas

******

0830 - 1030 **Technical Session 9 - Methods in Applied Mathematics - Commons Room**

Chairperson: James L. Overholt, U.S. Army Tank-Automotive Command, Warren, Michigan

ON COMPUTING THE SINGULAR BEHAVIOR OF SOLUTIONS OF THE CAUCHY SINGULAR INTEGRAL EQUATION
Ram P. Srivastav, State University of New York at Stony Brook, New York

SOLVING INTEGRAL EQUATIONS OF THE FIRST KIND BY ITERATION WITH IMPLICIT REGULARIZATION
Mark G. Vangel, U.S. Army Materials Technology Laboratory, Watertown, Massachusetts

ON THE ANALYSIS OF SUBHARMONIC OSCILLATIONS
Julian J. Wu, U.S. Army Research Office, Research Triangle Park, North Carolina

ASYMPTOTICALLY UNIFORM PIECEWISE LINEAR INTERPOLATION
Royce Soanes, Benet Laboratories, Watervliet, New York

ON CONSTRUCTING RATIONAL APPROXIMATIONS TO REAL-VALUED FUNCTIONS OF A REAL VARIABLE
Ram P. Srivastav, State University of New York at Stony Brook, New York

PARALLEL ALGORITHMS, CHAOTIC MIXING, AND SHOCK WAVE DIFFRACTION
Yuefan Deng, James Glimm, John Grove, and Yi Wang, State University of New York at Stony Brook, New York

Paper 1: **HIGH-T_C SUPERCONDUCTIVITY AND THE PHOTOELECTRIC EFFECT**

Paper 2: **QUANTUM THEORY OF TIME AND THERMODYNAMICS**
Richard A. Weiss, U.S. Army Engineer Waterways Experiment Station, Vicksburg, Mississippi

1030 - 1100 **Break**

1100 - 1200 **General Session VII - Seminar Room**

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

ROBUST CONTROL
Gunter Stein, Honeywell Corporation, Minneapolis, Minnesota

1200 - 1215 **ADJOURNMENT**

A Note on the Aspect Angle formed between the Convex Hull and its Interior Points, in the Context of the Euclidean Traveling Salesman Problem

T.M. Cronin
CECOM Center for Signals Warfare
Warrenton VA 22186-5200

Abstract. For the Euclidean traveling salesman problem (ETSP), it has long been known that the relative order of the cities comprising the convex hull is preserved within an optimal tour. It is thus natural during ETSP problem solving to utilize the hull as an initial tour. The main result of this paper is an extension of this concept, which proves that all interior cities which form a disjoint, maximally obtuse angle with the convex hull may also be inserted into the baseline tour (a disjoint, maximally obtuse angle is one larger than any other obtuse angle which a city may form with the hull). Furthermore, any cities which form a disjoint, maximally obtuse angle with the resultant structure may *also* be inserted. The only caveat is that each city inserted in this fashion must be periodically retested, to check that the maximally obtuse condition remains valid. The geometric rationale for the technique was developed in an earlier paper, in which it was shown that passing through each hull vertex is a hyperbola, the purpose of which is to discriminate the specific hull segment to be perturbed when inserting a city into the tour. With regard to performance, the entire process just described may be achieved in a preprocessing step with time complexity O[n log n], where n is the number of cities being processed. In the best case, if all interior cities form an obtuse angle with the hull, an instance of the problem is solved in O[n log n] time. In the worst case, when no interior cities form an obtuse angle with the hull, no improvement is obtained. The technique is demonstrated for two databases with proven certificates of optimality: the 127-city University of Augsburg dataset, and the 532-city Bell Laboratories dataset. For these examples, the partial tours produced by the technique bear a marked structural resemblance to a complete optimal tour.

Background: the Euclidean Traveling Salesman Problem.

The Euclidean traveling salesman problem [ETSP] is a long-standing problem in optimization, having roots and primary development in the field of operations research, with ancillary developments in the fields of computational geometry and graph theory. As is the case with many obtuse problems in mathematics, the ETSP may be succinctly stated. Given a set of cities and the distances between each pair, the objective is to find the shortest tour which visits each city exactly once, except the start city, which is revisited at tour's end. A tour is simply a closed loop connecting all the cities; the formal mathematical name for a tour is a Hamiltonian cycle. One of the interesting facts discovered early on is that a tour is not permitted to cross itself [F1]. There are (n-1)! / 2 possible tours through n cities, which is a combinatorially prohibitive number of operations to perform by brute force, so it is therefore desirable to find an algorithm which arrives at a solution in polynomial time. The ETSP is a special case of the general traveling salesman problem, the former bearing the distinction that the metrics involved are Euclidean distances rather than arbitrary costs or weights.

To date, the Euclidean traveling salesman problem remains unsolved. By "unsolved", it is meant that no one has developed a formal proof of optimality for a polynomial-time algorithm guaranteed to produce the shortest tour. In the mid-seventies, it was proven that the ETSP is NP-hard [G1]. This is a somewhat more favorable complexity result than that obtained for the general traveling salesman problem, which belongs to the NP-complete class of problems [G2]. There have been two camps of researchers working on the Euclidean version of the problem, with the earliest computational work dating back to the end of the second world war [L1]. The first camp has striven to produce an exact solution to the problem, and in doing so has pioneered advances in the field of linear programming, including such techniques as the simplex algorithm, branch-and-bound, and branch-and-cut [P1]. An exact approach favors precision at the cost of performance. The second camp of researchers has settled for an approximate approach, by resorting to heuristics which produce high quality solutions per unit of processing time. The principal heuristic techniques are k-opt edge exchange (the most advanced of which is the iterated Lin-Kernighan), simulated annealing, genetic algorithms, elastic bands, and neural nets [J1]. Generally, the approximate techniques

develop a solution with more speed than exact approaches, at the cost of precision. However, even this generality is moot, because some of the heuristic approaches render solutions orders of magnitude faster than others, with only marginally inferior results.

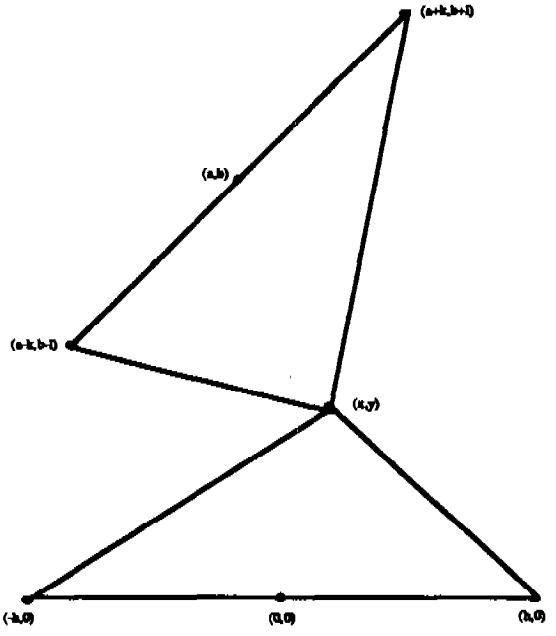
Verifying the Optimality of a Tour.

To test a ETSP algorithm (whether it be exact or approximate) against large databases, it is necessary to have at hand some technique to verify an optimal solution in polynomial time. For city databases of size one hundred or less, it is possible to use a variant of branch-and-bound to check optimality in reasonable computer time [J1]. However, when n becomes much larger than one hundred, certifying optimality begins to consume unreasonable amounts of time. It is for this reason that a technique based on computing a lower bound on optimal tour length has been developed [H1]. This quantity, known as the Held-Karp lower bound, is computable in polynomial time, and empirical results indicate that it is consistently within two percent of optimal [J1]. Scientists in the field of operations research have made good use of the bound. Rather than strive for an optimal tour, researchers instead attempt to come within a reasonable neighborhood of the Held-Karp bound.

The Discovery of the Non-linear Search Space for the ETSP.

Despite over forty years of intense study by computer scientists and operations research analysts, the search space for the Euclidean traveling salesman problem remained unspecified as of 1990 (i.e., it was not known whether the mathematics of tour construction was linear, non-linear, or transcendental in the number of cities). This lack of knowledge prompted the author to conduct experiments during the winter of 1990, in an attempt to characterize the space by leveraging the recently developed field of computational geometry upon the problem. In 1968, researchers at the Johns Hopkins University reported upon a slight modification to a theorem due to Barachet to show that an optimal tour must preserve the order of the convex hull of cities - the shortest tour must contain these cities in the order in which they appear about the perimeter [B1, B2]. This fact suggested that an experiment which inserts an arbitrary city into a hull could serve as a valuable testbed in which to discover the geometric locus of equal hull perturbation. A *perturbation* is a subtour which leads into the interior of the hull through two adjacent hull vertices, to capture cities which do not lie on the hull. In conjunction with a perturbation we introduce the *elliptic distance* between a segment and a point p, which is defined to be the sum of the distances from the endpoints of the segment to p, minus the length of the segment (Figure 1).

When comparing a perturbed hull segment against another perturbed segment, one is actually comparing a confocal system of ellipses against another system, under a continuous spectrum of elliptic distances. The foci of the two systems of ellipses are respectively the two endpoints of the hull segments being perturbed. In Army research at the CECOM Center for Signals Warfare performed during the 1990 fiscal year, it was discovered that the search space induced by the intersection of the two confocal systems of ellipses is in general fourth order (quartic), and in special cases hyperbolic. These facts become apparent when one realizes that reasoning about shortest tours is a process which inherently involves the intersection of a pair of ellipses, the foci of which are defined by pairs of cities. Ellipse intersection is an operation which in the worst case produces a fourth-order equation (quartic). In the special case in which two ellipses share a focus, the locus is a semi-hyperbola. The same non-linear behavior is manifested as more cities are added to the interior, which means that the general search space is quartic regardless of the number of cities added to the tour from within the hull. Dynamic programming immediately suggested itself as an approach to the problem which might provide the framework to keep track of the quartic and hyperbolic boundaries of equal tour perturbation when a new city is added to the existing space. Armed with the new information about the non-linear search space, the author has proceeded to develop a dynamic programming algorithm to maintain incremental optimality when building shortest Euclidean tours. Since the algorithm is designed to probe inwards from the convex hull, it is apparent that efficient hull generating algorithms are required, as well as any algorithms which might exactly extend the hull in a preprocessing step, to produce an optimal baseline tour containing the cities on the hull and any cities which form a shallow angle with it.



$$\sqrt{(x-h)^2+y^2}+\sqrt{(x+h)^2+y^2}-2h=$$
$$\sqrt{(x-(a-k))^2+(y-(b-l))^2}+$$
$$\sqrt{(x-(a+k))^2+(y-(b+l))^2}-2\sqrt{k^2+l^2}$$

Figure 1. The equation for elliptic distance, with associated geometry. The locus is in general a quartic (fourth-order polynomial).

A Brief History of Quartic Curves.

Before proceeding, it is perhaps instructive to pause for a short history of the development of higher order plane curves. Most of the material in this section has been paraphrased by the author from a variety of historical sources. Of particularly broad scope and insight is reference [K3].

The existence of quartics, or fourth-order polynomials, has been known since antiquity. In attempting to trisect the angle or duplicate the cube, the ancient Greeks produced a body of mathematics (and sometimes built devices) which resulted in the development of simple quartic curves. Examples include the Conchoid of Nicomedes and the Kampyle of Eudoxus. Nicomedes and Eudoxus were empirical scientists, overshadowed by the ingenious Archimedes, who devised sophisticated techniques to invent new quartics. The quartic curves discovered during this era were quite simple to visualize, given the limited tools available, and the Greeks obviously exploited symmetry to facilitate progress. That is probably why most of the quartics handed down to us in the twentieth century are even-valued functions. It is important to note that the introduction of odd-powered components serves to skew the quartics with asymmetric artifacts, and also introduces singularities such as cusps and multiple points.

Detailed knowledge about the existence of certain kinds of quartic curves did not provide a method to solve the general quartic equation (see below). It was not until the sixteenth century that a technique was discovered. The method followed immediately after an algorithm was successfully devised by Italian researchers to solve the general cubic equation. Ferrari's technique to solve the quartic (actually a specific quartic known as the biquadratic) equation appeared in Cardano's Ars Magna in 1545. Success in the algebraic solution of the cubic and quartic equations resulted in a centuries-long surge of mathematical research designed to solve by radicals the general equation of the n^{th} degree. However, this activity was destined for failure, because in 1826 Abel proved that an equation of degree five or higher is in general insoluble by algebraic means.

$$ax^4+by^4+cx^3y+dxy^3+ex^2y^2+fx^3+gy^3+hx^2y+ixy^2+jx^2+ky^2+lxy+mx+ny+o=0$$

In the interim, work had proceeded on cataloguing both cubics and quartic curves. Sir Isaac Newton performed a remarkable study of cubics in the latter part of the seventeenth century, and succeeded in enumerating most of the species of third degree curves known to us today. This work was published in 1704, in an appendix to his Opticks. It is reprinted at [W2]. Subsequent work by other researchers contributed a handful of other generic cubic forms to the knowledge store.

Work in algebraic geometry, particularly that of Cayley and Plucker in the nineteenth century, added significantly to knowledge about the quartics. Plucker succeeded in predicting the number of singularities and inflection points in an algebraic curve as a function of the degree of the curve. Salmon summarized and extended this work [S1]. The task of generic classification of the quartics continued into the twentieth century, but with diminished intensity. The last major work containing a detailed taxonomy of quartic curves was published by Hilton [H2]. A good source for the empirical scientist is [L2], although this work is concerned primarily with even-valued quartic functions. Of recent vintage, particularly in the area of singularities of cubics and quartics, is a work by Clemens [C2]. See also [K1] for some interesting results pertaining to self-inversion of cubics and quartics. It is hoped that interest in the higher order plane curves will be rekindled, to address not only the geometry of shortest tours in the plane, but the generic problem of non-linear optimization.

Background: the Convex Hull.

In the plane, the convex hull is the smallest bounding polygon which contains all the points of the problem domain. As indicated above, the relative order of the cities contained on the hull is preserved in an optimal tour. It has been proven that the convex hull is optimally computed in O[n log h] time, where n is the number of points, and h is the number of cities which actually comprise the vertices of the hull [K2]. If one prefers to compute the entire nested hull decomposition, sometimes called the onion, it has been shown [C1] that the structure is optimally computable in O[n log n] time.

The fact that the orientation of the hull is preserved in a an optimal tour suggests that the hull is a good baseline tour from which to add additional cities from the interior. This strategy has been adopted by a number of researchers in the operations research community. An intuitively obvious procedure is to incrementally add to the hull those interior cities which essentially preserve the shape of the hull, in order to least deform the baseline tour.

Shallow Angles as a Heuristic for the Euclidean Traveling Salesman Problem.

In the 1970s it was speculated that a city which forms a maximal angle with a tour leg is a good candidate to be inserted into the tour between the two cities at the endpoints of the tour leg [S2]. This technique attempts to preserve the shape of the existing tour. Various versions of this concept have been implemented, although there has been no work to characterize the admissibility of the technique; i.e., whether or not such an insertion is optimal. One variant is the Golden-Stewart CCAO heuristic algorithm [G3] which is outlined below, with the maximal angle selection step highlighted.

1. Form the convex hull of cities, to be used as the baseline subtour.

2. (Insertion) For each city k not yet contained in the subtour, obtain the two adjacent cities i_k and j_k on the subtour such that $dis(i_k, k) + dis(j_k, k) - dis(i_k, j_k)$ is minimized.

3. (Selection) Select the city k^* that maximizes the angle between edges $\{i_k, k\}$ and $\{k, j_k\}$ in the subtour and insert it between i^{k^*} and $j_k{}^*$.

4. Repeat steps 2 and 3 until a Hamiltonian cycle is obtained.

5. Apply the Or-opt procedure to the tour generated in steps 1-4, and iterate until no improvements are forthcoming.

We will during the remainder of the paper attempt to characterize what it means for a city to form a shallow angle with an existing tour; in particular, in order to get a handle on the problem, we will restrict our study to only those cities that form an *obtuse* angle with the tour.

Geometry of a City which forms an Obtuse Angle with the Current Tour.

One result known in antiquity is that an angle inscribed in a semi-circle is a right angle. For our purposes, suppose that a circle is centered at the midpoint of a tour leg, and that the cities at the endpoints define a diameter of the circle. Now, with the exception of the endpoints of the tour leg, any cities lying on the circle form a right angle with the two cities lying at the endpoints of the diameter. Consider the disk bounded by the circle. It defines an obtuse condition on the tour leg, since any point properly contained within the disk must form an obtuse angle with the two cities at the leg endpoints. Conversely, any cities lying properly outside the disk form an acute angle with the tour leg.

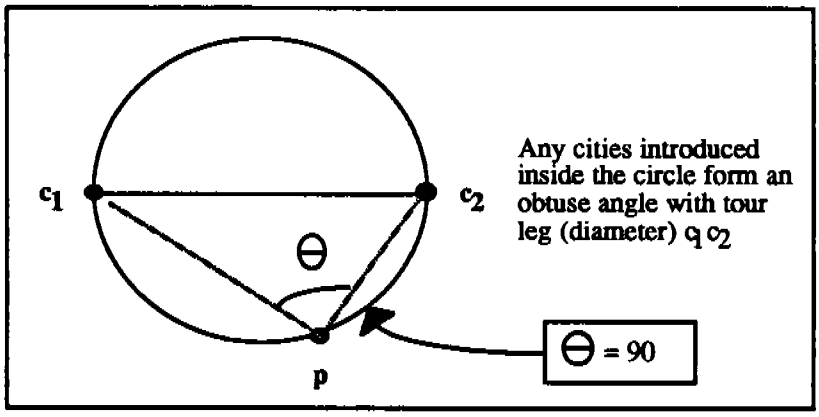


Figure 2. The Obtuse Condition Defined on a Tour Leg

We formalize as follows. Let p_1 and p_2 be the coordinates of the cities which lie at the endpoints of a specific tour leg (note: since a circle is rotationally invariant, we do not bother to rotate and translate the tour leg to an axis).

$$m = midp(p_1, p_2) = (h, k)$$
$$r = dis(m, p_1)$$
Right – angle condition:
$$(x-h)^2 + (y-k)^2 = r^2$$
Obtuse condition:
$$(x-h)^2 + (y-k)^2 < r^2$$
Acute condition:
$$(x-h)^2 + (y-k)^2 > r^2$$

Therefore, if a newly introduced city's coordinates lie within the semi-disk about an existing tour leg, the city forms an obtuse angle with the current tour. What remains to be shown is under what criteria the obtuse angle condition is sufficient to insert a new city optimally into the tour. Obviously, for all other segments in the tour, the discriminating quartic must not pass within the obtuse region of the segment under consideration.

The Effect of Tour Leg Translation upon the Quartic Locus of the ETSP.

To gauge the effect upon the quartic locus of the relative orientation of one tour leg's obtuse region with respect to another, a set of experiments was designed to monitor the transformations undergone by the quartic space when one tour leg is held fixed, while the other is systematically translated to a new position in the plane. A tour leg is defined to be a pair of cities which are currently connected. In the experiments, without loss of generality, the longer of two tour legs is assumed to lie upon the x-axis, with the shorter initially lying across it, at an oblique angle. The shorter segment is then translated in the positive ordinate direction until it lies totally beyond the circumscribing circle of the longer segment. During this process, the question is posed regarding which of the two segments is less costly to perturb when introducing an arbitrary city into the space. Cost in this instance is the elliptic distance, which is defined to be the sum of the distances from the endpoints of a segment to the new city, minus the length of the segment. During the translation process, the quartic locus of equal perturbation is observed at the extremal positions of both the shorter tour leg and its obtuse region. An extremal position for the shorter leg is defined to be the collinearity of one of its endpoints with the longer segment. An extremal position for its obtuse region is defined to be a tangency, either internal or external, with the obtuse region of the longer segment.

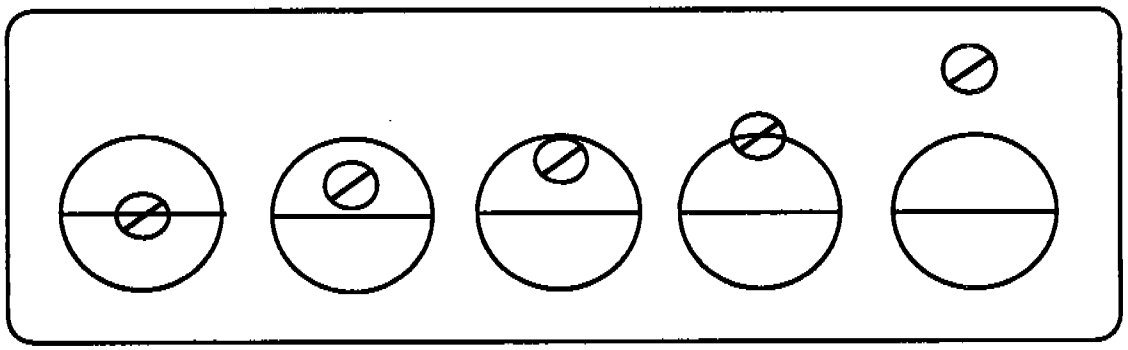


Figure 3. Tour leg translation in the positive ordinate direction.

We know from previous work that the locus of equal perturbation is fourth-order [C3]; the current effort attempts to specify what type of quartic arises for different positions of the segments. There are three situations to consider. First, the obtuse region of the shorter leg may be totally contained within the obtuse region of the longer leg, the point of internal tangency being the extreme. Secondly, the obtuse regions may partially overlap, with extremes at the internal and external tangencies. Finally, the obtuse regions may be disjoint. In the experiments, it was demonstrated that the genus of the quartic curve changes from two to one at the instant when the shorter segment becomes tangent to the longer, and from one to zero when the circumscribing circle about the shorter segment becomes internally tangent to the circumscribing circle of the longer. The empirical evidence for these results is contained in Appendix A. In the majority of the experiments, the direction of translation was in the positive ordinate direction. The exceptions are at graphics A10--A12. A summary of the experimental results appears below.

Situation 1. Segment CD's obtuse region is properly contained within that of segment AB.

Case 1.1. Tour leg CD lies properly across tour leg AB.

The locus is of genus two and its shape approximates a figure eight. The lobes of the figure eight are proportioned to the relative sizes of the pieces of tour leg CD defined by the crossing. For practical purposes, it has been shown that this condition cannot occur in an optimal tour [F1]. However, for the sake of completeness, it is included here.

Case 1.2. CD is properly tangent to AB (C collinear with AB; C=A and C=B).

The quartic locus is pear-shaped, with the point of the pear at C. This extremal condition corresponds to one lobe of the figure eight being lopped off, and changes the genus of the locus from two to one.

Case 1.3. CD lies properly within one of the semicircles which straddle leg AB.

If CD is nearer AB than AB's circle of obtuseness, the locus is pear-shaped. As CD nears the obtuse circle about segment AB, the cusp of the pear becomes smoother, and the locus resembles a quartic ellipsoid. The major axis of the ellipsoid approaches the medial axis of the tour legs as a limiting condition.

Situation 2. Segment CD's obtuse region partially overlaps that of segment AB.

Case 2.1. Tour leg CD lies properly across tour leg AB.

If CD lies within the obtuse circle of AB, then the description at case 1.1 applies. However, if CD extends outside the obtuse circle of AB, then one of the lobes of the figure eight opens up, and the locus is similar to the loop branch of Durer's conchoid. The remaining lobe either envelopes the section of CD which does not protrude beyond AB, or the section of AB which is nearest CD.

Case 2.2a. CD is properly tangent to AB (C collinear with AB; $C \neq A$ and $C \neq B$).

The quartic locus is a paraboloid, with a cusp at the point where C touches AB.

Case 2.2b. CD is improperly tangent to AB (C collinear with AB; C=A). Without loss of generality, we assume C=A. The quartic locus degenerates to a hyperbola, as proven at [C3].

Case 2.3. CD lies properly within one of the semicircles which encompass leg AB.

The quartic locus is a paraboloid, with a spectrum of behaviors. If CD is roughly parallel to AB, the locus is similar to the bullet nose; however, if one of the endpoints of CD is pointed at AB, the locus is cusped or sharply lobed about the endpoint. The cusp or lobe smooths out as CD's obtuse region moves away from segment AB proper and approaches the point of internal tangency with that of AB.

Situation 3. Segment CD's obtuse region is disjoint from that of segment AB.

The locus ranges in shape from a quartic paraboloid when CD's obtuse region is relatively near that of AB, to a quartic hyperboloid when the regions become remote. The point at which the change from paraboloid to hyperboloid occurs as yet remains unspecified. As the obtuse regions become increasingly remote, the locus resembles a branch of the classic quartic known as the Kampyle of Eudoxus, which may crudely be described as a hyperbola with inflection points.

The third situation is the one which we will ultimately exploit in the preprocessing algorithm. We require to know under what conditions the quartic locus is disjoint from the obtuse region of the longer tour leg, which will be developed in a separate section below.

The Lisp Function Utilized to Plot the Quartic Locus during the Tour Leg Translation Experiments.

The tour leg translation experiments were conducted on a Macintosh IIfx workstation with 8MB of RAM memory, using a version of the Lisp language called Macintosh Allegro Lisp. Since this version of Lisp does not yet support bitmap operations, the author developed a Lisp function to dump the contents of a window to a global variable, which in turn is passed to a Laserwriter printer netted to the computer. The function which displays the locus is called "plot-loci"; a hardcopy listing of the source code appears below. The logic is essentially a double do loop: the outer loop throttles both the position of the tour leg and the program termination condition, while the inner controls the locus plot for a specified position of the shorter tour leg. Some of the quartic ellipsoids were of such extensive area that only a small portion of them could be displayed on the screen. It is conceivable for one of these ellipsoids to be infinitely long just prior to reaching the point where the smaller obtuse region becomes internally tangent to that of the longer, when the genus of the locus is altered from one to zero, and the locus opens into the shape of a paraboloid.

```
(defun  plot-loci ()
  (prog (i j p1 p2 p3 p4 per1 per2 p mp diff 8-set max newp lastp
         twoback nexttolastp anchor (cnt 0) (passcnt 0) leftflag bothflag m1 m2 m)
(putprop (cadr citydata) 'xy (cons (car (getprop (cadr citydata) 'xy))
                    (cdr (getprop (car citydata) 'xy))))
(display-cities11 citydata)
(dc)
(dc1)
(setq p1 (getprop (car citydata) 'xy))
(setq p2 (getprop (cadr citydata) 'xy))
outerloop (setq passcnt (1+ passcnt))
(setq p3 (getprop (caddr citydata) 'xy))
(setq p4 (getprop (cadddr citydata) 'xy))
(setq m (midpoint p3 p4))

(setq i (round (car m)))
(print i)
(setq j 0)
(setq max 1000000)
(setq cnt 0)
(setq bothflag nil)
(setq leftflag nil)

jloop (setq j (1+ j))
(cond ((= j 600)(go init)))
(setq per1 (per-points (list p1 (cons i j) p2)))
(setq per2 (per-points (list p3 (cons i j) p4)))
(setq diff (abs (- per1 per2)))
(cond ((< diff max)(setq max diff)
(setq anchor (cons i j))))
(go jloop)

init
(cond ((> max .05)(setq i (1+ i))(print i)(setq j 0)(go jloop)))
(setq mp (make-point (car anchor)(cdr anchor)))
(ask tsw (move-to mp))(ask tsw (line-to mp))
(setq p anchor)

anchor (setq cnt (1+ cnt))
; Function 8-set finds the eight digital neighbors of coordinate p
(setq 8-set (8-set p))
(setq 8-set (delete lastp 8-set :test #'equal))
(setq 8-set (delete nexttolastp 8-set :test #'equal))
(setq 8-set (delete twoback 8-set :test #'equal))
(setq max 1000000)
; Function per-points computes the perturbation length of three points, with the new point in
the middle
(imapc 8-set (function (lambda (x)
(setq per1 (per-points (list p1 x p2)))
(setq per2 (per-points (list p3 x p4)))
(setq diff (abs (- per1 per2)))
(cond ((< diff max)(setq max diff)(setq newp x)))
)))
(cond ((eq cnt 1)(setq newp (caddr 8-set)))
    (leftflag (setq newp (caddr (reverse 8-set)))
          (setq leftflag nil)(setq bothflag t)))
(ask tsw (move-to (make-point (car newp)(cdr newp))))
(ask tsw (line-to (make-point (car newp)(cdr newp))))
bypass (setq twoback nexttolastp)
(setq nexttolastp lastp)
(setq lastp p)
(setq p newp)
```

```
(cond ((eq passcnt 20)(dc1)(return))
    ((equal p anchor)(setq p nil)(setq anchor nil)(setq twoback nil)
                (setq nexttolastp nil)(setq lastp nil)(tr 2)(go outerloop))
    ((> (car p) 600)(setq leftflag t)(setq p anchor)(setq lastp nil)
     (setq twoback nil)(setq nexttolastp nil)(go anchor))
     ((> (car p) 600)(setq leftflag t)(setq p anchor)(setq lastp nil)
      (setq twoback nil)(setq nexttolastp nil)(go anchor))
      ((< (cdr p) 0)(cond (bothflag (setq p nil)(setq lastp nil)
                                          (setq anchor nil)(setq twoback nil)
                                  (setq nexttolastp nil)(tr 20)(go outerloop))
                                      (t (setq leftflag t)(setq p anchor)
                                (setq lastp nil)(setq twoback nil)
                                          (setq nexttolastp nil)(go anchor))))
      ((< (car p) 0)(cond (bothflag (setq p nil)(setq lastp nil)
                                      (setq anchor nil)(setq twoback nil)
                                  (setq nexttolastp nil)(tr 20)(go outerloop))
                                      (t (setq leftflag t)(setq p anchor)
                                  (setq lastp nil)
                                  (setq twoback nil)
                                   (setq nexttolastp nil)
                                  (go anchor)))))
(go anchor)
))
```

The Coordinate Data for the Tour Leg Translation Experiments.

Table 1 below records the positions of the four coordinates representing the endpoints of the two tour legs utilized in the experiments. The Allegro Lisp environment employs a windowing system in which the upper left corner of a window is the origin, so a translation in what is conventionally considered to be the positive ordinate direction produces an ordinate of lesser magnitude.

Graphics A1-A9 portray experiments in which the shorter tour leg is translated in the positive ordinate direction, in a series of iterated steps which vary in size. Graphics A10 through A12 involve a translation along the longer tour leg, in either the positive or negative direction. In selecting the coordinates, an attempt was made to develop a dataset representative of a variety of quartic behaviors, although the sclection process was not exhaustive.

Graphic	*Position of C1*	*Position of C2*	*C3's start*	*C4's start*	*C3's finish*	*C4's finish*	*Stepsize*
A1	(186, 490)	(520, 490)	(326, 485)	(340, 477)	(326, 335)	(340, 327)	5
A2	(150, 547)	(510, 547)	(302, 546)	(332, 510)	(302, 312)	(332, 276)	2
A3	(206, 579)	(408, 579)	(290, 592)	(326, 567)	(290, 32)	(326, 7)	20
A4	(136, 536)	(506, 536)	(275, 563)	(349, 440)	(275, 323)	(349, 200)	30
A5	(256,562)	(328, 562)	(276, 571)	(318, 556)	(276, 31)	(318, 16)	20
A6	(175, 521)	(507, 521)	(238, 510)	(453, 495)	(238, 310)	(453, 295)	20
A7	(140, 512)	(482, 512)	(292, 512)	(233, 401)	(292, 450)	(233, 349)	1
A8	undoc.	undoc.	undoc.	undoc.	undoc.	undoc.	undoc.
A9a-i	undoc.	undoc.	undoc.	undoc.	undoc.	undoc.	undoc.
A10	(6, 379)	(208, 379)	(160, 379)	(196, 354)	(167, 379)	(203, 354)	1
A11	(182, 477)	(457, 477)	(209, 477)	(263, 389)	(159, 477)	(213, 389)	-5
A12a-c	(287, 449)	(521, 449)	(298, 470)	(322, 412)	(300,270)	(324,412)	1

Table 1. Initial and Terminal Positions of the Tour Leg Translation City Data.

Explanation of the Graphics contained in Appendix A.

Appendix A contains a series of computer graphics which visually depict the effect of translating a shorter tour leg while holding a longer one fixed, while at the same time requesting a plot of the corresponding quartic locus of equal tour leg perturbation. Recall that a perturbation is a synthetic operation which produces two new tour edges by constructing segments from the endpoints of a tour leg to a new city, while at the same time discarding the edge currently defined by the tour leg. The elliptic length of the perturbation is the sum of the lengths of the two new edges, minus the length of the old edge. The locus of equal tour leg perturbation is the set of points where the elliptic lengths are the same for two tour legs. In the graphics, the longer tour leg is oriented along the x-axis, and the shorter one is systematically translated to some other position in the plane.

Although upon first inspection it may appear that the translation process is non-robust because the genus of the quartic locus may suddenly change during a translation of a single pixel in a specific direction, it should be pointed out that the translation process is limited by the grain size (screen resolution) of the monitor. If one is permitted to zoom in on the graphics to view the locus at a finer resolution, there is actually an infinitely long spectrum of behavior between shifts in the genus of the locus. The zooming operation can be effectively achieved by simply scaling up the coordinates of the four cities by a nominal factor, and redisplaying the data (or a portion of it) in the window.

In many of the graphics contained in Appendix A, the iterated positions of the obtuse region for the shorter tour leg are seen as a series of circles plotted in what appears to be a cylindrical formation. This information tends to occlude the quartic locus in some instances, but it was decided to include it so that the reader might get a more intuitive appreciation of the position of the locus as a function of the location of the smaller obtuse region.

Graphic A1. This experiment translates a smaller tour leg up and away from a relatively large one. The locus is seen to evolve from a small piriform through a family of ever-larger quartic ellipsoids, culminating in an open paraboloid beyond the point at which the smaller obtuse region becomes internally tangent to that of the larger.

Graphic A2. The smaller tour leg is longer than in the first experiment, which produces a noticeably wider funnel of quartic ellipsoids during the translation process.

Graphic A3. The shorter tour leg is translated well beyond the obtuse region of the longer, to highlight the quartic paraboloids and hyperboloids which appear when the tour legs become remote. In the initial position, the tour legs cross, which produces a locus in the form of a figure eight.

Graphic A4. In this case, the shorter tour leg is approximately equal to the length of the radius of the obtuse region about the longer leg. The locus demonstrates a typical change in genus produced by the figure eight evolving into the piriform, followed by a smoother ellipsoid (the cusp of the piriform is modulated), and finally by a genus zero family of paraboloids.

Graphic A5. The shorter tour leg is now longer than the radius of the longer leg, and also more parallel to the longer leg. The locus consequently becomes flattened, with a bulletnose quartic behavior predominating over the piriform.

Graphic A6. A more detailed look at the bulletnose behavior exhibited by graphic G5, with an emphasis on the change from genus one to genus zero as the point of internal tangency is encountered during translation.

Graphic A7. A somewhat detailed look at the piriform behavior of the locus as a shorter tour leg's obtuse region is translated from internal tangency with that of the longer, to a position slightly beyond the obtuse circle of the longer.

Graphic A8. A remote view of the quartic ellipsoids encountered in an undocumented translation experiment.

Graphics A9a-i. For the sake of clarity, each locus is plotted to its own sheet of paper as a tour leg is translated from a crossing with the longer segment (eight figure locus) through tangency with the longer segment (piriform locus) to a position well beyond the circumscribing circle of the longer leg (locus has evolved into a quartic hyperboloid).

Graphic A10. We change the direction of translation by laying one endpoint of the shorter leg upon the longer leg, and then walking the shorter leg to the right (positive abscissa direction). The piriform bends radically in the direction of translation until it opens at the point of internal tangency.

Graphic A11. This time a tour leg is started inside the larger obtuse region, with one endpoint upon the longer leg, and then walked to the left (negative abscissa direction) until it passes outside. At the point when the endpoint coincides with that of the longer, the locus is the familiar second degree hyperbola. In a neighborhood about this point, the locus appears to be a serpentine (cubic).

Graphics A12a-c. This experiment dramatically illustrates the effect of translating a tour leg by one pixel (screen coordinate), to radically alter the appearance of the quartic locus. The shorter tour leg is walked from left to right in this case, and the tour legs cross. A folium-shaped quartic results at the original position of the tour leg, with a lobe wrapped around an endpoint of the longer leg. The second position of the leg continues to produce a lobe about the same endpoint. However, the very next translation of one pixel causes the lobe of the quartic folium to move over and wrap about an endpoint of the shorter segment. It should be emphasized that if one were afforded the luxury of an infinitely high resolution graphics screen, there would be an infinite number of quartic behaviors displayed between the lobe shifts. Machine imprecision can cause processes to appear non-robust, simply because in a discrete process one is not permitted to select a small enough input scale to portray a continuous phenomenon.

Position of Tour Leg CD with respect to Leg AB	***Shape of Quartic Locus for ETSP***	***Genus of Quartic Locus***	***Ideal Example from Antiquity***
Leg CD properly crosses leg AB; CD's obtuse region inside that of AB	Figure eight	2	Eight curve $x^4 = a^2x^2 - a^2y^2$
Leg CD properly crosses leg AB; CD's obtuse region intersects that of AB	Degenerate figure eight (one lobe open)	1	Durer's Conchoid with $a > b$
CD's obtuse region properly internal to that of AB; CD properly tangent to AB	Pear-shaped with cusp	1	Piriform $a^4y^2 = b^2x^3 (2a - x)$
CD's obtuse region properly internal to that of AB; no crossing	Pear-shaped near AB; Quartic ellipsoid near AB's circle	1	Piriform / Ellipsoid $a^4y^2 = b^2x^3 (2a - x)$ $a^4y^2 = b^2x^4$
CD's obtuse region intersects that of AB; no crossing	Quartic Paraboloid	0	Bullet nose $a^2/x^2 - b^2/y^2 = 1$
AB and CD share endpoint	Hyperbola	0	Hyperbola $x^2/a^2 - y^2/b^2 = 1$
CD outside AB's circle	Quartic Hyperboloid	0	Kampyle of Eudoxus $x^4 = a^2x^2 + a^2y^2$

Table 2. Effect of Tour Leg Translation on the Quartic Search Space of the Euclidean Travelling Salesman Problem

Summary of the Effect of Tour Leg Translation Upon the Quartic Locus.

When arbitrating which of two tour legs to perturb when inserting a new city into an existing tour, one can predict the position of the quartic locus, based upon the relative positions of the circles of obtuseness drawn about each tour leg. The predicted quartic behavior is summarized at Table 2. If the two obtuse regions are disjoint, then one may invoke a simple check to verify that the discriminating quartic does not intersect the larger obtuse region, and make the indicated insertion, when reasoning about the Euclidean Traveling Salesman Problem. If the smaller obtuse region is properly contained within a semicircular region about the larger, then we know that the locus is of genus one.

Some Practical Considerations Concerning the Quartic Locus.

Some of the quartic loci observed in the tour leg translation experiments are not encountered in practice, when actually constructing shortest tours. For example, it has been shown that tour legs cannot cross in an optimal tour [F1], so we need not be concerned with the eight-curve or the folium when building shortest tours. Also, the pure form of the piriform which occurs when a tour leg endpoint is collinear with another tour leg cannot happen, since that endpoint would in fact be optimally absorbed into the other tour leg. In general, it may be said that the more extreme forms of quartics (those which are of higher genus, or contain cusps, or multiple singularities) need not be treated when constructing optimal tours, since there exists some other tour connection which is optimal, with a simpler quartic available to arbitrate the decision.

The Intersection of the Quartic Locus with the Circle of Obtuseness of a Tour Leg.

To exploit the condition specifying that the quartic locus of equal tour leg perturbation does not intersect the obtuse region about a particular tour leg, we are required to find the values of x and y for the limiting case in which the locus is tangent to the circle which circumscribes the tour leg. Any interior points which lie between the quartic and tour leg (within the obtuse region) are then safe to insert into the tour. Conversely, any points lying beyond the locus (i.e., on the other leg's side) cannot be inserted into the candidate tour leg. The point of tangency lies upon both the quartic locus and the circumscribing circle of the tour leg under consideration. If a quartic locus does not intersect the obtuse region of a tour leg as in the figure below, it is safe to insert any interior cities which happen to fall within the obtuse region. One must ensure that the quartic locus does not encroach into the circle of obtuseness, or some other segment would then be the source of optimal perturbation for cities bounded below by the quartic and above by the obtuse region.

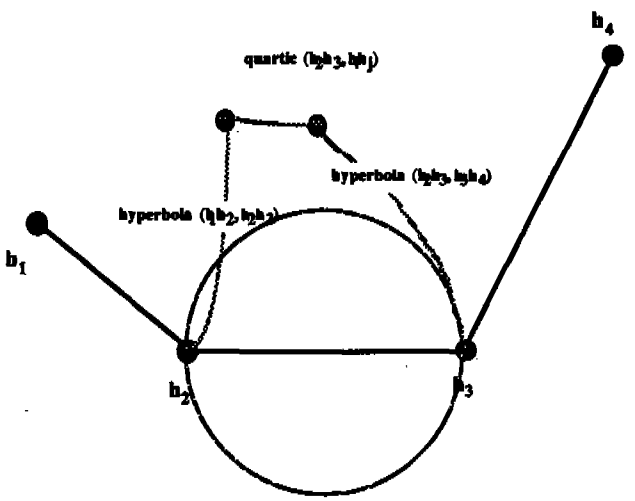


Figure 4. Maximal obtuseness.

Simultaneous Solution of the Quartic Locus and the Obtuse Circle about a Tour Leg

The quartic locus:

$$\sqrt{(x+h)^2+y^2}+\sqrt{(x-h)^2+y^2}-2h=$$
$$\sqrt{(x-a)^2+(y-b)^2}+\sqrt{(x-c)^2+(y-d)^2}-\sqrt{(a-c)^2+(b-d)^2} \qquad [1a]$$

The obtuse condition:

$$x^2+y^2=h^2 \qquad [2a]$$
$$\Rightarrow\ y^2=h^2-x^2 \qquad [3a]$$
$$\Rightarrow\ y=\pm\sqrt{h^2-x^2} \qquad [4a]$$

Substitution of [3a] in [1a] produces:

$$\sqrt{2h(h+x)}+\sqrt{2h(h-x)}-2h=$$
$$\sqrt{-2ax-2by+a^2+b^2+h^2}+\sqrt{-2cx-2dy+c^2+d^2+h^2}-\sqrt{(a-c)^2+(b-d)^2} \qquad [5a]$$

We allow a parameter z to represent the quantity on the left side of [5a]:

$$Let\ \ z=\sqrt{2h(h-x)}+\sqrt{2h(h+x)}-2h \qquad [6a]$$
$$Then\ \ z+2h=\sqrt{2h(h-x)}+\sqrt{2h(h+x)} \qquad [7a]$$
$$\Rightarrow\ z^2+4hz+4h^2=2h^2-2hx+2h^2+2hx+2\sqrt{(2h^2-2hx)(2h^2+2hx)} \qquad [8a]$$
$$\Rightarrow\ z^2+4hz=4h\sqrt{h^2-x^2} \qquad [9a]$$
$$\Rightarrow\ y=\frac{z^2+4hz}{4h} \qquad [10a]$$

Also, by squaring both sides of equation [9a], we obtain:

$$z^4+8hz^3+16h^2z^2=16h^2(h^2-x^2) \qquad [11a]$$
$$\Rightarrow\ x^2=\frac{16h^4-16h^2z^2-8hz^3-z^4}{16h^2} \qquad [12a]$$

Therefore, in terms of z, the parametric equations of the locus are:

$$x=\frac{\pm\sqrt{16h^4-16h^2z^2-8hz^3-z^4}}{4h}\ ; \qquad [13a]$$
$$y=\frac{z^2+4hz}{4h} \qquad [14a]$$

But x is real

$$\Leftrightarrow \quad 16h^4 - 16h^2z^2 - 8hz^3 - z^4 \geq 0 \qquad [15a]$$

$$\Leftrightarrow \quad 16h^4 \leq 16h^2z^2 + 8hz^3 + z^4 \qquad [16a]$$

$$\Leftrightarrow \quad 16h^4 \leq z^2(16h^2 + 8hz + z^2) \qquad [17a]$$

$$\Leftrightarrow \quad 16h^2 + 8hz + z^2 \geq \frac{16h^4}{z^2} \qquad [18a]$$

$$\Leftrightarrow \quad \text{the LHS is positive; } z \neq 0.$$

But $16h^2 + 8hz + z^2 = 0$ [19a]

$$\Leftrightarrow \quad (z+4h)^2 = 0 \qquad [20a]$$

$$\Leftrightarrow \quad z = -4h \qquad [21a]$$

Therefore the LHS of [18a] is positive

$$\Leftrightarrow \quad z > -4h \qquad [22a]$$

$\therefore$ There are no real solutions to [13a]

$$\Leftrightarrow \quad z < -4h \ . \qquad [23a]$$

If the left side of equation [1a] is plotted as a parameter in three dimensions, with both x and y ranging within the interval [-10, 10), while fixing h at 1, then the graphic depicted at Figure 5 results. This illustration was computed using the Plot3D function available with the software tool Mathematica 2.0, copyrighted by Wolfram Research, Inc. for the Apple Macintosh family of computers.

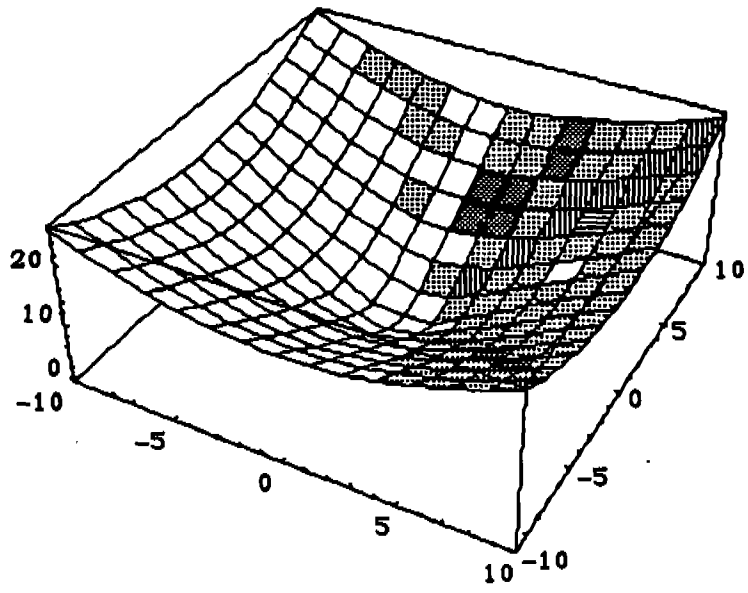


Figure 5. A Parametric Plot of the Left Side of Equation [1a], with h equal to 1.

Now turning to the right side of [5a], *we solve for the same real parameter z:*

$$z+\sqrt{(a-c)^2+(b-d)^2}=\sqrt{-2ax-2by+a^2+b^2+h^2}+\sqrt{-2cx-2dy+c^2+d^2+h^2} \qquad [24a]$$

Squaring both sides of [24a] *produces:*

$$z^2+2z\sqrt{(a-c)^2+(b-d)^2}+a^2+b^2+c^2+d^2-2ac-2bd=$$
$$-2ax-2cx-2by-2dy+a^2+b^2+c^2+d^2+2h^2+$$
$$2\sqrt{(-2ax-2by+a^2+b^2+h^2)(-2cx-2dy+c^2+d^2+h^2)} \qquad [25a]$$

which simplifies to:

$$z^2+2z\sqrt{(a-c)^2+(b-d)^2}+2ax+2cx+2by+2dy-2ac-2bd+2h^2=$$
$$2\sqrt{\begin{array}{l}4acx^2+4adxy-2ac^2x-2ad^2x-2ah^2x\\+4bcxy+4bdy^2-2bc^2y-2bd^2y-2bh^2y\\-2a^2cx-2a^2dy+a^2c^2+a^2d^2+a^2h^2\\-2b^2cx-2b^2dy+b^2c^2+b^2d^2+b^2h^2\\-2ch^2x-2dh^2y+c^2h^2+d^2h^2+h^4\end{array}} \qquad [26a]$$

If we square both sides once again we obtain:

$$z^4+4z^3\sqrt{(a-c)^2+(b-d)^2}+4axz^2+4cxz^2+4byz^2+4dyz^2-4acz^2-4bdz^2+4h^2z^2$$
$$+4z^2\sqrt{(a-c)^2+(b-d)^2}+8axz\sqrt{(a-c)^2+(b-d)^2}+8cxz\sqrt{(a-c)^2+(b-d)^2}$$
$$+8byz\sqrt{(a-c)^2+(b-d)^2}+8dyz\sqrt{(a-c)^2+(b-d)^2}-8acz\sqrt{(a-c)^2+(b-d)^2}$$
$$-8bdz\sqrt{(a-c)^2+(b-d)^2}-8h^2z\sqrt{(a-c)^2+(b-d)^2}+4a^2x^2+8abxy-8abdx$$
$$+4c^2x^2+8cdxy-8bcdx+4b^2y^2-8abcy+4d^2y^2-8acdy+8abcd+8ach^2+8bdh^2=$$
$$8acx^2+8adxy-8ad^2x+8bcxy+8bdy^2-8bc^2y-8a^2dy+4a^2d^2+4a^2h^2-8b^2cx$$
$$+4b^2c^2+4b^2h^2+4c^2h^2+4d^2h^2 \qquad [27a]$$

The Set of Non-linear Constraints to Assure that the Obtuse Condition is Sufficient to Guarantee Optimality.

We have developed the rudiments of a preprocessing algorithm, since we know under what geometric conditions it is safe to insert a city into a tour if it happens to form an obtuse angle with some existing tour leg. The constraints which must be included in the formal design specification for the algorithm are stated below:

Maximally obtuse:	$(x-h)^2+(y-k)^2 < r_j^2$;
Left – hyperbolic conformable:	$x > H_l$;
Right – hyperbolic conformable:	$x < H_r$;
Semi – positive:	$y \geq 0$;
ETSP quartic conformable:	$y < \Phi_i^4(x,y)$, $\Phi_i \in \{\Phi^4(x,y)\}$; $i = 1,...k$.

The algorithm may then be outlined as follows:

0. Begin with a baseline tour consisting of the convex hull.

1. Sort all interior cities based on the maximal angle formed with the (extended) hull, and form an open list. If there are no interior cities or obtuse angles, return the extended hull structure.

2. Select the city at the head of the open list as a candidate to be inserted. If the open list is null go to step 1.

3. If the candidate's obtuse region is disjoint from that of all non-neighboring tour legs, insert the candidate; otherwise use the quartic locus for the decision.

4. Recheck all previously inserted cities for obtuseness and reorder if necessary, and go to step 2.

Results for two Certified Databases.

The new preprocessing algorithm has been applied to a variety of small to moderate size databases, the largest being the 127-city University of Augsburg dataset [R1], and the 532-city Bell Laboratories dataset [P1]. Both of these datasets have been certified to optimality by means of a version of the branch-and-cut algorithm. For the 127-city database, 35 cities are preprocessed by the new algorithm, and for the 532-city database, 151 cities are preprocessed. For each of these two instances, over a quarter of the database is successfully preprocessed into a tour which is optimal for the cities it contains. Graphics of the complete optimal tours for the datasets, and for the partial optimal tours produced by the preprocessing algorithm are contained in Appendix B. An explanation of the graphics is as follows:

Graphic B1. The locations of 127 beer gardens in the city of Augsburg, Germany.

Graphic B2. The best tour found with branch-and-cut, by researchers at the University of Augsburg.

Graphic B3. The baseline tour found by the new preprocessing algorithm which exploits quartic loci that do not pass through the obtuse region of a current tour leg.

Graphic B4. The quartic Voronoi diagram for the partial baseline tour. See [C3] for a discussion of the Voronoi diagram for the ETSP. The diagram is a connectivity map which shows how to attach a new city to the existing tour. If the new city lies properly within one of the cells depicted by the diagram, it should

be connected to the endpoints of the tour leg about which the cell wraps, while detaching the old connection. If a city lies at a Voronoi junction (where three quartics intersect) there are three optimal tours; if it lies uniquely on one quartic, there are two optimal tours.

Graphic B5. The locations of 532 Bell telephone offices in the contiguous United States.

Graphic B6. The best tour found with branch-and-cut, by researchers at New York University and the University of Rome.

Graphic B7. The baseline tour found by the new preprocessing algorithm.

It should be made clear that although the partial tours produced by the algorithm are optimal for the cities which they contain, a complete optimal tour may in fact appear quite different in shape than the partial tours produced by the preprocessing algorithm. For example, suppose a cluster of internal cities remains unprocessed after the algorithm runs its course because each of the cities in the cluster forms an acute angle with the extended hull. It is possible for the cluster to combine two perturbations of the hull produced by the preprocessing algorithm into a single optimal subtour originating from some other hull segment, thereby radically altering the shape of the tour produced by the preprocessing algorithm.

Summary.

For the Euclidean traveling salesman problem, an algorithm has been presented which preprocesses any cities which form a disjoint, maximally obtuse angle with the convex hull, or for that matter with the resultant structure. The utility of the obtuse condition is to ensure that the interior cities which satisfy the criterion lie upon the appropriate side of the quartic locus which discriminates the tour leg perturbation of minimal length. With this enhancement, the hull is extended until only interior cities at an acute angle remain to be inserted into the tour. The algorithm has time complexity O[n log n], where n is the number of cities. In the best case, if all interior cities form a disjoint, maximally obtuse angle with the hull or its extended structure, an instance of the Euclidean traveling salesman problem is solved in O[n log n] time. At the other extreme, if no cities meet the criterion, then no advantage is obtained. During the development of the algorithm, an experiment was conducted to monitor the effect of tour leg translation on the quartic search space of the ETSP. The effect of a translation is to change the orientation of the obtuse regions surrounding two tour legs, so that a newly introduced city may lie within one or the other, within both, or within neither. Empirical observation suggests that there are three genuses of quartic curves manifested during shortest tour construction, only two of which are admissible as legitimate constructs. For these two, the genus is seen to change from one to zero at the point where the circumscribing circle of the shorter tour leg is internally tangent to that of the longer leg. The preprocessing algorithm exploits the condition for which the genus is zero, while simultaneously there is no real intersection of the locus with the obtuse region of the longer leg.

Acknowledgments.

A note of gratitude to Dr. Ken Clark of the Army Research Office for an invitation to present some of the results in a special session on large-scale optimization at the Ninth Annual Army Conference on Applied Mathematics and Computing. Thanks to Dr. Jack Robertson of the US Military Academy for his assistance in leveraging the computer software tools MACSYMA and Mathematica against the elliptic perturbation equation discussed at the beginning of the paper. I would also like to thank Professor C.T. Kelley of North Carolina State University for hosting an excellent conference on Numerical Methods in Differential Equations and Control, at which a number of the results were presented in a poster session.

Valuable suggestions from many scientists and engineers have been incorporated into the research over the last two years. Among those whose comments have been appreciated are Gerald Andersen, Richard Antony, Jacob Barhen, Robert Bixby, Chris Bogart, Roger Brockett, Paul Broome, Mel Brown, Ken Clark, Jagdish Chandra, Doug Chubb, Ben Cummings, Michael Dillencourt, Francis Dressel, Herbert Edelsbrunner, Geoffrey Fox, Ray Freeman, Martin Groetschel, Andrew Harrell, Bob Hein, William

Jackson, David S. Johnson, Shen Lin, Andrew Logan, Sanjoy Mitter, John Pfaltz, Carl Russell, Jay Setheraman, Robert Somoano, J. Michael Steele, Andrew Thompson, Paul Tseng, Franz-Erich Volter, and David Willshaw.

The early sections of the paper dealing with the background of the Euclidean traveling salesman problem, verifying the optimality of a tour, and the Army discovery of the non-linear search space are in large part borrowed from [C4], and are reproduced here as a means of boilerplate introduction to the problem.

Bibliography

[B1] Barachet, L.L., "Graphic Solution of the Traveling Salesman Problem", **Operations Research 5**, 1957, pp. 841-845.

[B2] Bellmore, M., and G.L. Nemhauser, "The Traveling Salesman Problem: A Survey", **Operations Research 16**, 1968, pp. 538-558.

[C1] Chazelle, B., "On the Convex Layers of a Convex Set", **IEEE Trans. Inform. Theory IT-31**, 1985, pp. 509-517.

[C2] Clemens, C.H., A Scrapbook of Complex Curve Theory, Plenum Press, New York NY, 1980.

[C3] Cronin, T.M., "The Voronoi Diagram for the Euclidean Traveling Salesman Problem is Piecemeal Quartic and Hyperbolic", Transactions of the Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Army Research Office Report ARO91-1, June 1990.

[C4] Cronin, T.M., "Maintaining Incremental Optimality when Building Shortest Euclidean Tours", publication pending, Proceedings of the Twenty-Sixth Army Conference on the Design of Experiments, University of Delaware, Newark DE, October 1990.

[F1] Flood, M.M., "The Traveling Salesman Problem", **Operations Research 4**, 1956, pp. 61-75.

[G1] Garey, M.R., R.L. Graham, and D.S. Johnson, "Some NP-complete Geometric Problems", Eighth Annual Symp. on Theory of Comp., May 1976, pp. 10-22.

[G2] Garey, M.R., and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness, W.H. Freeman and Company, New York NY, 1979.

[G3] Golden, B.L., and W.R. Stewart, "Empirical Analysis of Heuristics", Chapter 7 in Lawler, E.L., J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys (eds.), The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization, John Wiley and Sons, New York NY, 1985.

[H1] Held, M., and R.M. Karp, "The Traveling Salesman Problem and Minimum Spanning Trees: Part II", **Mathematical Programming 1**, 1971, pp. 6-25.

[H2] Hilton, H., Plane Algebraic Curves, Oxford at the Clarendon Press, London, 1920.

[J1] Johnson, D.S., Private communication, and set of viewgraphs entitled "How to Beat Lin-Kernighan", *Workshop on Computational Aspects of the Traveling Salesman Problem*, Rice University, Houston TX, April 1990.

[K1] Kavanau, J.L., Curves and Symmetry, Vol. I, Science Software Systems, Los Angeles CA, 1982.

[K2] Kirkpatrick, D.G., and R. Seidel, "The Ultimate Convex Planar Hull Algorithm?", **SIAM J. Computing 15**, 1986.

[K3] Kline, M., Mathematical Thought from Ancient to Modern Times, Oxford University Press, New York, 1972.

[L1] Lawler, E.L., J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys (eds.), The Traveling Salesman Problem: A Guided Tour of Combinatorial Optimization, John Wiley and Sons, New York NY, 1985.

[L2] Lawrence, J.D., A Catalog of Special Plane Curves, Dover Publications, New York NY, 1972.

[L3] Lin, S., and B.W. Kernighan, "An effective heuristic algorithm for the traveling salesman problem", **Operations Research 21**, 1973, pp. 498-516.

[P1] Padberg, M., and G. Rinaldi, "Optimization of a 532-city Symmetric Traveling Salesman Problem by Branch and Cut", **Operations Res. Let., Vol 6, Number 1**, March 1987.

[R1] Reinelt, G., "TSPLIB - A Traveling Salesman Problem Library", Institute of Mathematics, University of Augsburg, Augsburg Germany, 1989.

[S1] Salmon, G., A Treatise on the Higher Plane Curves, Photographic reprint of the Third Edition of 1879 (with a number of corrections), G.E. Stechert & Co., New York NY, 1934.

[S2] Stewart, Jr., W.R., "A Computationally Efficient Heuristic for the Traveling Salesman Problem", Proc. 13th Annual Mtg. S.E. TIMS, 1977, pp. 75-85.

[W1] Walker, R.J., Algebraic Curves, Dover Publications, Inc., New York NY, 1950.

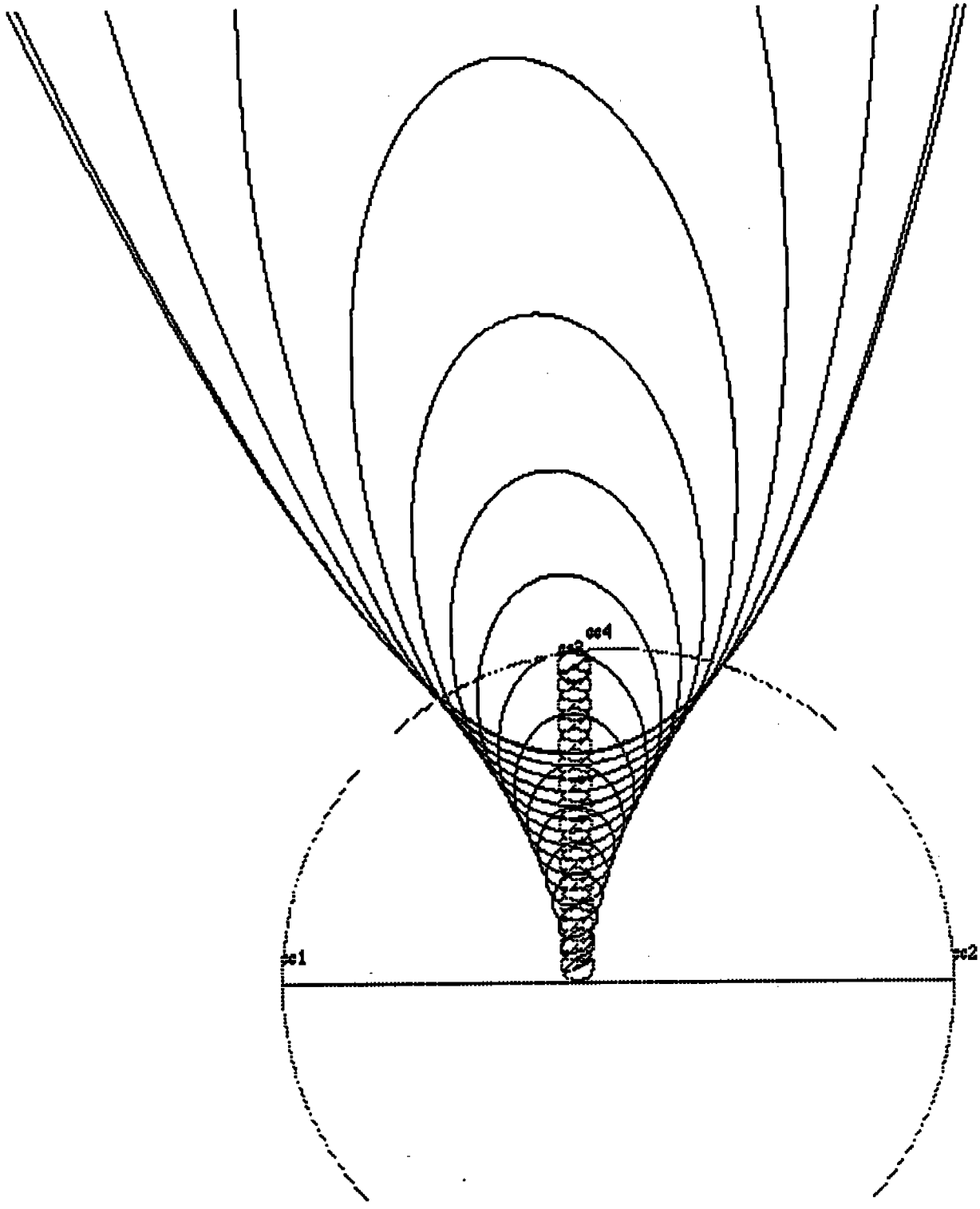
[W2] Whiteside, D., ed., The Mathematical Works of Isaac Newton, Vol. 2, Johnson Reprint Corporation, London, 1967.

Appendix A. Tour Leg Translation and the Quartic Locus.

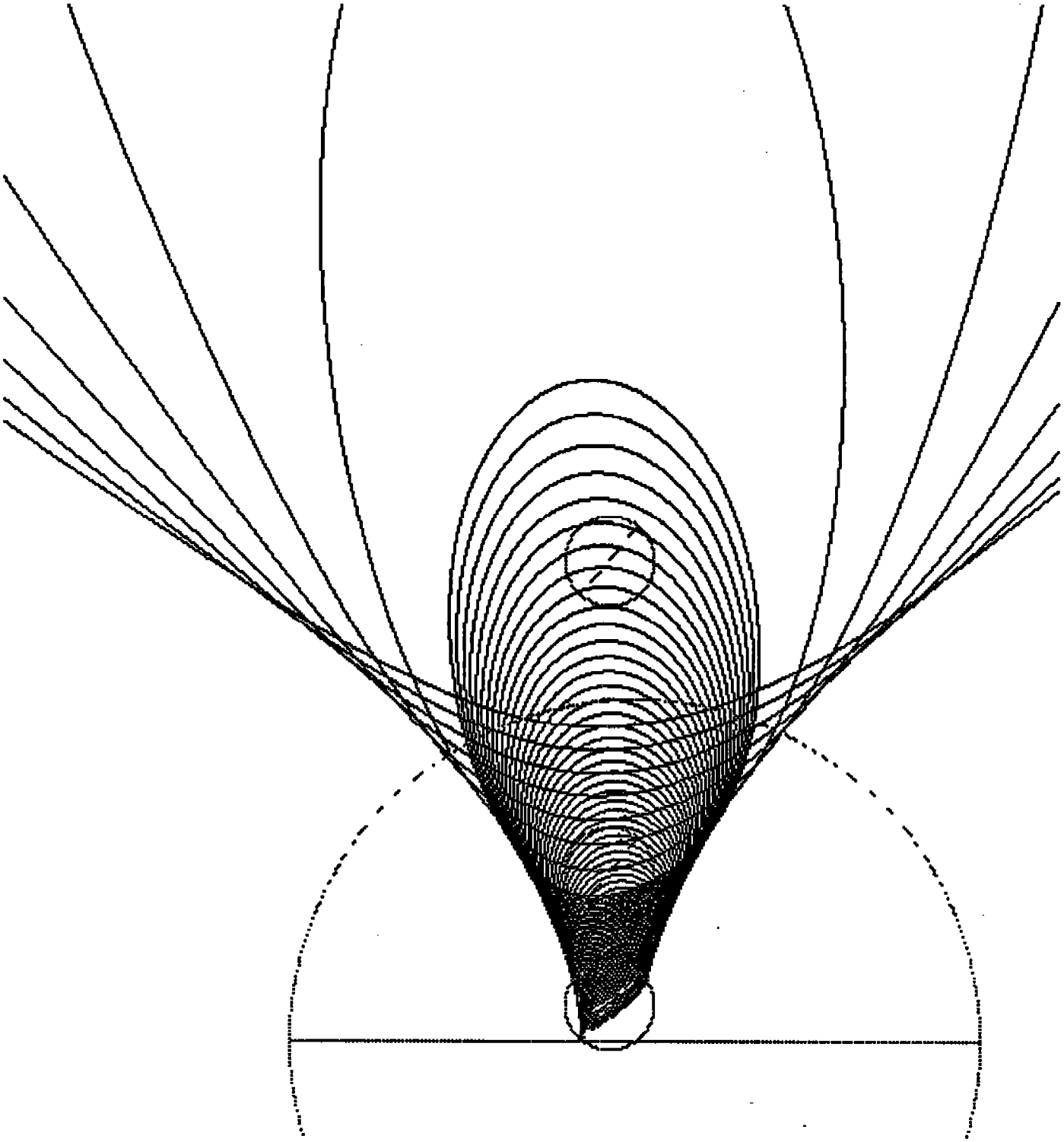
Appendix A contains a series of computer graphics which visually depict the effect of translating a shorter tour leg while holding a longer one fixed, while at the same time requesting a plot of the corresponding quartic locus of equal tour leg perturbation. Recall that a perturbation is a synthetic operation which produces two new tour edges by drawing segments from the endpoints of a tour leg to a new city, while at the same time discarding the edge currently defined by the tour leg. The elliptic length of the perturbation is the sum of the lengths of the two new edges, minus the length of the old edge. The locus of equal tour leg perturbation is the set of points for which the elliptic lengths are the same for two tour legs. In the graphics, the longer tour leg is oriented along the x-axis, and the shorter one is systematically translated to some other position in the plane. The reader is referred to the text for the details of each translation experiment.

Although upon first inspection it may appear that the translation process is non-robust because the genus of the quartic locus may suddenly change with a translation of a single pixel in the ordinate direction, it should be pointed out that the translation process is limited by the grain size (screen resolution) of the monitor. If one is permitted to zoom in on the graphics to view the locus at a finer resolution, there is actually an infinitely long spectrum of behavior between shifts in the genus of the locus. The zooming operation can be effectively achieved by simply scaling up the coordinates of the four cities by a nominal factor, and redisplaying the data (or a portion of it) to the window.

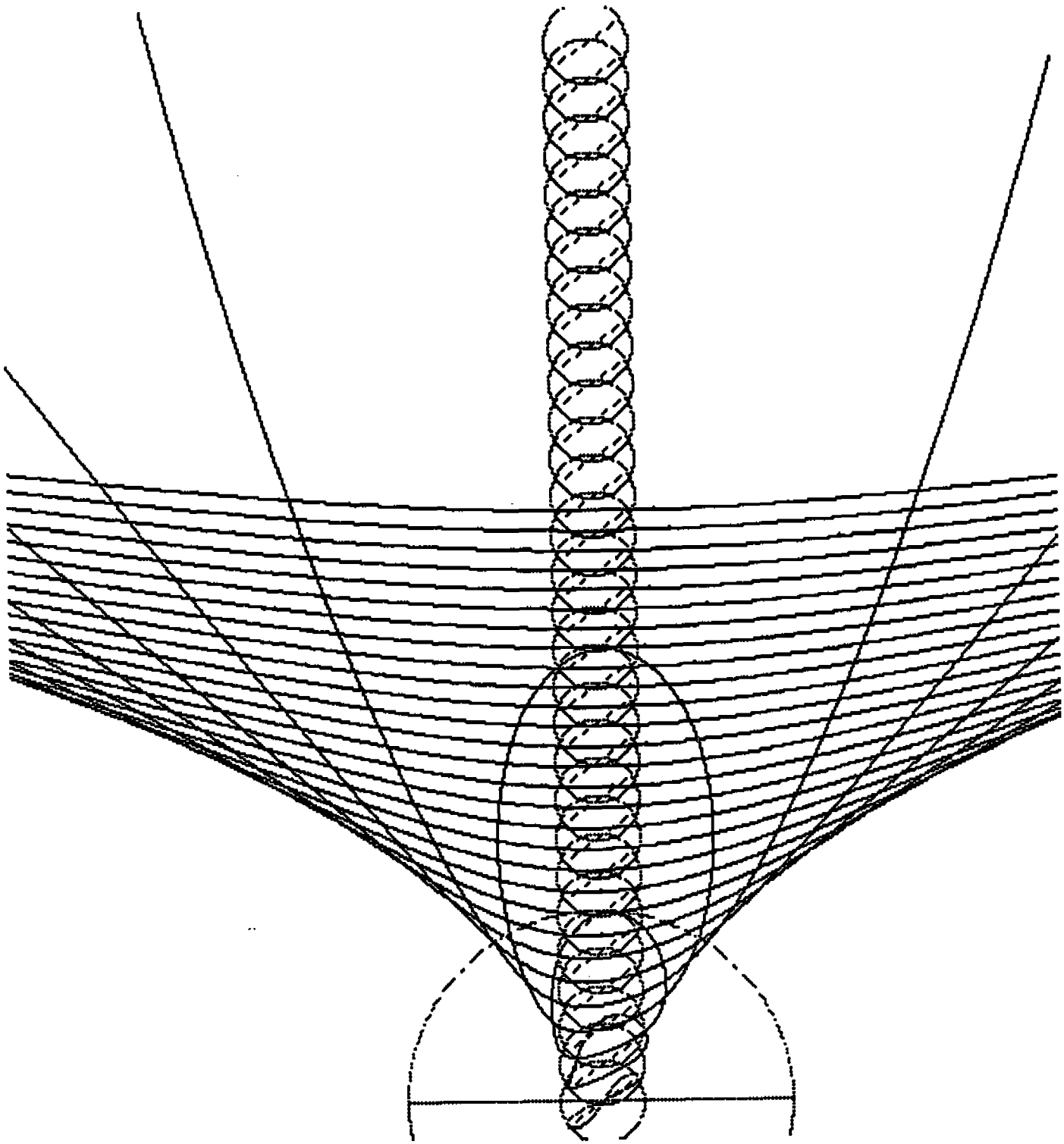
The tour leg translation experiments were conducted on a Macintosh IIfx workstation with 8MB of RAM memory, using a version of the Lisp language called Macintosh Allegro Lisp. Since this version of Lisp does not yet support bitmap operations, the author developed a Lisp function to dump the contents of a window to a global variable, which in turn is passed to a Laserwriter printer netted to the computer. The function which displays the locus is called "plot-loci"; a hardcopy listing of the source code accompanies the text. The logic is essentially a double do loop: the outer loop throttles both the position of the tour leg and the program termination condition, while the inner controls the locus plot for a specified position of the shorter tour leg. Some of the quartic ellipsoids were of such extensive area that only a small section of them could be displayed on the screen. It is conceivable for one of these ellipsoids to be infinitely long just prior to reaching the point where the smaller obtuse region becomes internally tangent to that of the longer, where the genus of the locus is altered from one to zero, and the locus opens into the shape of a paraboloid.



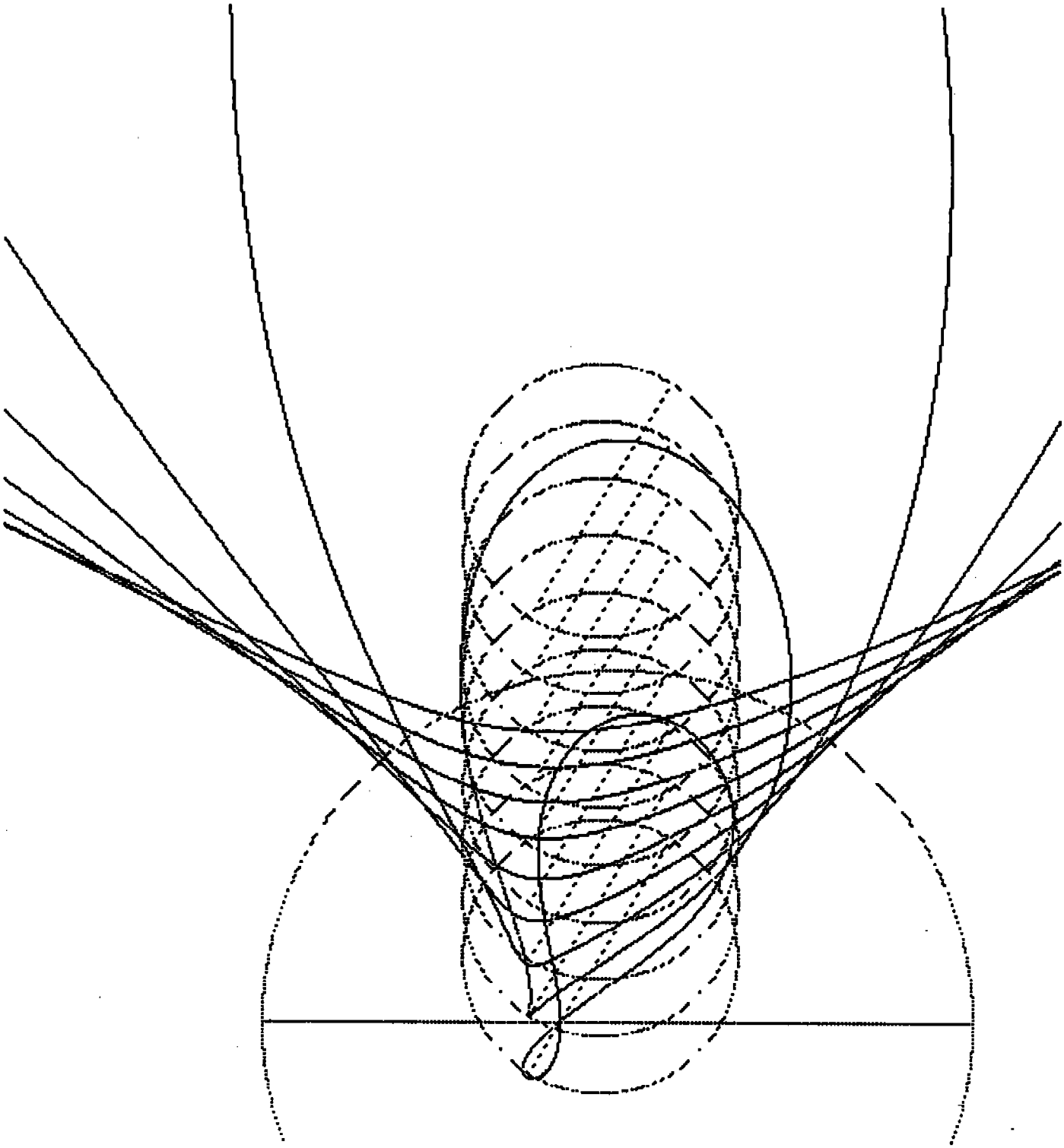
A-1



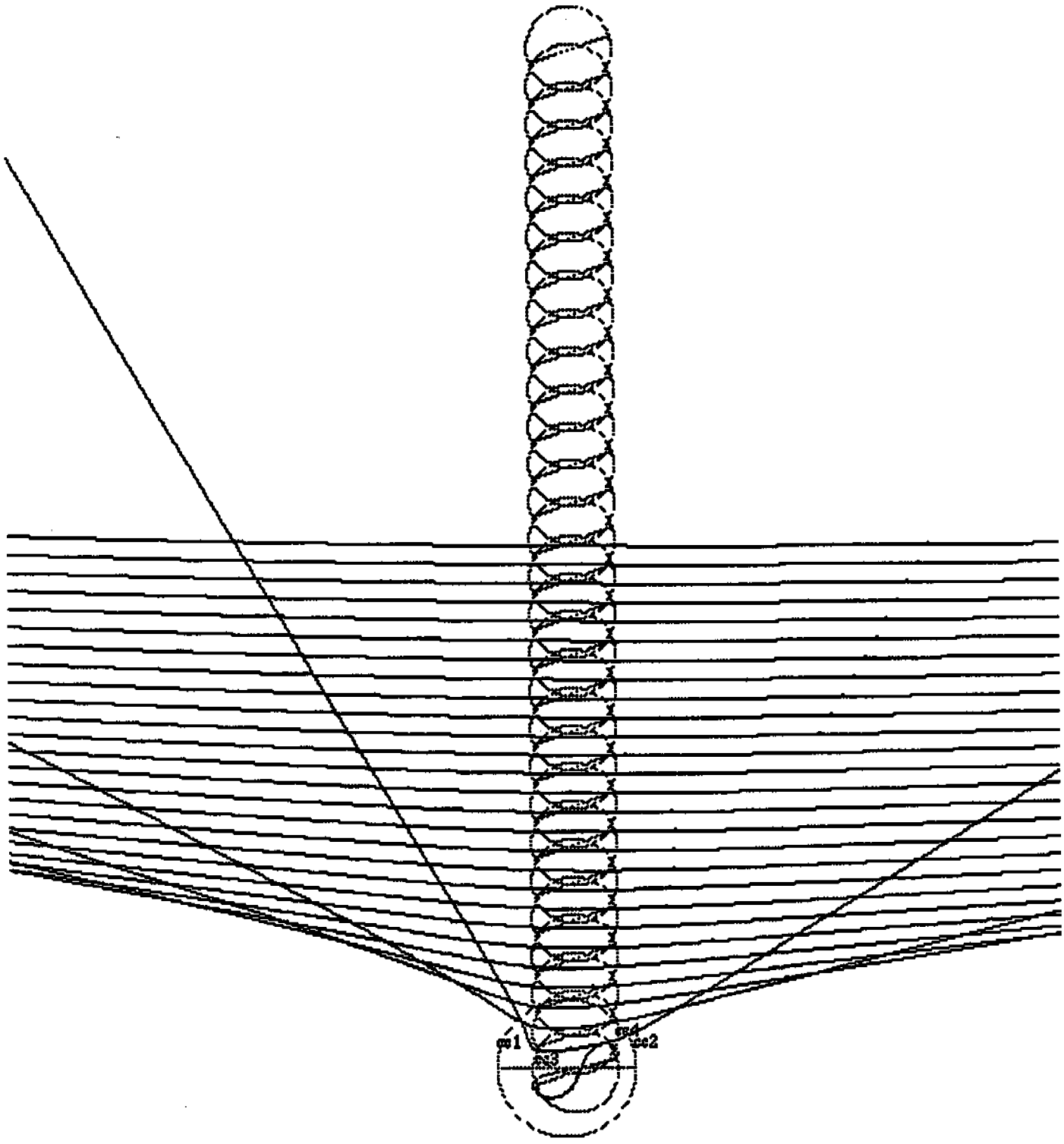
A-2



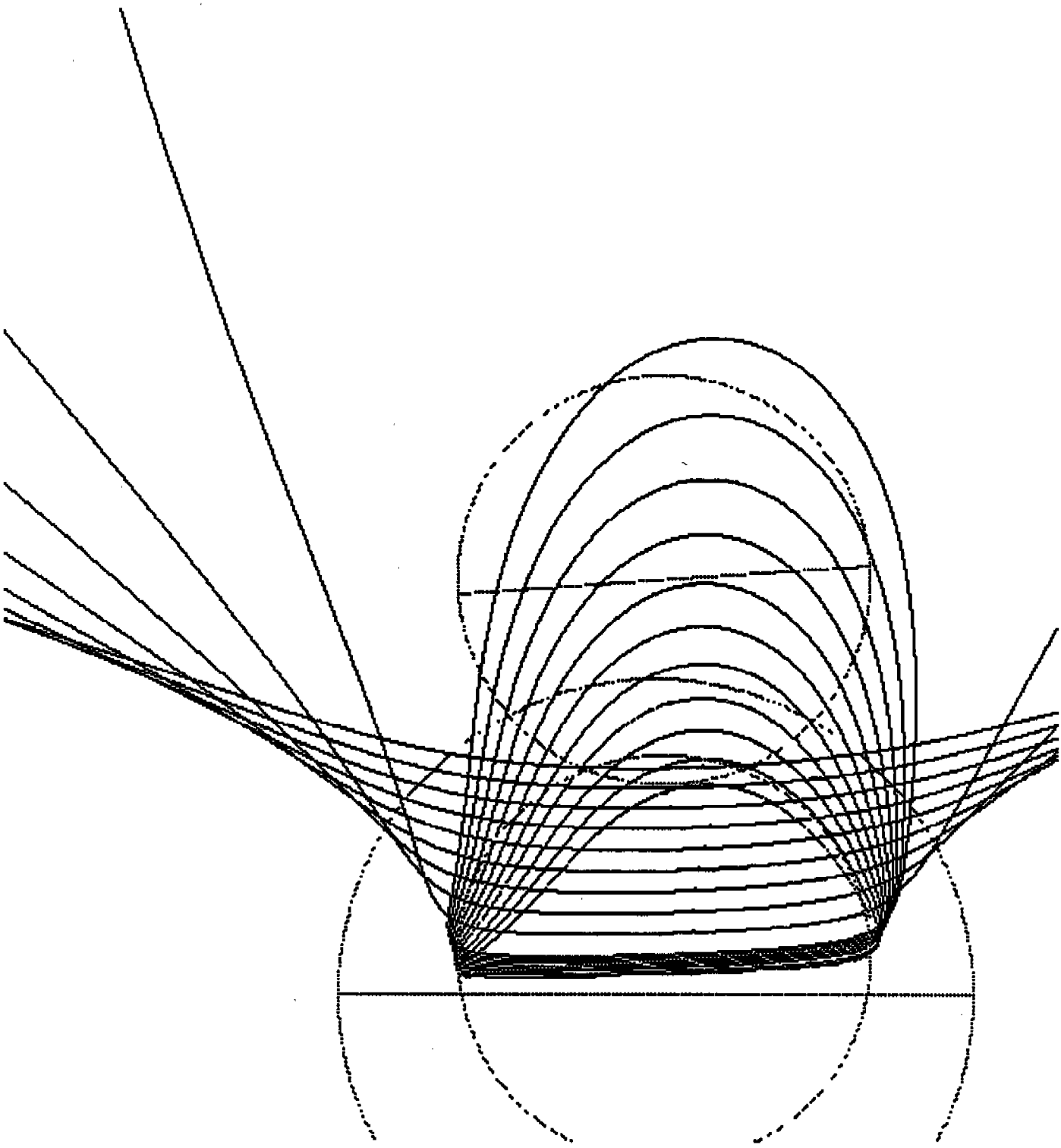
A-3



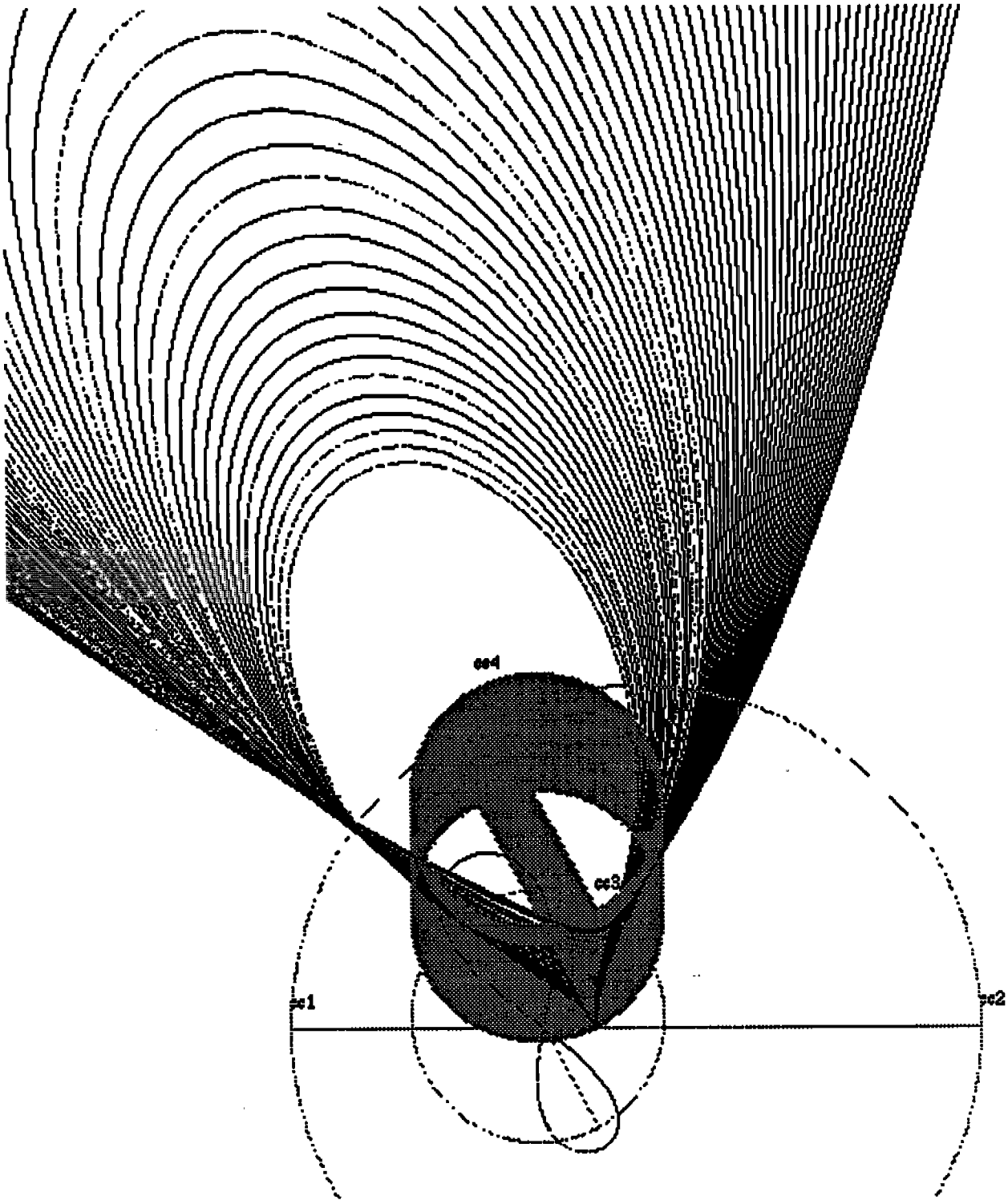
A-4



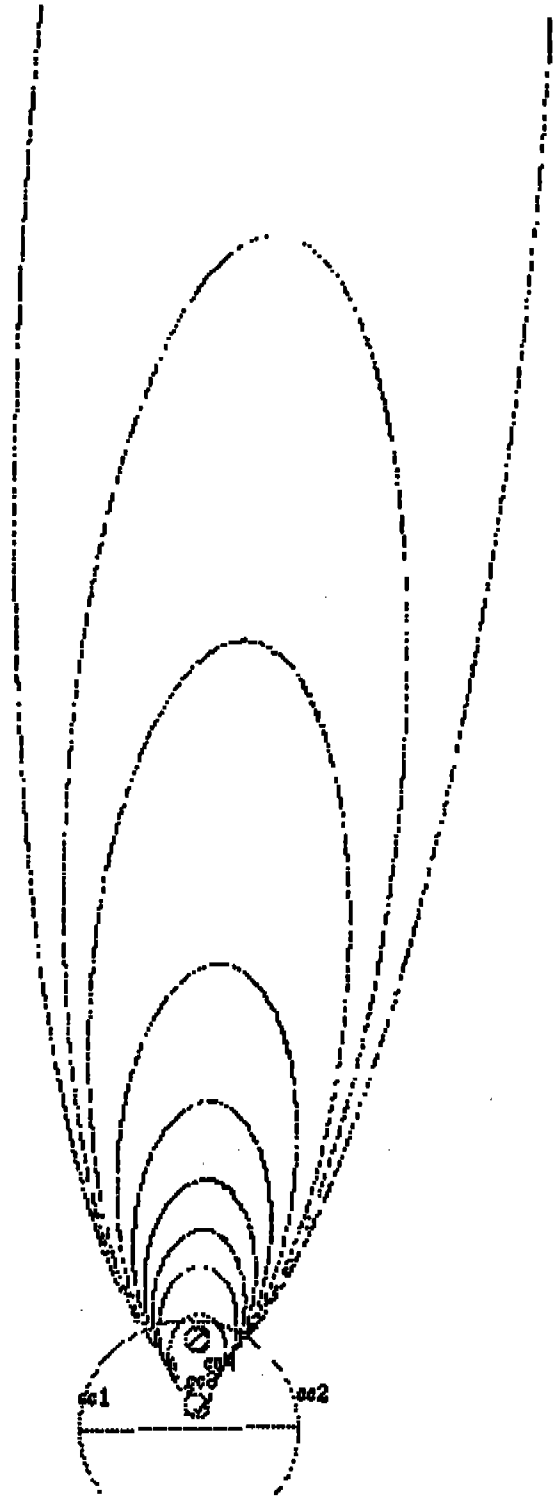
A-5



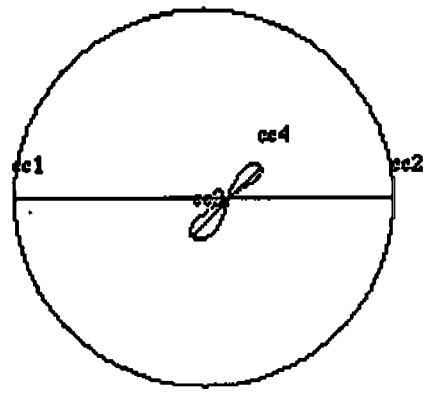
A-6



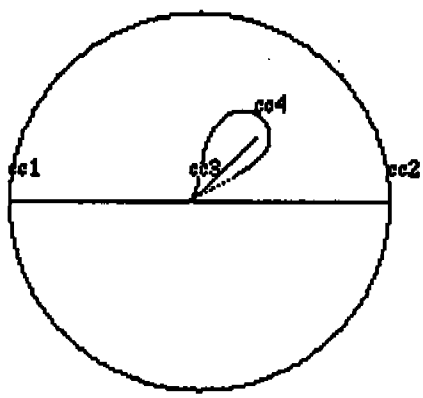
A-7



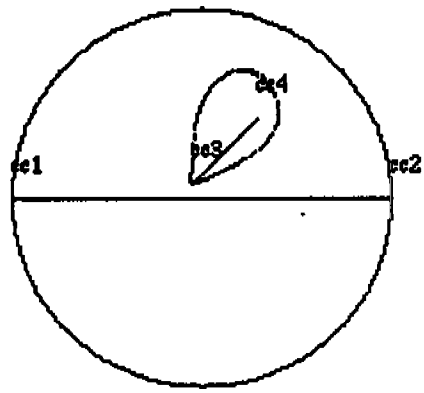
A-8



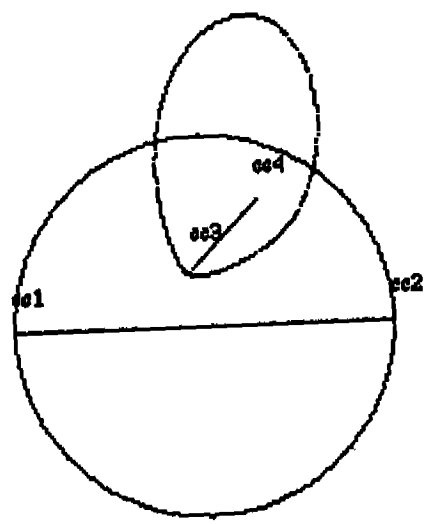
A-9a



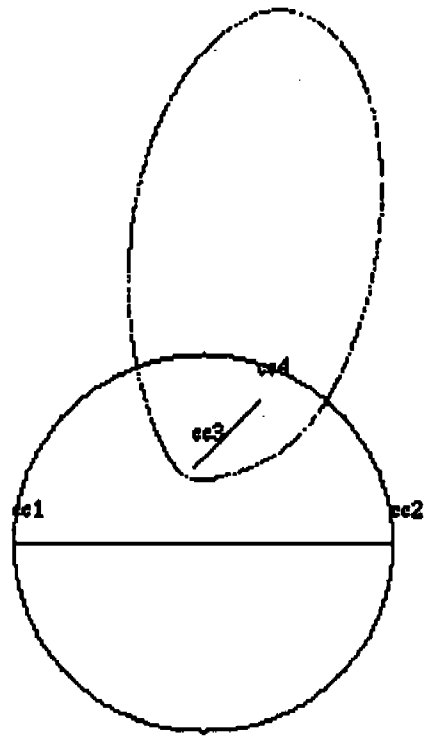
A-9b



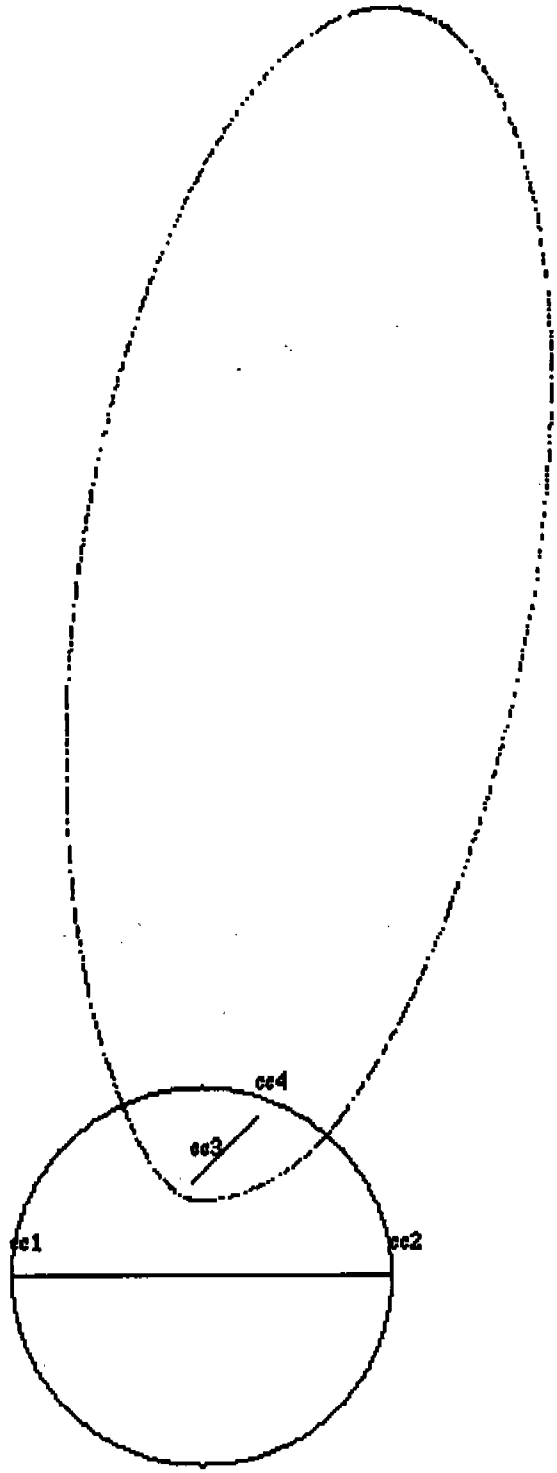
A-9c



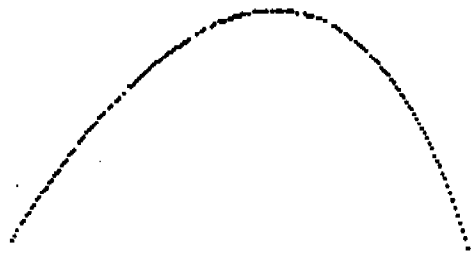
A-9d

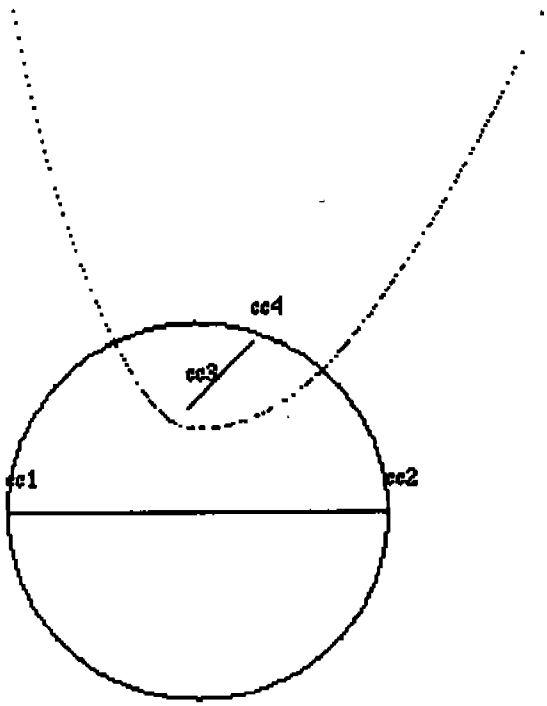


A-9e

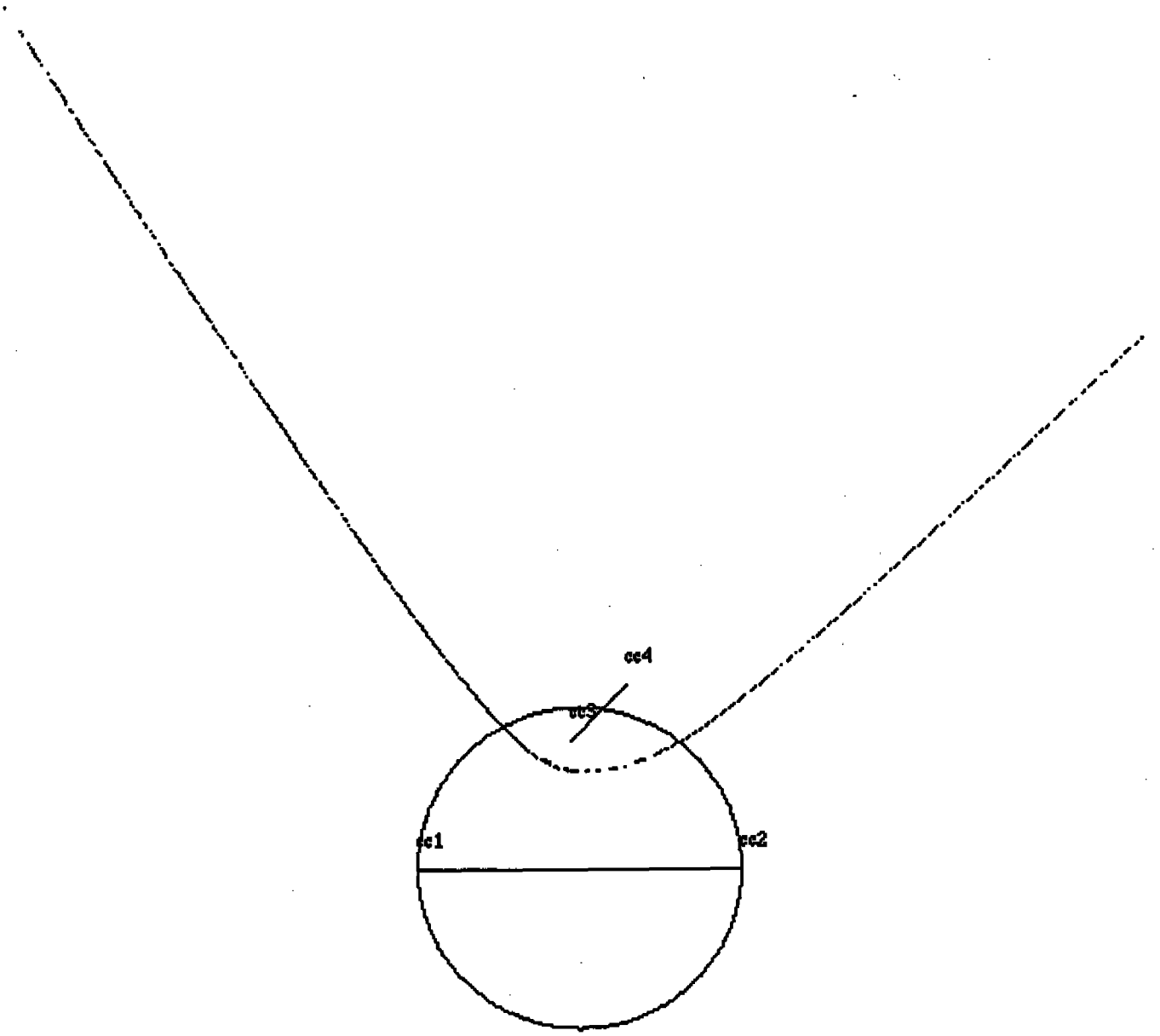


A-9f

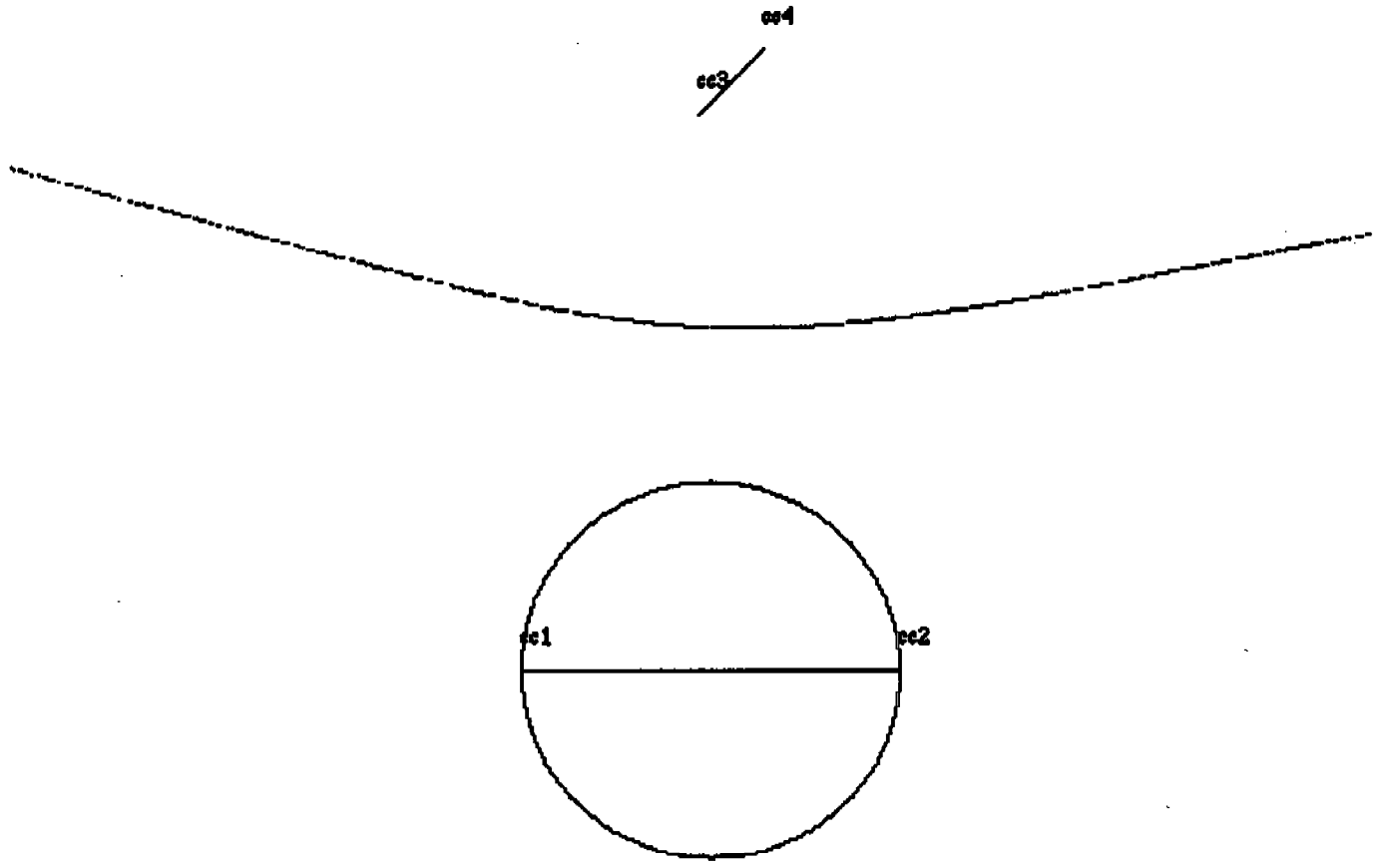




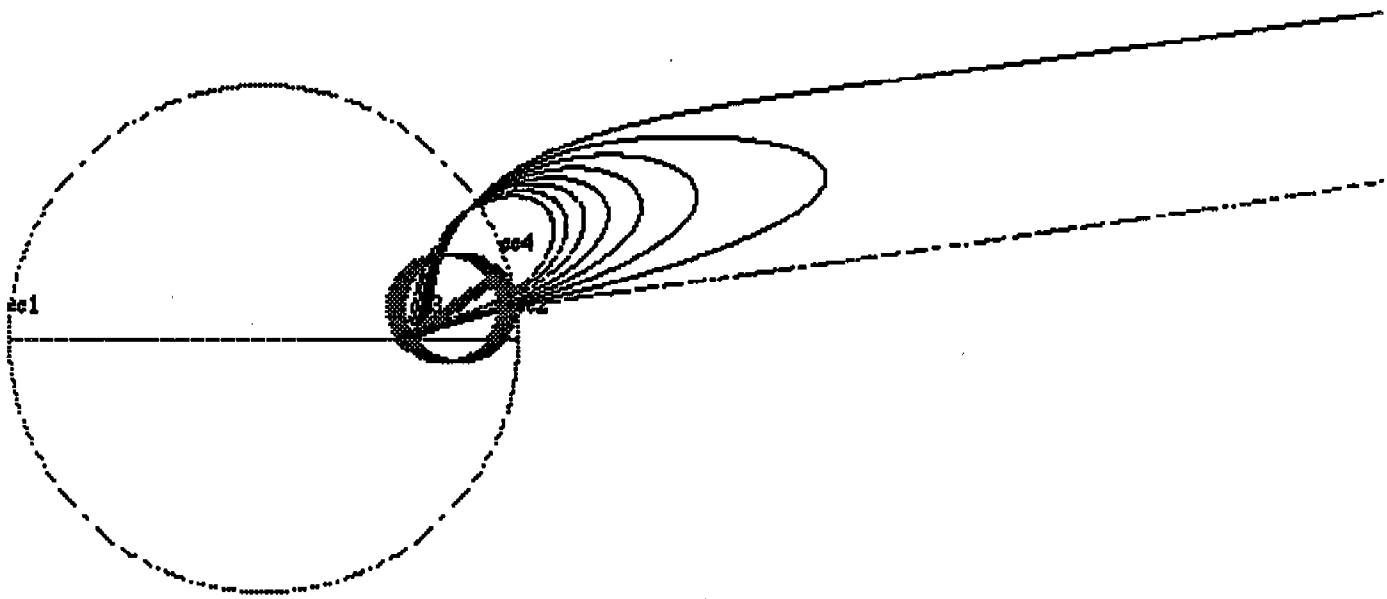
A-9g



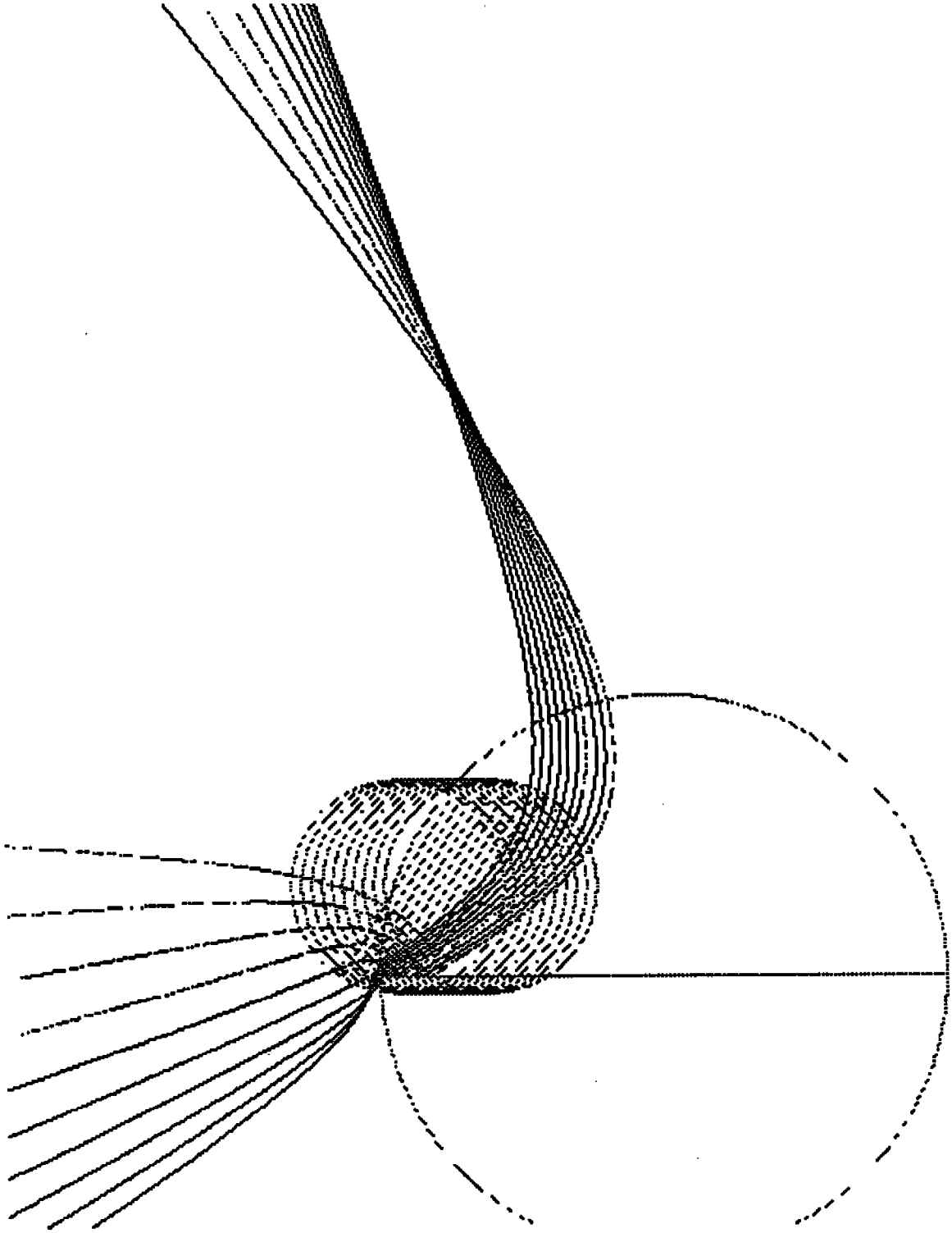
A-9h



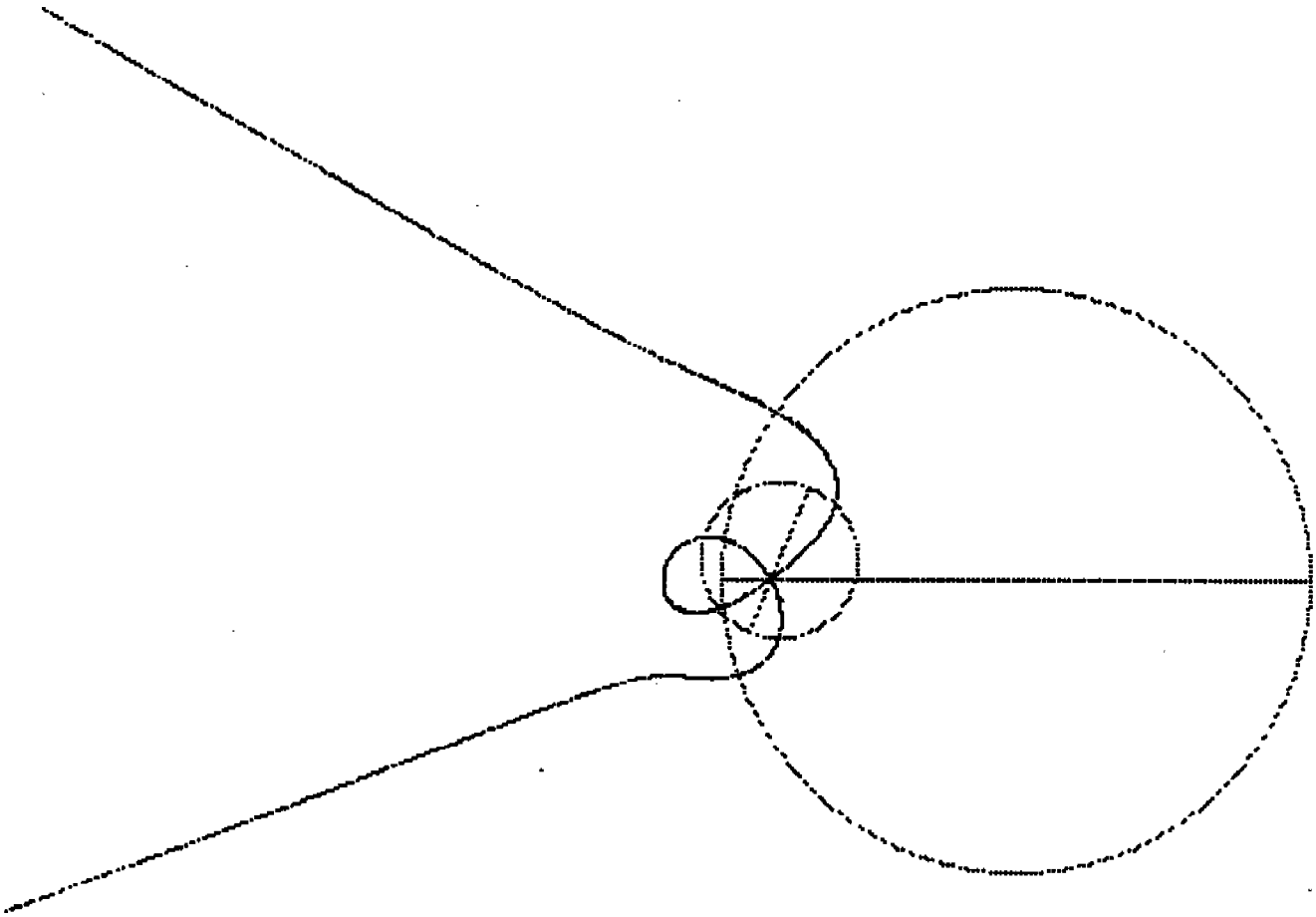
A-9i



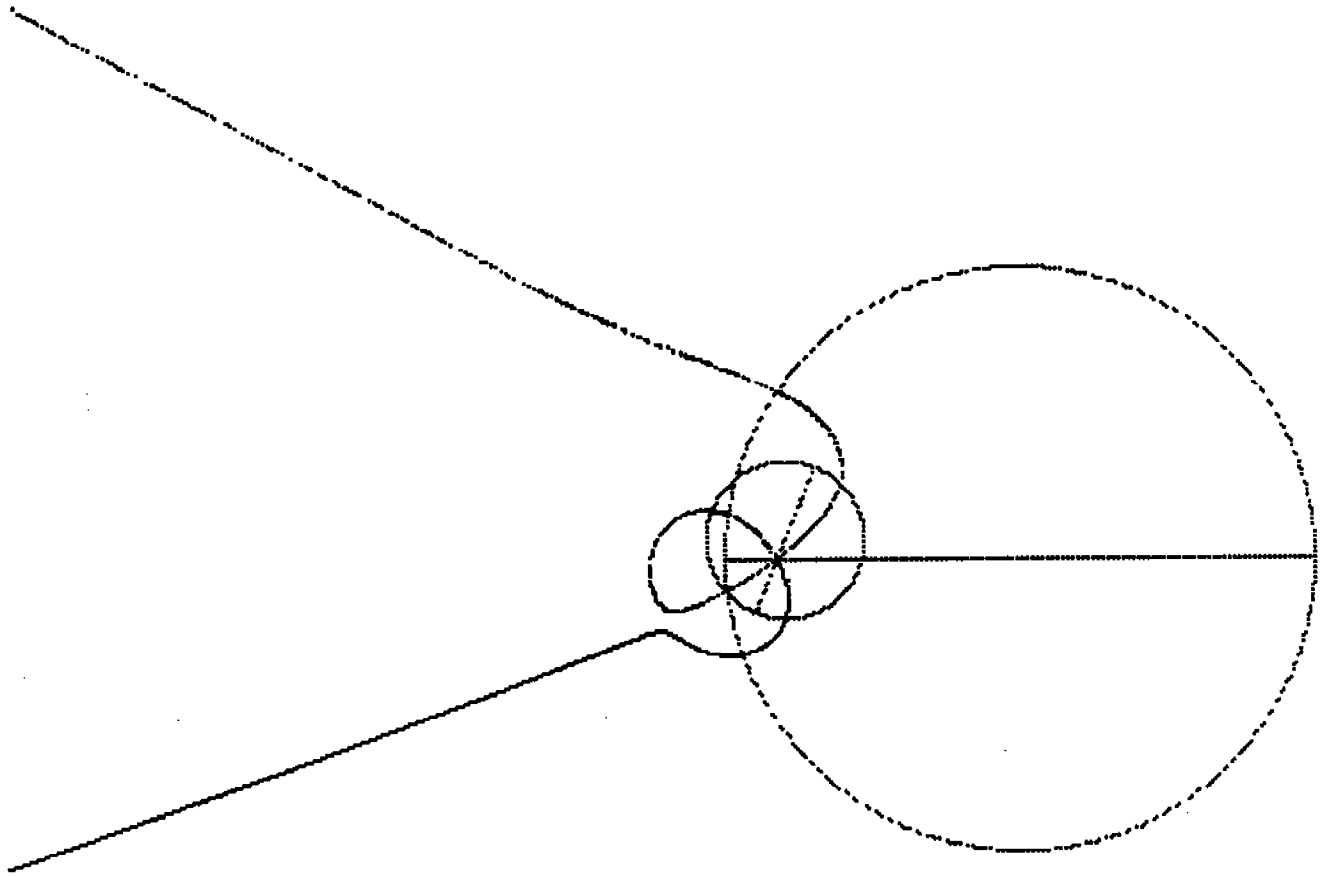
A-10



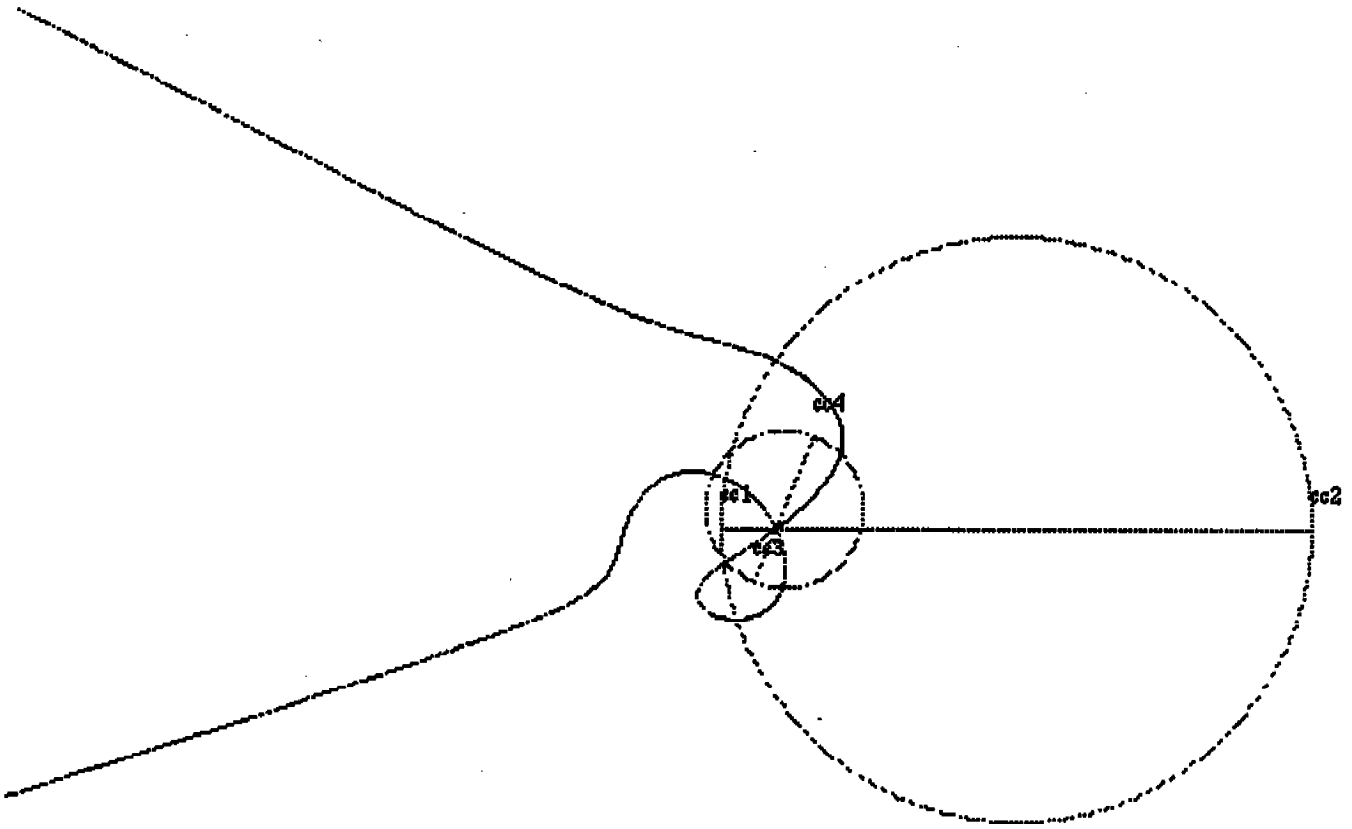
A-11



A-12a



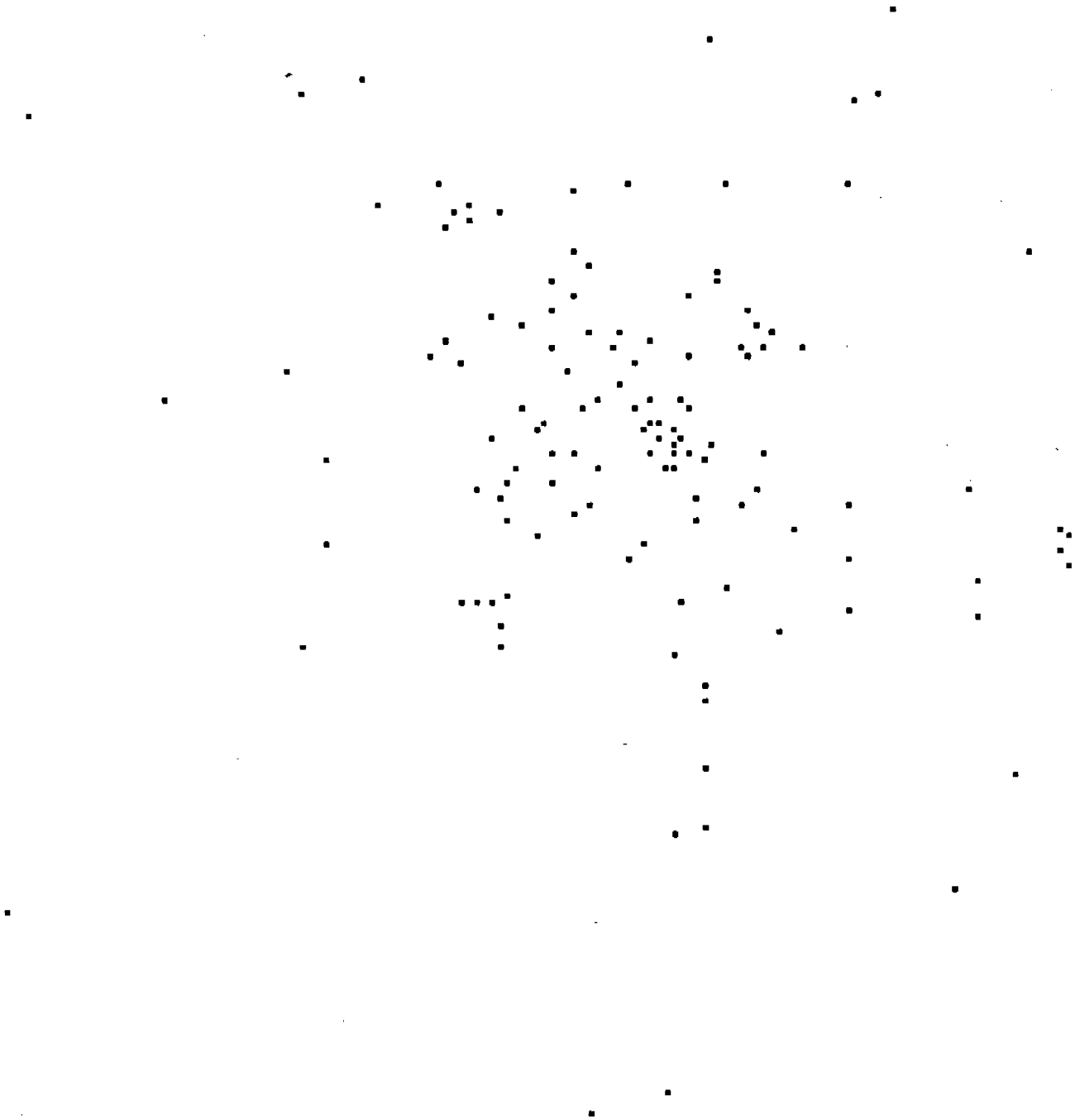
A-12b



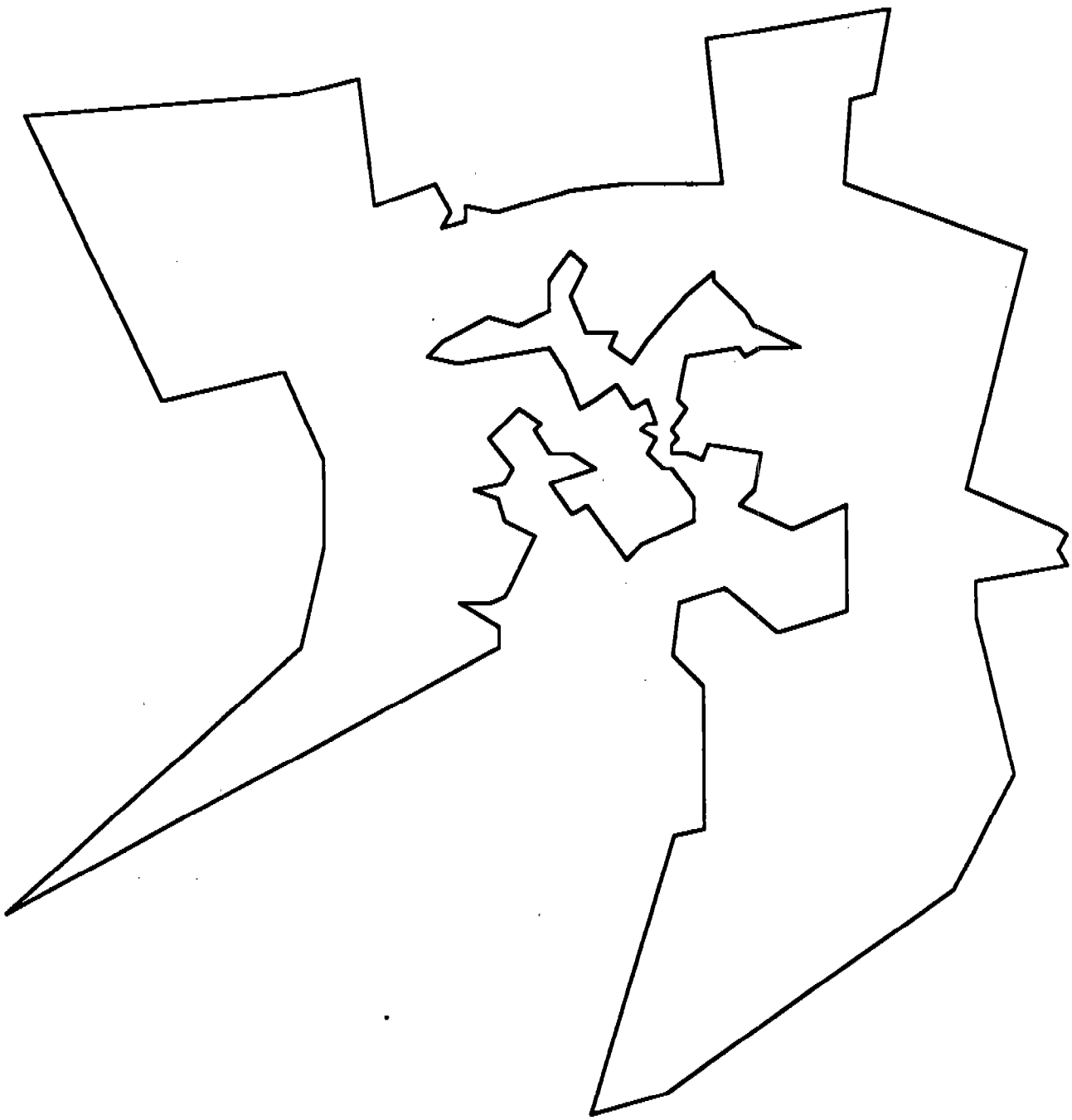
A-12c

Appendix B. The Preprocessing Algorithm Applied to Two Databases.

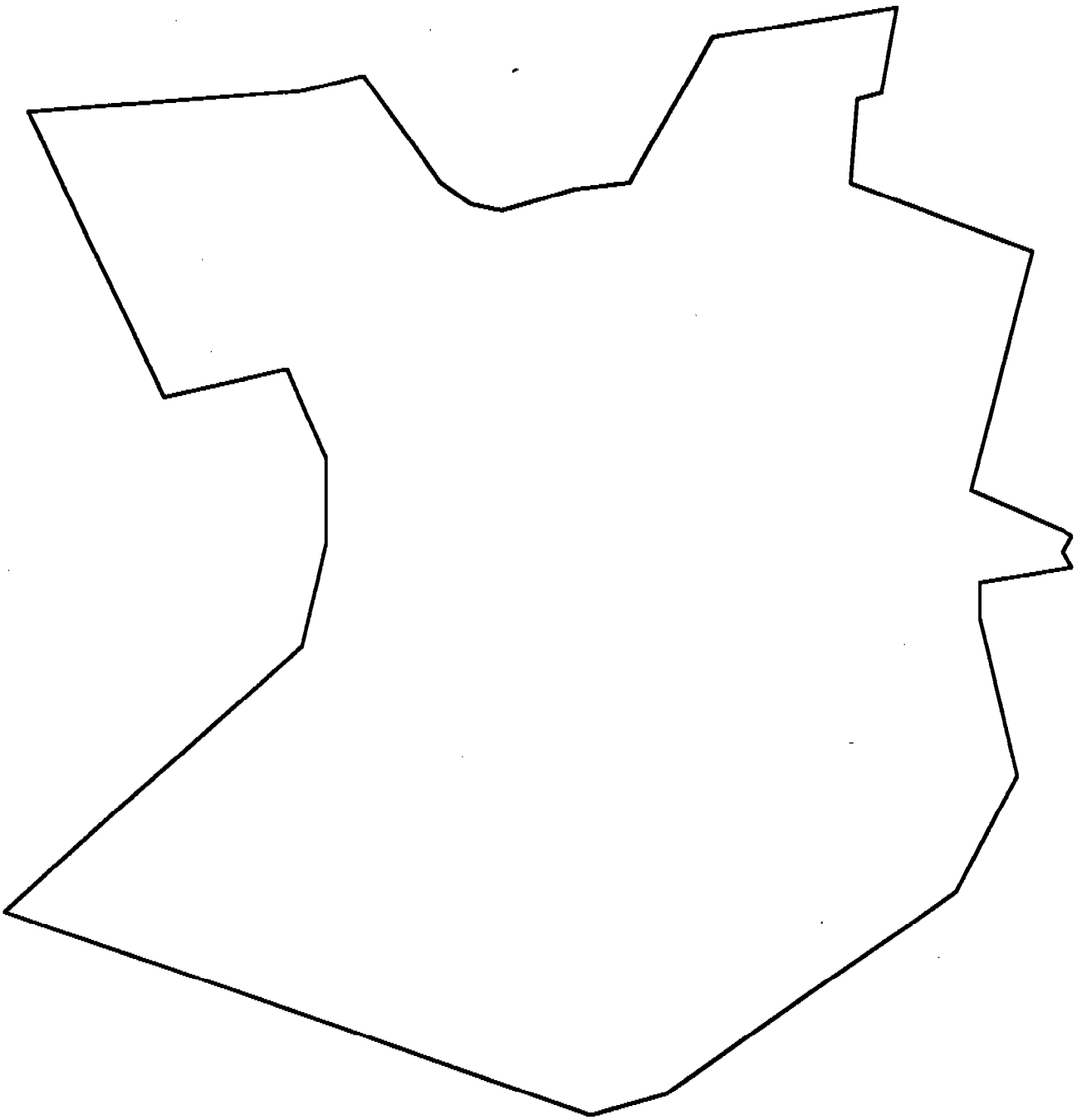
Appendix B consists of a series of seven computer graphics depicting the processing of cities for two certifiably optimal databases: the 127-city University of Augsburg dataset, and the 532-city Bell Laboratories dataset. For each of these datasets, a graphic is included to show respectively the positions of the cities, the certified optimal tour, and the optimal partial tour produced by the preprocessing algorithm described in the main body of the paper. In addition, for the 127 city dataset, the quartic Voronoi diagram is depicted for the baseline partial tour constructed by the preprocessing algorithm.



B-1

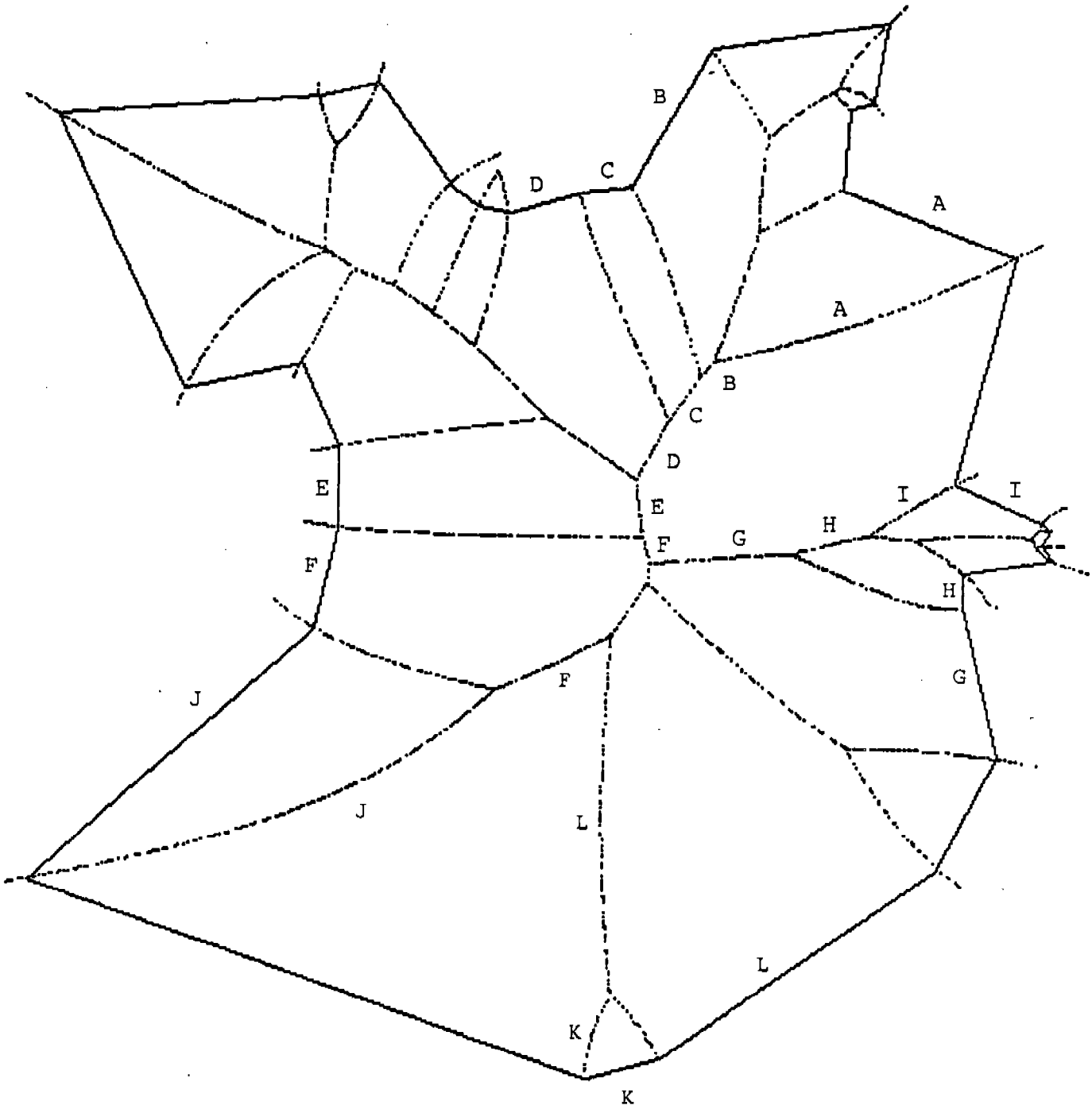


B-2

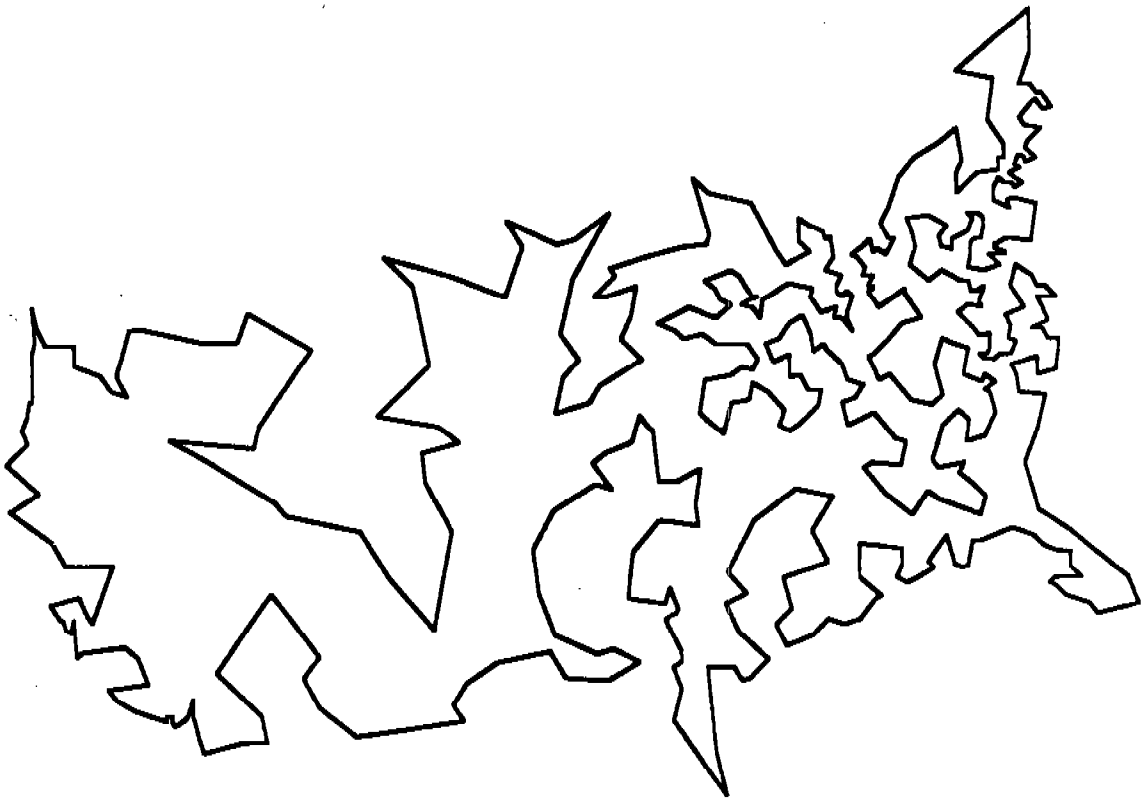


B-3

B-5



B-4



B-6

B-7

Manifold Method of Material Analysis

Gen-hua Shi
Independent Researcher,
El Cerrito, California

Abstract. The manifold method is a newly developed general method to analyze material response to external and internal changes in loads (stress). The method uses different displacement functions in different material domains. The function domains overlap each other, covering the whole material space to form a finite cover system. The large displacements of jointed or blocky materials of complex shape and moving boundaries can be computed in a mathematically consistent manner. Both the finite element method (FEM) for continua and the discontinuous deformation analysis (DDA) for block systems are special cases of the manifold method.

Mathematical Mesh and Physical Mesh of Manifolds. Physically, the shapes of material objects differ. When the material volume has fractures, blocks or different zones, the shape and boundaries become complex. Also, computations can be extremely time-consuming under conditions of large deformation and moving boundaries. The difficulty occurs because the representation via conventional analytical approximations by functions or series is feasible and useful only in a local continuous domain which represents only a small part of the material space.

Manifolds connect many individual overlapping domains together to cover the entire material volume. Then, the global behavior can be computed by functions defined in local covers. The term "manifold" in this paper is a generalization of the "differentiable manifold" which is the main subject of differential geometry and topology.

The manifold method has two independent meshes: the mathematical mesh and the physical mesh. The mathematical mesh, which is chosen by the user, consists of overlapping finite individual domains which cover the whole material space. Regular grids, finite element meshes or randomly distributed convergency regions of series can be combined to form overlapping domains of the mathematical mesh. The physical mesh includes the boundaries of the material volume, joints, blocks and the dividing lines of different zones. The physical mesh represents material conditions which can not be chosen artificially.

The mathematical mesh defines the displacement functions and the physical mesh limits the integration zones. For manifolds, the finite cover system is formed by both meshes. The finite cover system of the manifold is flexible enough to represent a wide variety of continuous or discontinuous materials located within moving boundaries.

In Figures 1 and 2, two circles and one rectangle (indicated by thin lines) delimit three domains U_1, U_2, U_3 to form the mathematical mesh. The thick lines indicate the material boundary and inner curved joints. In Figure 1, U_1 is divided by the physical mesh into two covers 1_1, 1_2 U_2 has two covers 2_1, 2_2 and U_3 has two covers 3_1, 3_2. The larger numbers refer to the mathematical domain numbers and the numerical subscripts refer to the physical zones.

Figure 2 shows a more complex mesh. Domain U_2 contains three curved lines, but only two totally disconnected covers, 2_1, 2_2, are formed. The upper curve (inside cover 2_1) does not cut through rectangle U_2 to form more covers, therefore cover 2_1 is a single cover. Similarly, since domain U_3 just intersects the end of the upper curves, covers 3_1 and 3_2 are formed. In both Figures 1 and 2, the common part of two or more covers are marked by its cover numbers.

Local Functions and Weight Functions on a Cover System. The local displacement functions are defined on individual covers which can be connected together to form a global displacement function for overlapped covers.

The local displacement functions $f_i(x,y)$ defined on cover U_i

$$f_i(x,y) \quad (x,y) \in U_i$$

can be constant, linear, high order polynomials or locally defined series. These local functions are connected together by the weight functions $w_i(x,y)$, where

$$w_i(x,y) \geq 0 \quad (x,y) \in U_i; \qquad w_i(x,y) = 0 \quad (x,y) \notin U_i;$$

$$\sum_{(x,y)\in U_j} w_j(x,y) = 1.$$

The purpose of the weight functions $w_i(x,y)$ is to take a percentage of each local function $f_i(x,y)$ for all U_i containing (x,y).

Using the weight functions $w_i(x,y)$ a global function $F(x,y)$ for the whole finite cover system is defined from the local functions

$$F(x,y) = \sum_{i=1}^{n} w_i(x,y) f_i(x,y).$$

Figure 3 is a one dimensional example: there are three domains or covers

$$U_1 = A_1A_2, \quad U_2 = B_1B_2, \quad U_3 = C_1C_2$$

$$f_1(x) = A_3A_4, \quad x \in U_1 \qquad w_1(x)f_1(x) = A_3A_5A_2, \quad x \in U_1$$

$$f_2(x) = B_3B_4, \quad x \in U_2 \qquad w_2(x)f_2(x) = B_1B_5B_4B_2, \quad x \in U_2$$

$$f_3(x) = C_3C_4, \quad x \in U_3 \qquad w_3(x)f_3(x) = C_2C_5C_4, \quad x \in U_3.$$

The global function F(x) is

$$F(x) = \sum_{i=1}^{n} w_i(x) f_i(x) = A_3A_5B_5B_4C_5C_4.$$

Displacement Functions of Manifolds are Suitable for Both Continuous and Discontinuous Deformations. For material analysis, four basically different methods are often used In order of their development, analytical solutions (AS) are the earliest, then came finite difference (FD), the finite element method (FEM), and most recently the distinct element method (DEM), and the discontinuous deformation analysis (DDA). The DEM and DDA methods are perhaps more convenient and more realistic. The convenience of the latter methods is due to the usage of more geometric information. The AS approach does not use geometry at all. The FD method uses grids with equal spaces and as such, is more general than the AS method. The FEM was a revolution, it shifted from differential equations to integral equations; from the smooth functions to the piecewise smooth functions. The generally shaped mesh of the FEM can give good results for continuous materials. The latest DEM and DDA methods are for block systems which are totally discontinuous. Displacement functions of DEM and DDA are defined for individual blocks of general shape which are completely disconnected from block to block.

The one dimensional example, represented in Figure 4, shows the relative ability of those different methods to approach a natural function (thin curves) which is discontinuous at a point. The thick smooth curve of Figure 4(a) is the approximation from the AS and FD methods. The thick piecewise

smooth segments of Figure 4(b) are the approximation from FEM. The one dimensional finite elements are defined by the line segments.

$$x_0x_1, x_1x_2, x_2x_3, x_3x_4, x_4x_5.$$

The disconnected segments of Figure 4(c) are the approximations from the DEM and DDA methods. The one dimensional blocks are $y_0x_1, y_1x_2, y_2x_3, y_3x_4, y_4x_5$ which have more unknowns than the previous methods. Figure 4(d) and Figure 5 show the approximation of the manifold method. There are seven one dimensional covers $U_1 = x_0x_1$, $U_2 = x_0x_2$, $U_3 = x_1x_3$, $U_4 = x_2x_3$, $U_5 = y_3x_4$, $U_6 = y_3x_5$, $U_7 = x_4x_5$. Since the natural function has a jump at the point $x_3 = y_3$, the cover x_2x_4 was split into two covers $U_4 = x_2x_3$ and $U_5 = y_3x_4$.

$$\begin{aligned}
w_1(x)f_1(x) &= A_0x_1, && x \in U_1\\
w_2(x)f_2(x) &= x_0A_1x_2, && x \in U_2\\
w_3(x)f_3(x) &= x_1A_2x_3, && x \in U_3\\
w_4(x)f_4(x) &= x_2A_3, && x \in U_4\\
w_5(x)f_5(x) &= B_3x_4, && x \in U_5\\
w_6(x)f_6(x) &= y_3A_4x_5, && x \in U_6\\
w_7(x)f_7(x) &= x_4A_5, && x \in U_7
\end{aligned}$$

The global function

$$F(x) = \sum_{i=1}^{n} w_i(x)f_i(x) = A_0A_1A_2A_3B_3A_4A_5$$

is very close to the original natural curve. The global displacement functions of the manifolds are capable of representing large deformations of fractured or blocky materials until the ultimate damage stage in a unified mathematical form.

Finite Cover Systems Formed by Finite Element Meshes. The manifold method can perform the computations of the finite element method (FEM) for continuous material and the discontinuous deformation analysis (DDA) for block systems in a unified algorithm.

The FEM meshes can be used to define domains or mathematical meshes for the manifold method. For any node, all finite elements having this node form

a domain (called "star" in algebraic topology). In Figures 6 and 7, the domain U_5 of node 5 has three elements 2 4 5, 2 5 3, and 3 5 6. The domain U_1 of node 1 has only one element 1 2 3 which is the only element having node 1. Any element is the common area of the domains of its nodes. For example, domain U_5 of node 5 is the area defined by the polygon 2 4 5 6 3; domain U_2 of node 2 is the area defined by the polygon 1 2 4 5 3; domain U_3 of node 3 is the area defined by the polygon 1 2 5 6 3. The common part of domains U_5, U_2, and U_3 are element 5 2 3.

The physical mesh of Figure 6 and 7 conforms to the thick lines. The covers of Figure 6 and 7 are

Covers of Figure 6			Covers of Figure 7		
U_1	1_1		U_1	1_1	
U_2	2_1	2_2	U_2	2_1	2_2
U_3	3_1	3_2	U_3	3_1	
U_4	4_1	4_2	U_4	4_1	4_2
U_5	5_1	5_2	U_5	5_1	
U_6	6_1	6_2	U_6	6_1	

Each point inside the material boundary lies in an "element" which is a common part of exactly three covers.

The following important conclusions can be proven and can also be seen directly from Figures 6 and 7:

[1] the elements are irregularly shaped;

[2] each element has three cover numbers;

[3] these three covers have one element as their common area;

[4] the three covers can be seen as three "nodes" of the element;

[5] the adjacent element has the same nodes along the common edge;

[6] two elements divided by fractures or boundaries have totally different nodes.

The "elements" and "nodes" are the extensions of their FEM counterparts. Under the new nodes and elements, the joints can open and slide, the blocks can move away and the continuous area of the material body can still be connected. The proof of these important conclusions come directly from the definition of the finite cover systems and the local and global displacement functions of the general manifold method.

For the DDA method, the material body is simply the individual blocks.

Each block is a domain, and each domain is a cover. The mathematical mesh and the physical mesh are the same where the covers are not overlapped. Therefore the DDA method is the totally discontinuous case of the manifold method.

Assuming there are n covers (or nodes) in the finite cover, the simultaneous equilibrium equations have the form:

$$\begin{pmatrix} [K_{11}] & [K_{12}] & [K_{13}] & \cdots & [K_{1n}] \\ [K_{21}] & [K_{22}] & [K_{23}] & \cdots & [K_{2n}] \\ [K_{31}] & [K_{32}] & [K_{33}] & \cdots & [K_{3n}] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ [K_{n1}] & [K_{n2}] & [K_{n3}] & \cdots & [K_{nn}] \end{pmatrix} \begin{pmatrix} \{D_1\} \\ \{D_2\} \\ \{D_3\} \\ \vdots \\ \{D_n\} \end{pmatrix} = \begin{pmatrix} \{F_1\} \\ \{F_2\} \\ \{F_3\} \\ \vdots \\ \{F_n\} \end{pmatrix} \qquad (1)$$

Because each node or cover has two degrees of freedom in a 2-d FEM manifold, each element $[K_{ij}]$ in the coefficient matrix given by equation (1) is a 2 x 2 submatrix. $\{D_i\}$ and $\{F_i\}$ are 2 x 1 submatrices.

Displacement Function. For the manifold method, the integration zones have general shapes, therefore the integrations are more difficult than the integrations of the FEM. Analytical solutions were found for many cases of the manifold method. At least all numerical integrations can be avoided for the FEM meshes within the manifold method. FEM computes the integrations of complex functions in simple domains; the manifold method computes the integrations of simple functions in complex domains.

For a triangular element, denote i_i : (x_i, y_i) the coordinates of nodes: i = 1,2,3, and the related nodal displacements as follows:

coordinates	*displacements*
i_1: (x_1, y_1)	(u_1, v_1);
i_2: (x_2, y_2)	(u_2, v_2);
i_3: (x_3, y_3)	(u_3, v_3).

The displacement field can be approximated as:

$$u = a_1 + b_1 x + c_1 y;$$
$$v = a_2 + b_2 x + c_2 y, \qquad (2)$$

$$\begin{pmatrix} a_1 \\ b_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix}^{-1} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}; \quad \begin{pmatrix} a_2 \\ b_2 \\ c_2 \end{pmatrix} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix}.$$

Denote:

$$\begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{pmatrix} = \begin{pmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{pmatrix}^{-1}, \qquad \Delta = \begin{vmatrix} 1 & x_2 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix},$$

$$(f_1 \quad f_2 \quad f_3) = (1 \quad x \quad y) \begin{pmatrix} f_{11} & f_{12} & f_{13} \\ f_{21} & f_{22} & f_{23} \\ f_{31} & f_{32} & f_{33} \end{pmatrix},$$

then

$$u = (f_1 \quad f_2 \quad f_3) \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}; \qquad v = (f_1 \quad f_2 \quad f_3) \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix},$$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} f_1 & 0 & f_2 & 0 & f_3 & 0 \\ 0 & f_1 & 0 & f_2 & 0 & f_3 \end{pmatrix} \begin{pmatrix} u_1 \\ v_1 \\ u_2 \\ v_2 \\ u_3 \\ v_3 \end{pmatrix}$$

$$= [T_e]\{D_e\},$$

$$[T_e] = ([T_1] \quad [T_2] \quad [T_3]); \qquad \{D_e\} = \begin{pmatrix} \{D_1\} \\ \{D_2\} \\ \{D_3\} \end{pmatrix}, \tag{3}$$

$$[T_i] = \begin{pmatrix} f_i & 0 \\ 0 & f_i \end{pmatrix}; \qquad \{D_i\} = \begin{pmatrix} u_i \\ v_i \end{pmatrix}, \qquad i = 1,2,3.$$

Stiffness Matrix. The relationship between stress and strain, is given by:

$$\begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} = \frac{E}{1-\nu^2} \begin{pmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1-\nu}{2} \end{pmatrix} \begin{pmatrix} \epsilon_x \\ \epsilon_y \\ \gamma_{xy} \end{pmatrix} = [E] \begin{pmatrix} \epsilon_x \\ \epsilon_y \\ \gamma_{xy} \end{pmatrix},$$

where, E, and ν are Young's modulus and Poisson's ratio respectively. Let

$$[B_i] = \begin{pmatrix} f_{2i} & 0 \\ 0 & f_{3i} \\ f_{3i} & f_{2i} \end{pmatrix}, \qquad i = 1,2,3.$$

Then

$$\begin{pmatrix} \epsilon_x \\ \epsilon_y \\ \tau_{xy} \end{pmatrix} = [B_e]\{D_e\} = ([B_1] \quad [B_2] \quad [B_3]) \begin{pmatrix} \{D_1\} \\ \{D_2\} \\ \{D_3\} \end{pmatrix}. \tag{4}$$

The strain energy Π_e, caused by the elastic stresses of element e, is:

$$\Pi_e = \iint_A \frac{1}{2} (\epsilon_x \quad \epsilon_y \quad \gamma_{xy}) \begin{pmatrix} \sigma_x \\ \sigma_y \\ \tau_{xy} \end{pmatrix} dx\,dy \tag{5}$$

$$= \frac{1}{2}\{D_e\}^T \left(S^e [B_e]^T [E] [B_e] \right) \{D_e\},$$

where S^e is the area of that element. Then

$$S^e[B_r]^T[E][B_s] \quad \rightarrow \quad [K_{i(r)i(s)}]; \qquad r,s = 1,2,3; \tag{6}$$

$$i(\ell) = \begin{cases} i_1, & \ell = 1; \quad \text{first node} \\ i_2, & \ell = 2; \quad \text{second node} \\ i_3, & \ell = 3; \quad \text{third node.} \end{cases}$$

Initial Stresses Matrix. For the element e, the potential energy of the initial constant stresses $\{\sigma_e^0\} = (\sigma_x^0 \quad \sigma_y^0 \quad \tau_{xy}^0)^T$ is

$$\begin{aligned} \Pi_\sigma &= \iint_A (\epsilon_x \quad \epsilon_y \quad \gamma_{xy}) \begin{pmatrix} \sigma_x^0 \\ \sigma_y^0 \\ \tau_{xy}^0 \end{pmatrix} dx\,dy \\ &= S^e\{D_e\}^T[B_e]^T\{\sigma_e^0\}. \end{aligned} \tag{7}$$

where S^e is the area of that element. Then

$$-S^e[B_r]^T \begin{pmatrix} \sigma_x^0 \\ \sigma_y^0 \\ \tau_{xy}^0 \end{pmatrix} \quad \rightarrow \quad \{F_{i(r)}\}; \qquad r = 1,2,3. \tag{8}$$

Point Loading Matrix. The point loading force $(F_x \quad F_y)^T$ acts on point (x,y) of element e. And

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} u(x,y) \\ v(x,y) \end{pmatrix}.$$

The potential energy due to the point loading is

$$\Pi_p = -(u \quad v) \begin{pmatrix} F_x \\ F_y \end{pmatrix} = -\{D_e\}^T[T_e(x,y)]^T \begin{pmatrix} F_x \\ F_y \end{pmatrix}. \tag{9}$$

Then

$$[T_r]^T \begin{pmatrix} F_x \\ F_y \end{pmatrix} \quad \rightarrow \quad \{F_{i(r)}\}; \qquad r = 1,2,3. \tag{10}$$

Body Loading Matrix. Assuming that $(f_x \quad f_y)^T$ is the constant body force loading acting on the volume of element e. The potential energy due to the body loading is

$$\Pi_w = -\iint_A (u \quad v) \begin{pmatrix} f_x \\ f_y \end{pmatrix} dx\,dy = -\{D_e\}^T \left[\iint_A [T_e]^T dx\,dy\right] \begin{pmatrix} f_x \\ f_y \end{pmatrix},$$

$$\iint_{\mathcal{A}} [T_e]^T \, dx\, dy = \begin{pmatrix} S_i & 0 \\ 0 & S_i \end{pmatrix},$$
$$S_i = f_{1i} S^e + f_{2i} S_x^e + f_{3i} S_y^e, \tag{11}$$

where

$$S^e = \iint_{\mathcal{A}} dx\, dy; \quad S_x^e = \iint_{\mathcal{A}} x\, dx\, dy; \quad S_y^e = \iint_{\mathcal{A}} y\, dx\, dy.$$

Then

$$\begin{pmatrix} S_r & 0 \\ 0 & S_r \end{pmatrix} \begin{pmatrix} f_x \\ f_y \end{pmatrix} \quad \rightarrow \quad \{F_{i(r)}\}; \qquad i = 1, 2, 3. \tag{12}$$

Fixed Point Matrix. As a boundary condition, some of the elements are fixed at specific points. This constraint can be applied by two very stiff springs. Assume the fixed point is (x,y) at element e and the stiffness of the springs is p. The spring forces are

$$\begin{pmatrix} f_x \\ f_y \end{pmatrix} = \begin{pmatrix} -pu(x,\ y) \\ -pv(x,\ y) \end{pmatrix}.$$

The strain energy of the spring is Π_m, then

$$\Pi_m = \frac{p}{2} \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \frac{p}{2} \{D_e\}^T [T_e]^T [T_e] \{D_e\}. \tag{13}$$

Therefore,

$$p f_r f_s \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \rightarrow \quad [K_{i(r)i(s)}]; \qquad r, s = 1, 2, 3. \tag{14}$$

Forces of Inertia Matrix. Denote $(u(t) \quad v(t))^T$ as the time dependent displacements of any point (x,y) of element e and M as the mass per unit area. The potential energy of the inertia force of element e is

$$\Pi_i = -\iint_{\mathcal{A}} \begin{pmatrix} u & v \end{pmatrix} \begin{pmatrix} f_x \\ f_y \end{pmatrix} dx\, dy = \iint_{\mathcal{A}} \mathcal{M} \begin{pmatrix} u & v \end{pmatrix} [T_e] \frac{\partial^2 \{D_e(t)\}}{\partial t^2} \, dx\, dy.$$

Assume $\{D_e(0)\} = \{0\}$ as the element displacements at the beginning of the time step, $\{D_e(\Delta)\} = \{D_e\}$ as the displacements at the end of the time step, and Δ as the time interval of this time step. Then

$$\{D_e\} = \{D_e(\Delta)\} = \Delta \frac{\partial \{D_e(0)\}}{\partial t} + \frac{\Delta^2}{2} \frac{\partial^2 \{D_e(0)\}}{\partial t^2},$$
$$\frac{\partial^2 \{D_e(0)\}}{\partial t^2} = \frac{2}{\Delta^2} \{D_e\} - \frac{2}{\Delta} \frac{\partial \{D_e(0)\}}{\partial t} = \frac{2}{\Delta^2} \{D_e\} - \frac{2}{\Delta} \{V_e(0)\}, \tag{15}$$

where

$$\{V_e(0)\} = \frac{\partial\{D_e(0)\}}{\partial t},$$

is the velocity at the beginning of the time step. The potential energy then becomes:

$$\Pi_i = \mathcal{M}\{D_e\}^T \left[\iint_{\mathcal{A}} [T_e]^T [T_e]\, dx\, dy\right] \left(\frac{2}{\Delta^2}\{D_e\} - \frac{2}{\Delta}\{V_e(0)\}\right). \tag{16}$$

Then

$$\frac{2\mathcal{M}}{\Delta^2}\left[\iint_{\mathcal{A}} [T_r]^T [T_s]\, dx\, dy\right] \quad \rightarrow \quad [K_{i(r)i(s)}]; \qquad r, s = 1, 2, 3. \tag{17}$$

$$\frac{2\mathcal{M}}{\Delta}\left[\iint_{\mathcal{A}} [T_r]^T [T_s]\, dx\, dy\right]\{V_s(0)\} \quad \rightarrow \quad \{F_{i(r)}\}; \qquad \begin{cases} \text{r} & = 1,2,3; \\ \text{s} & = \text{tensor sum}, \end{cases} \tag{18}$$

where

$$\{V_s(0)\} = \frac{\partial}{\partial t}\begin{pmatrix} u_s(0) \\ v_s(0) \end{pmatrix}.$$

In the following we compute

$$\iint_{\mathcal{A}} [T_r]^T [T_s]\, dx\, dy$$
$$= \iint_{\mathcal{A}} \begin{pmatrix} f_r f_s & 0 \\ 0 & f_r f_s \end{pmatrix} dx\, dy = \begin{pmatrix} t_{rs} & 0 \\ 0 & t_{rs} \end{pmatrix};$$

where

$$t_{rs} = \iint_{\mathcal{A}} f_r f_s\, dx\, dy$$
$$= \iint_{\mathcal{A}} (f_{1r} + f_{2r}x + f_{3r}y)(f_{1s} + f_{2s}x + f_{3s}y)\, dx\, dy,$$

and

$$t_{rs} = f_{1r}f_{1s}S^e + (f_{1r}f_{2s} + f_{1s}f_{2r})S^e_x + (f_{1r}f_{3s} + f_{1s}f_{3r})S^e_y + f_{2r}f_{2s}S^e_{xx} + f_{3r}f_{3s}S^e_{yy} + (f_{2r}f_{3s} + f_{2s}f_{3r})S^e_{xy}, \tag{19}$$

where

$$S^e_{xy} = \iint_{\mathcal{A}} xy\, dx\, dy, \qquad S^e_{xx} = \iint_{\mathcal{A}} x^2\, dx\, dy, \qquad S^e_{yy} = \iint_{\mathcal{A}} y^2\, dx\, dy.$$

As this element of the manifold method is a generally shaped polygon, then

$$P_1\ P_2\ \ldots\ P_{m-1}\ P_m\ P_1, \quad P_i = (x_i, y_i)$$

are its ordered vertices rotating from axis x to axis y. Denoting $P_o = (x_o, y_o)$ as the arbitrary chosen point, the analytical solutions of these integrations are the following:

$$
\begin{aligned}
S^e &= \frac{1}{2}\sum_{i=1}^{m}\begin{vmatrix} 1 & x_0 & y_0 \\ 1 & x_i & y_i \\ 1 & x_{i+1} & y_{i+1} \end{vmatrix};\\
S_x^e &= \frac{S^e}{3}\sum_{i=1}^{m}(x_0 + x_i + x_{i+1});\\
S_y^e &= \frac{S^e}{3}\sum_{i=1}^{m}(y_0 + y_i + y_{i+1});\\
S_{xx}^e &= \frac{S^e}{6}\sum_{i=1}^{m}(x_0^2 + x_i^2 + x_{i+1}^2 + x_i x_0 + x_{i+1}x_0 + x_i x_{i+1});\\
S_{yy}^e &= \frac{S^e}{6}\sum_{i=1}^{m}(y_0^2 + y_i^2 + y_{i+1}^2 + y_i y_0 + y_{i+1}y_0 + y_i y_{i+1});\\
S_{xy}^e &= \frac{S^e}{12}\sum_{i=1}^{m}(2x_0y_0 + 2x_iy_i + 2x_{i+1}y_{i+1}\\
&\quad + x_iy_0 + x_{i+1}y_0 + x_0y_i + x_0y_{i+1} + x_iy_{i+1} + x_{i+1}y_i);
\end{aligned}
\tag{20}
$$

Normal Contact Matrix. Assume P_1 is a vertex; P_2P_3 is the reference line and (x_i, y_i) and (u_i, v_i) are the coordinates and displacement of P_i, $i = 1,2,3$ respectively. If points P_1, P_2, and P_3 rotate in the same sense as the rotation of ox to oy (Figure 8), then the distance d from P_1 to line P_2P_3 is:

$$
\begin{aligned}
d = \frac{\Delta}{l} &= \frac{1}{l}\begin{vmatrix} 1 & x_1 + u_1 & y_1 + v_1 \\ 1 & x_2 + u_2 & y_2 + v_2 \\ 1 & x_3 + u_3 & y_3 + v_3 \end{vmatrix};\\
l &= \sqrt{(x_2 - x_3)^2 + (y_2 - y_3)^2}.
\end{aligned}
\tag{21}
$$

If P_1 passed edge P_2P_3, d will be zero. Let

$$
S_0 = \begin{vmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{vmatrix},
$$

and we have

$$
\Delta \approx S_0 + \begin{vmatrix} 1 & u_1 & y_1 \\ 1 & u_2 & y_2 \\ 1 & u_3 & y_3 \end{vmatrix} + \begin{vmatrix} 1 & x_1 & v_1 \\ 1 & x_2 & v_2 \\ 1 & x_3 & v_3 \end{vmatrix}. \tag{22}
$$

Let

$$\{H\} = \begin{pmatrix} \{H_1\} \\ \{H_2\} \\ \{H_3\} \end{pmatrix} = \frac{1}{l} \begin{pmatrix} [T_1]^T(x_1, y_1) \\ [T_2]^T(x_1, y_1) \\ [T_3]^T(x_1, y_1) \end{pmatrix} \begin{pmatrix} y_2 - y_3 \\ x_3 - x_2 \end{pmatrix};$$

$$\{G\} = \begin{pmatrix} \{G_1\} \\ \{G_2\} \\ \{G_3\} \end{pmatrix} = \frac{1}{l} \begin{pmatrix} [T_1]^T(x_2, y_2) \\ [T_2]^T(x_2, y_2) \\ [T_3]^T(x_2, y_2) \end{pmatrix}^T \begin{pmatrix} y_3 - y_1 \\ x_1 - x_3 \end{pmatrix}$$

$$+ \frac{1}{l} \begin{pmatrix} [T_1]^T(x_3, y_3) \\ [T_2]^T(x_3, y_3) \\ [T_3]^T(x_3, y_3) \end{pmatrix}^T \begin{pmatrix} y_1 - y_2 \\ x_2 - x_1 \end{pmatrix}.$$

Then (21) becomes

$$d = \{H\}^T\{D_i\} + \{G\}^T\{D_j\} + \frac{S_0}{l}. \tag{23}$$

The potential energy of the normal spring is:

$$\begin{aligned} \Pi_k &= \frac{p}{2} d^2 \\ &= \frac{p}{2} \left(\{H\}^T\{D_i\} + \{G\}^T\{D_j\} + \frac{S_0}{l} \right)^2 \\ &= \frac{p}{2} \Big[\{D_i\}^T\{H\}\{H\}^T\{D_i\} + \{D_j\}^T\{G\}\{G\}^T\{D_j\} + 2\{D_i\}^T\{H\}\{G\}^T\{D_j\} \\ &+ 2\left(\frac{S_0}{l}\right)\{D_i\}^T\{H\} + 2\left(\frac{S_0}{l}\right)\{D_j\}^T\{G\} + \left(\frac{S_0}{l}\right)^2 \Big]. \end{aligned} \tag{24}$$

Thus,

$$\begin{aligned} p\{H_r\}\{H_s\}^T &\rightarrow [K_{i(r)i(s)}]; \quad r, s = 1, 2, 3; \\ p\{H_r\}\{G_s\}^T &\rightarrow [K_{i(r)j(s)}]; \quad r, s = 1, 2, 3; \\ p\{G_r\}\{H_s\}^T &\rightarrow [K_{j(r)i(s)}]; \quad r, s = 1, 2, 3; \\ p\{G_r\}\{G_s\}^T &\rightarrow [K_{j(r)j(s)}]; \quad r, s = 1, 2, 3; \\ -p\left(\frac{S_0}{l}\right)\{H_r\} &\rightarrow \{F_{i(r)}\}; \quad r = 1, 2, 3; \\ -p\left(\frac{S_0}{l}\right)\{G_r\} &\rightarrow \{F_{j(r)}\}; \quad r = 1, 2, 3; \end{aligned} \tag{25}$$

where

for element i $\begin{cases} i(1) = i_1; \\ i(2) = i_2; \\ i(3) = i_3; \end{cases}$ for element j $\begin{cases} j(1) = j_1; \\ j(2) = j_2; \\ j(3) = j_3. \end{cases}$

Applications. The manifold method has been applied to a variety of important engineering problems. For example:

Figure 9 shows the ability of the manifold method to compute the deformation of a joint or fracture within a material.

Figure 10 shows the failure of an arch under the influence of a point load on the center and self weight.

Figure 11 is a simulation of sliding of rock blocks. Notice that the center sliding block separated two adjacent blocks during the sliding. This result is consistent with laboratory tests.

Figure 12 shows a soil slope which slides along a circular surface. The sliding computation satisfies all equilibrium conditions.

Figure 13 shows the failure of a gravity dam with a rock foundation. The loads are the upstream water pressure and the self weight of the dam.

Conclusions. This new theory, entitled the Manifold Method of Material Analysis, incorporates a multitude of simultaneous physical meshes (manifolds) which overlay the mathematical mesh. These (coupled) physical meshes provide the means to consider both jointed and continuous materials, and even different material phases (i.e. solid, gas, or liquid). At present, a fairly robust theory for the manifold method has been accomplished, as has a first generation 2-D dynamic computer code. The preliminary results are extremely encouraging (for example, the convergence of solutions has been established). Finite element and DDA formulations are special cases of this developing theory. A brief listing of a few of the advantages of the manifold method follows:

- Free surfaces and flexible boundaries
- Analysis not hindered by boundary conditions
- Free form elements (any shape)
- Conservation of energy
- Obeys Coulomb's Law
- Very, very small to very, very large deformations
- Statics and dynamics possible
- Analytically correct

- Continuum/discontinuum analysis

Acknowledgements. The development of the theory presented herein is the result of research for Work Unit No. 31700, "Special Studies for Civil Works Rock Problems," of the Civil Works-Materials /Rock Research and Development Program and Work Unit No. 32648, "Geomechanical Modelling for Stability of Gravity Structures," of the Repair, Evaluation, Maintenance and Rehabilitation Research Program sponsored by the US Army Corps of Engineers (USACE). The performing agency was the U.S. Army Engineer Waterways Experiment Station (USAEWES). The Chief of Engineers granted permission to publish this information. However, this study is ongoing and no policy or recommendations have been made: opinions, findings, and conclusions expressed are those of the author and do not necessarily reflect the views of the USACE or the USAEWES.

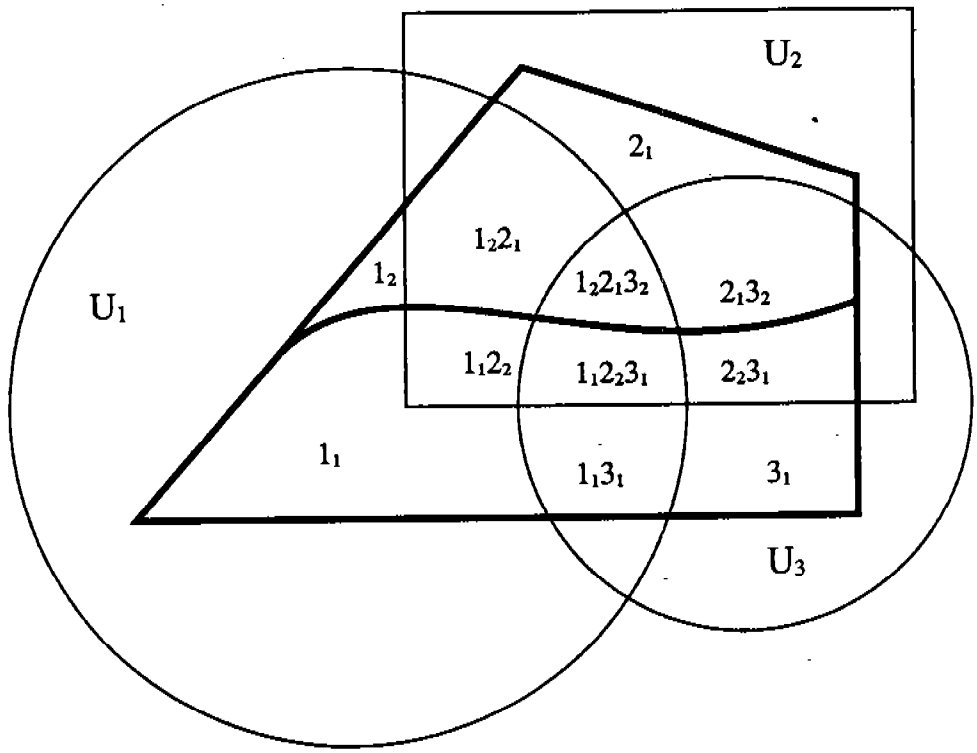


Figure 1.

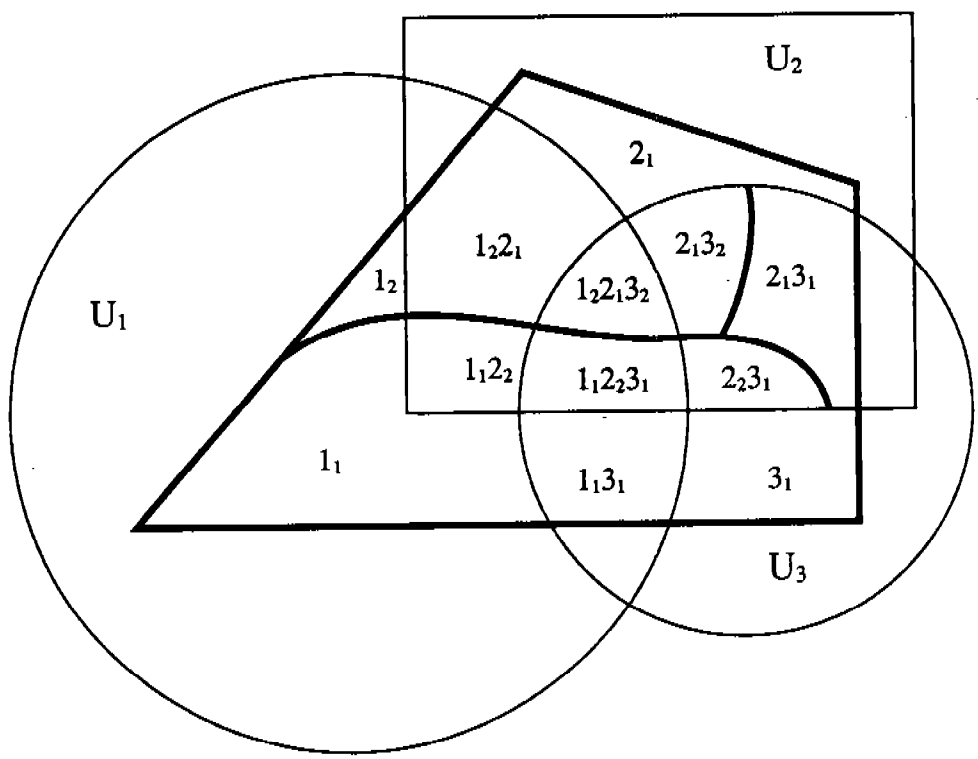


Figure 2.

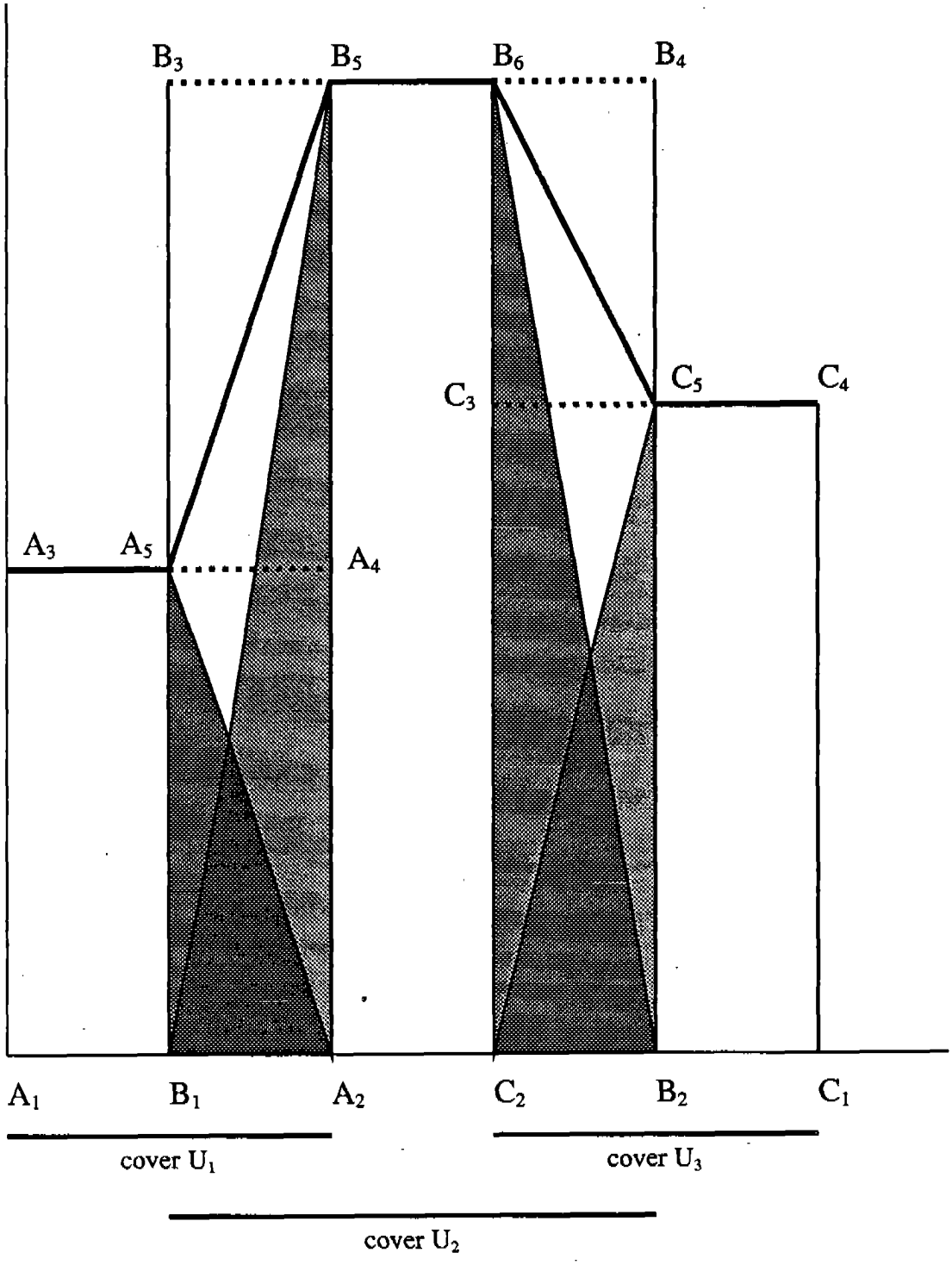

B3
B5
B6
B4
C3
C5
C4
A3
A5
A4
A1
B1
A2
C2
B2
C1
cover U1
cover U3
cover U2

Figure 3.

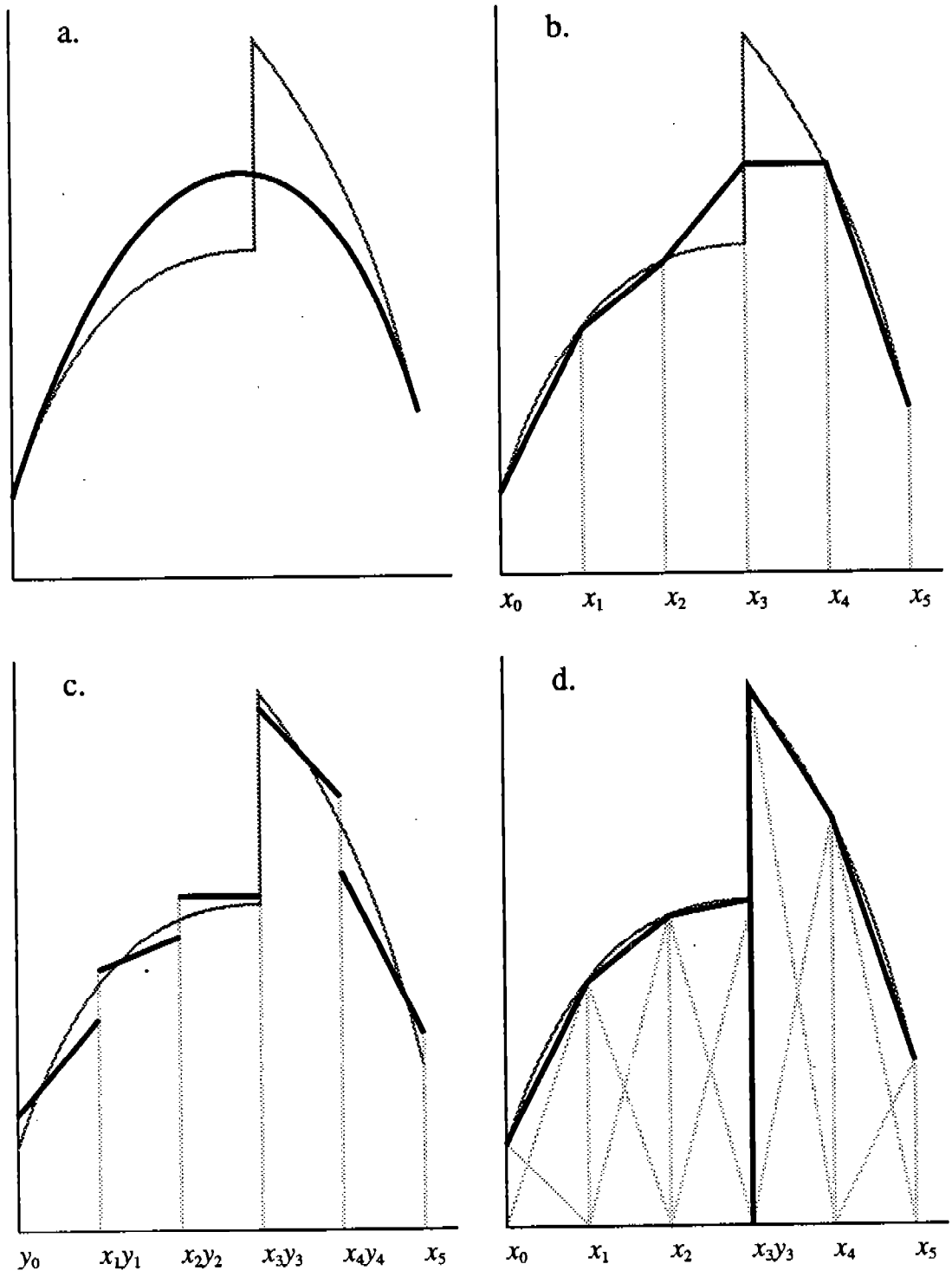

a.
b.
x_0 x_1 x_2 x_3 x_4 x_5
c.
d.
y_0 x_1y_1 x_2y_2 x_3y_3 x_4y_4 x_5
x_0 x_1 x_2 x_3y_3 x_4 x_5

Figure 4.

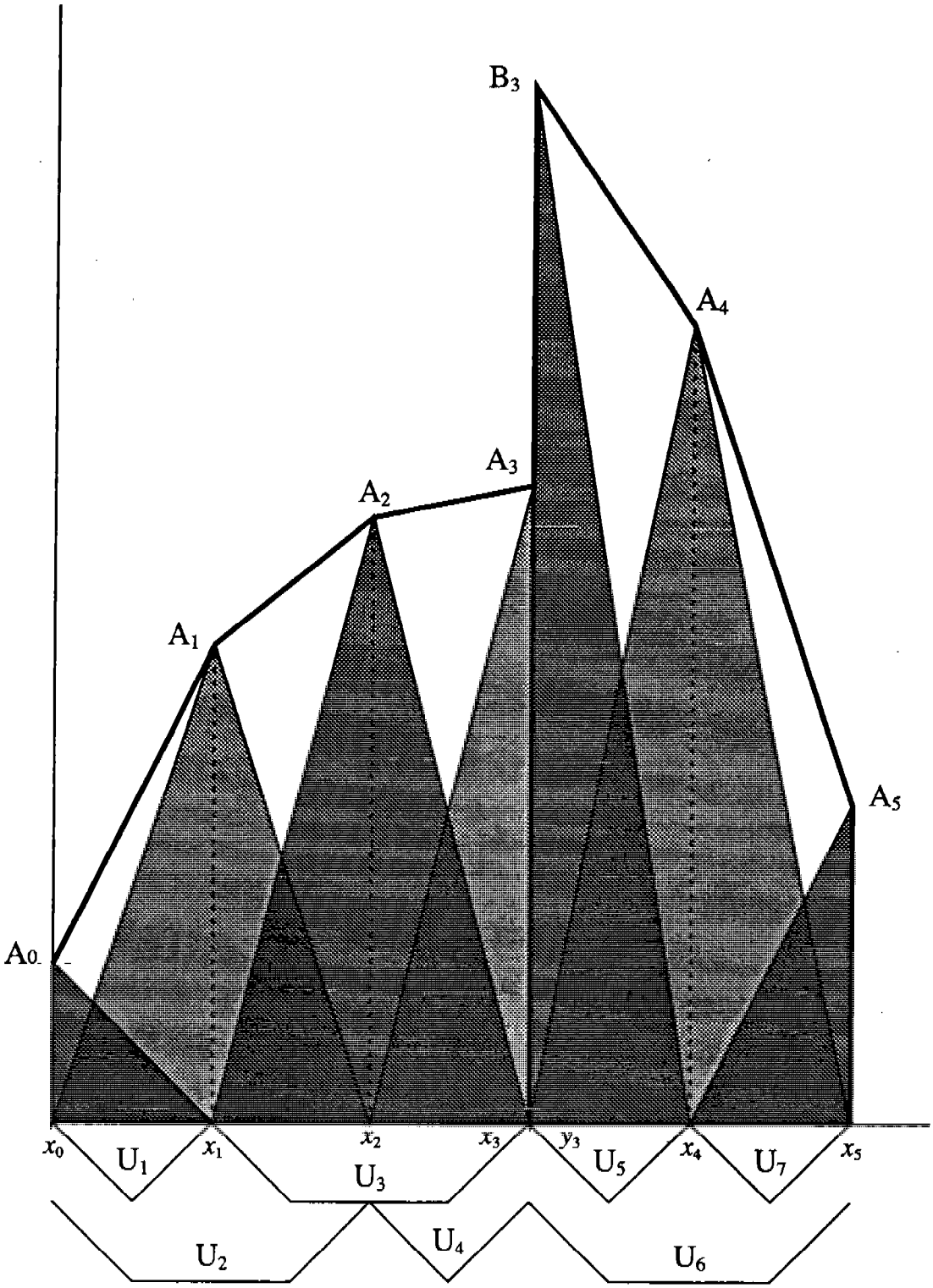


Figure 5.

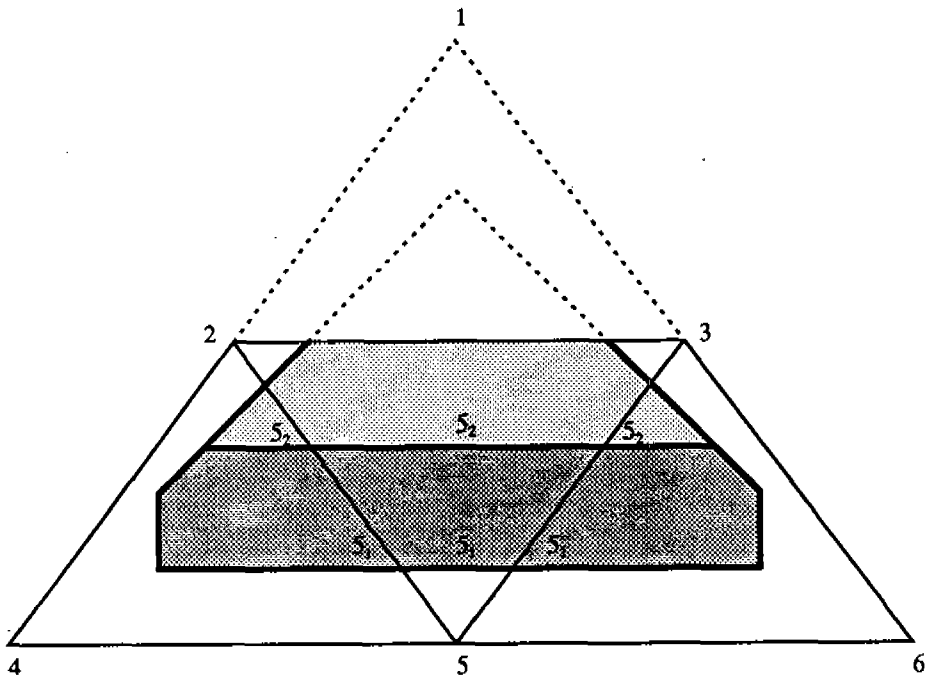


Figure 6a.

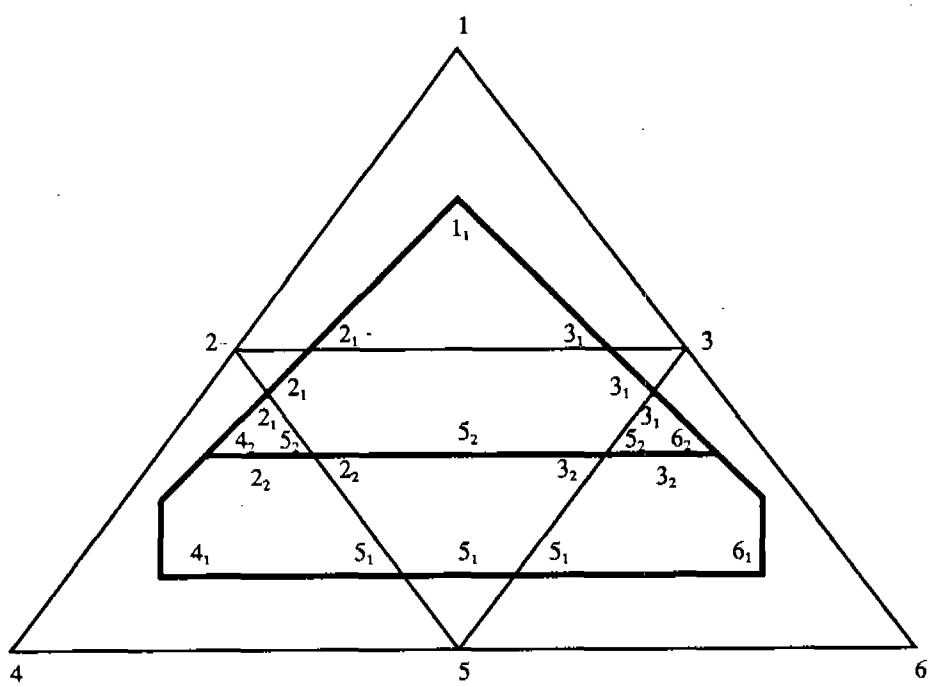


Figure 6b.

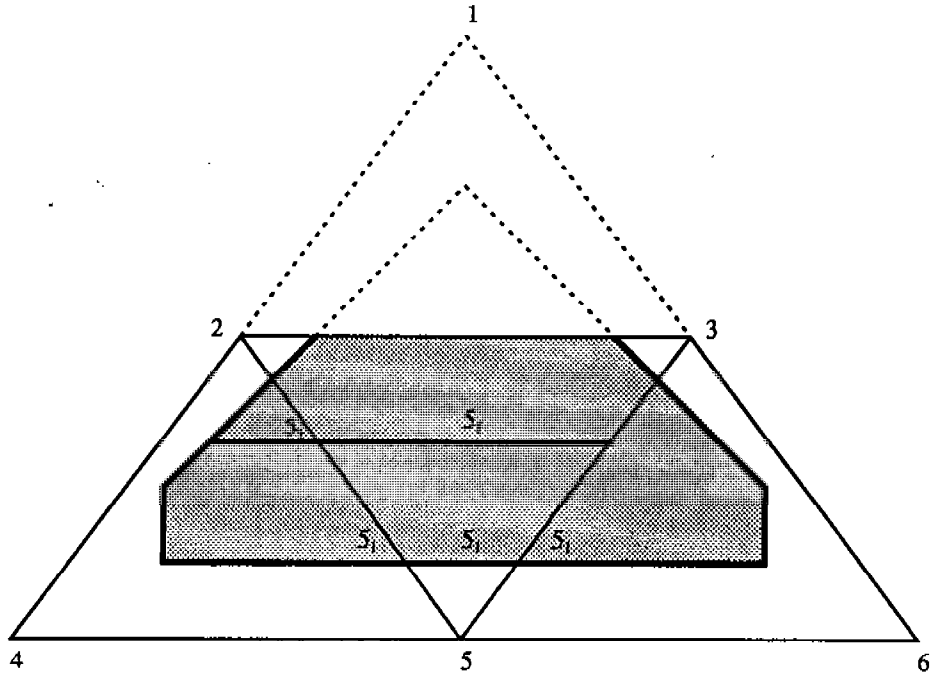


Figure 7a.

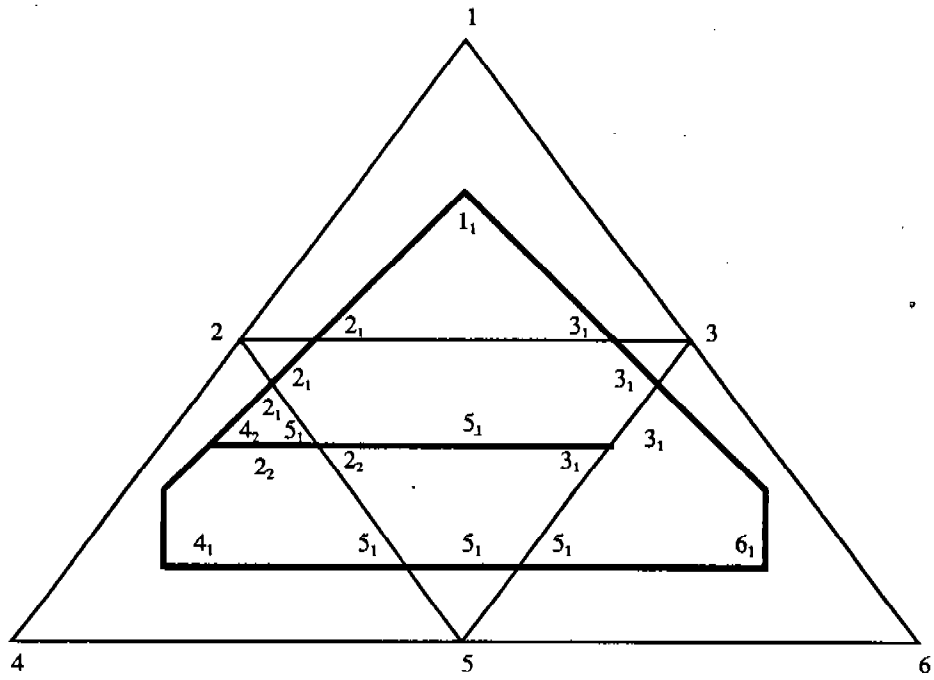


Figure 7b.

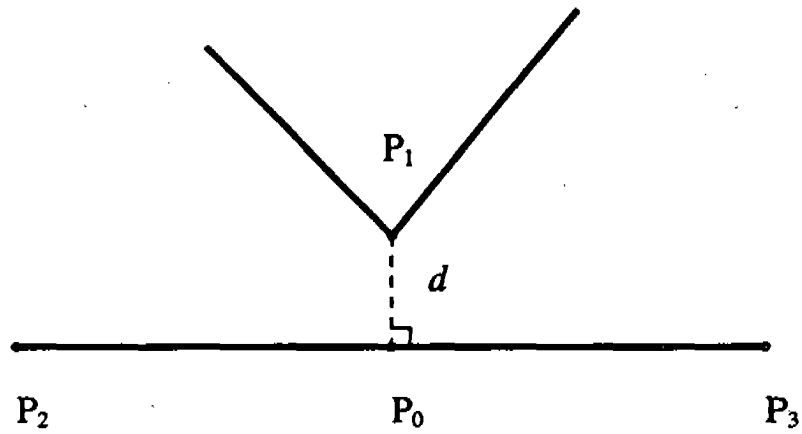


Figure 8.

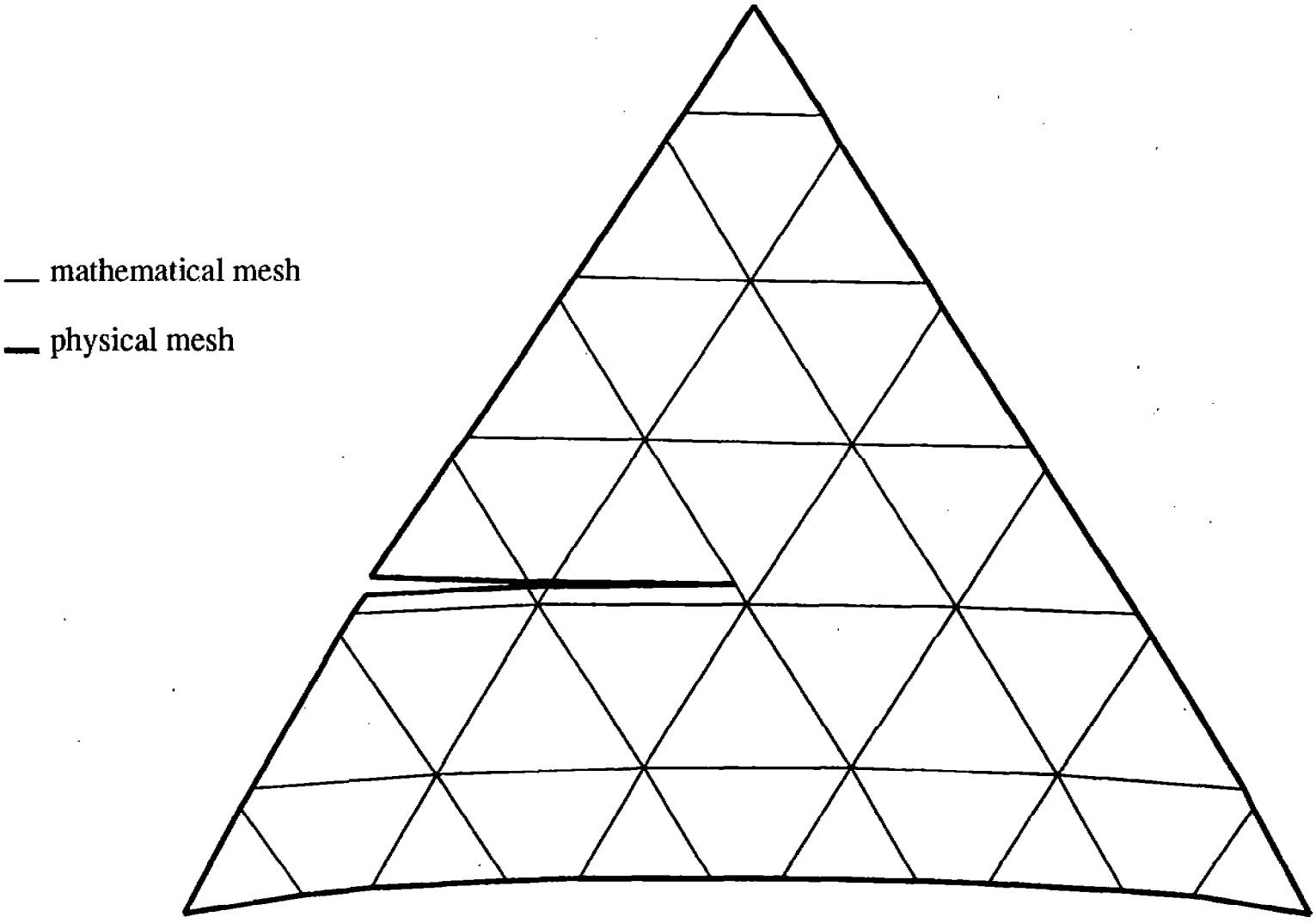


Figure 9.

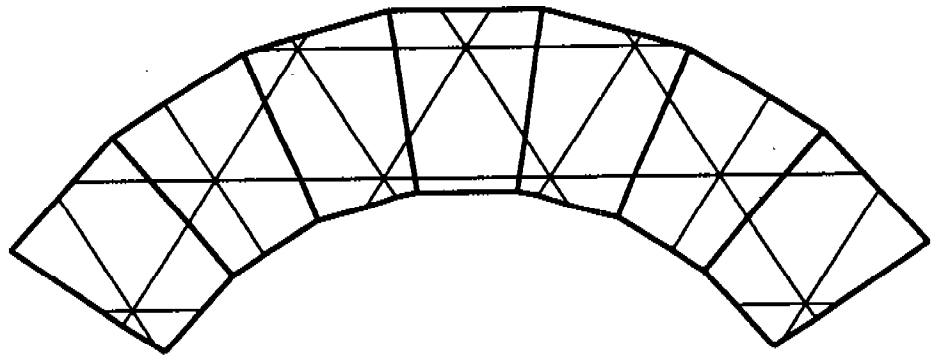


Figure 10a.

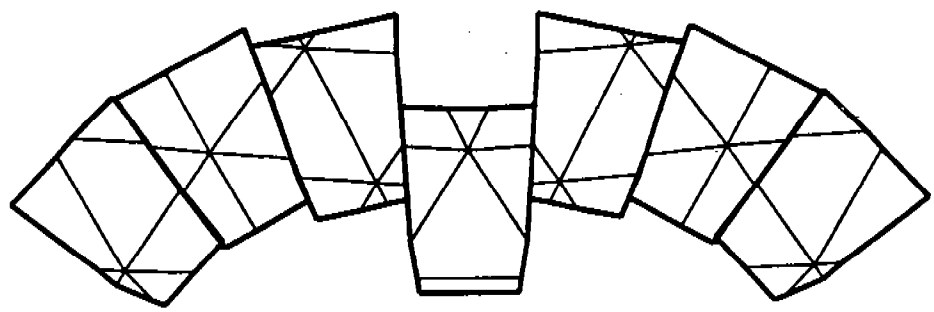


Figure 10b.

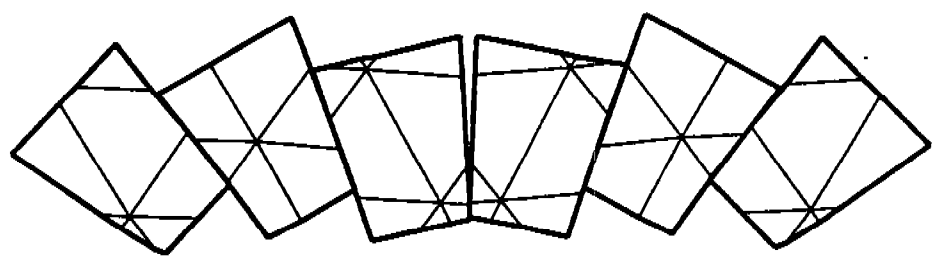
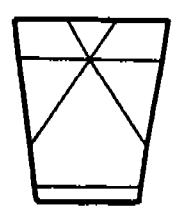


Figure 10c.



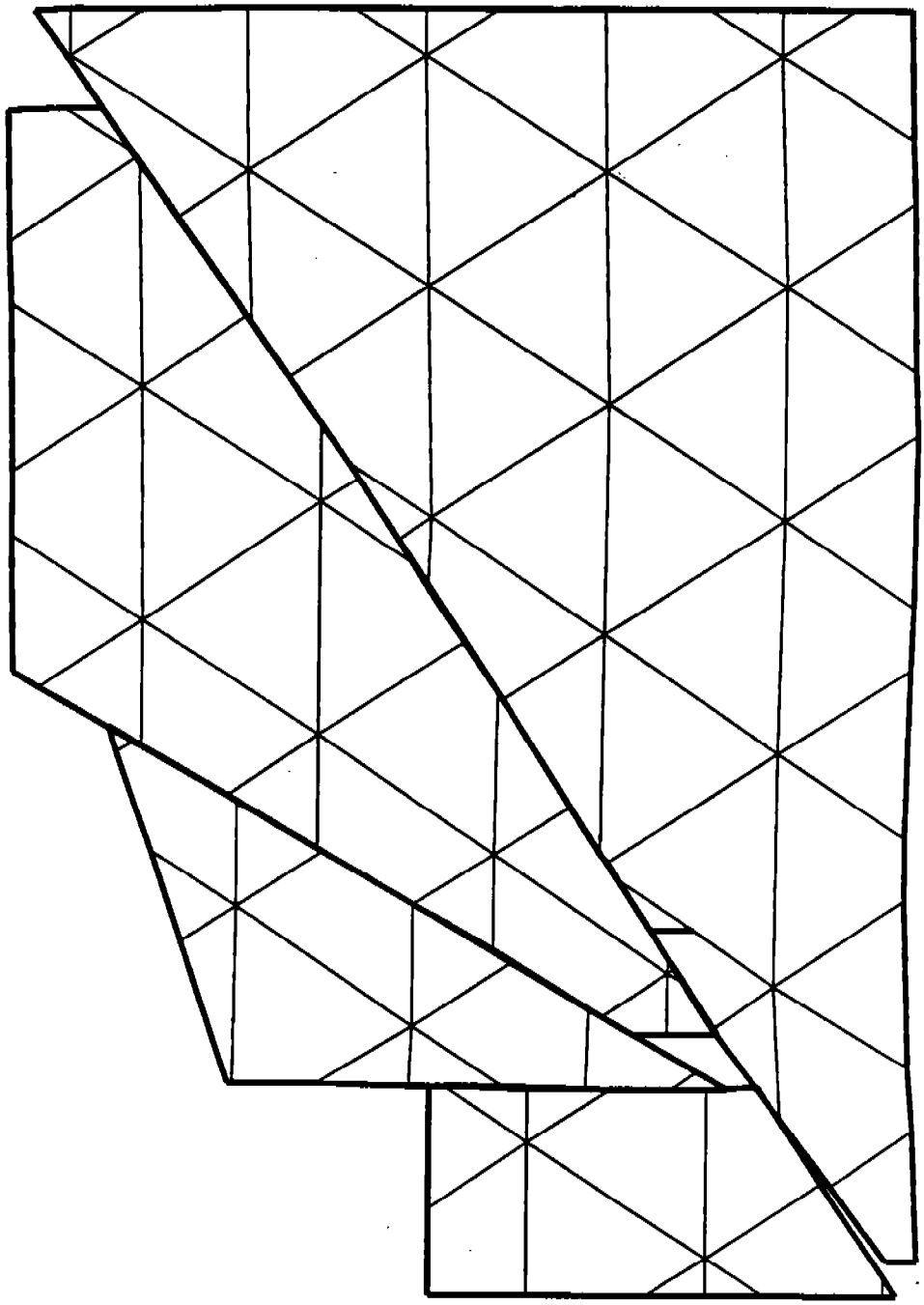


Figure 11.

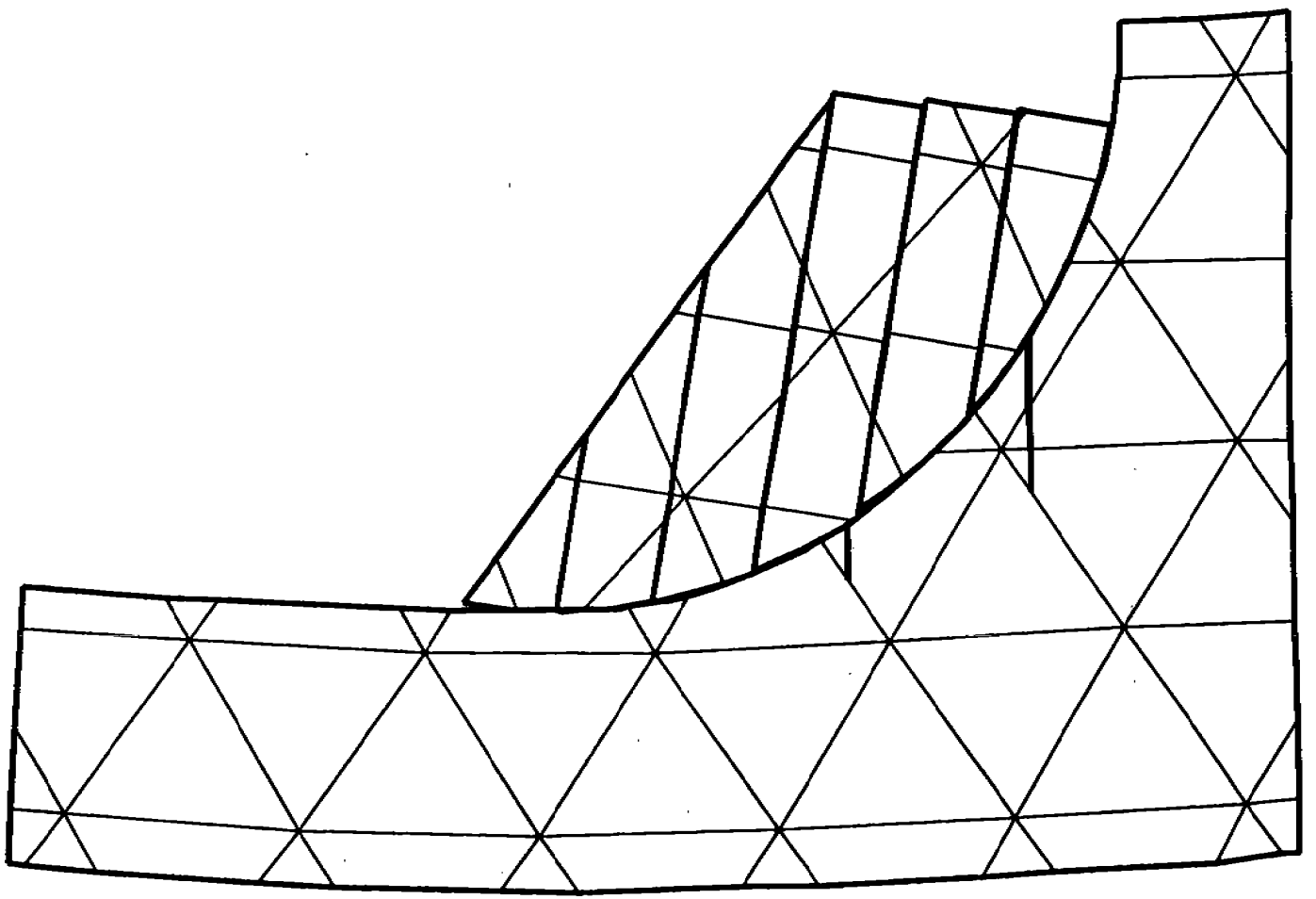


Figure 12.

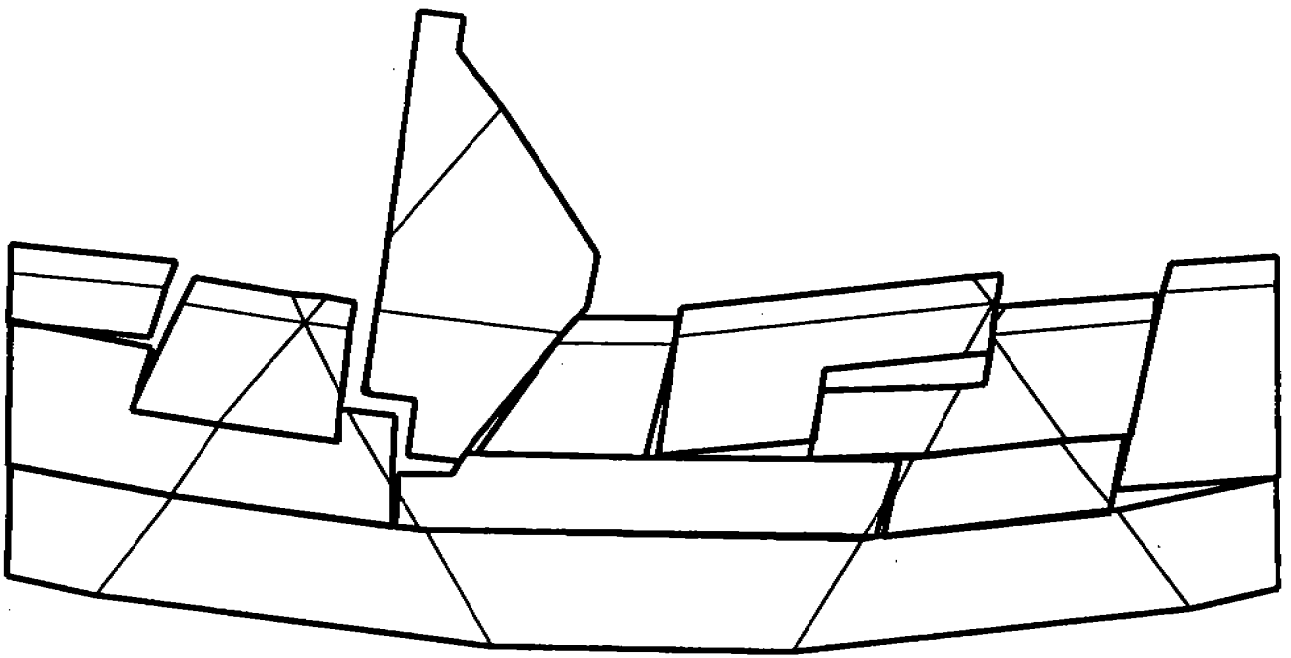


Figure 13.

ANALYTICAL SOLUTION OF ELASTIC-PLASTIC THICK-WALLED CYLINDERS WITH GENERAL HARDENING

Peter C.T. Chen
U.S. Army Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Laboratories
Watervliet, NY 12189-4050

ABSTRACT. This paper presents an analytical solution for the elastic-plastic behavior of thick-walled cylinders under internal pressure. The general hardening law employed in this investigation is a piecewise linear representation of arbitrary stress-strain curves in uniaxial form. Closed-form analytical solutions are developed for the stresses, the elastic and plastic strains, and the displacements by using Tresca's yield criterion and its associated flow rule. Experimental data obtained from cylinders made of either SAE 1045 steel, OFHC copper, or aluminum alloy 1100 are used to determine the material constants. Numerical results for partially-plastic and fully-plastic cylinders are presented for the radial distributions of plastic hoop strain, radial, and tangential stresses.

1. INTRODUCTION. Of all the available elastic-plastic solutions, the problem of pressurized thick-walled cylinders has received the greatest attention. This is because of the symmetric nature of the problem and its practical importance to pressure vessels and the autofrettage process of gun tubes. Many solutions for this problem have been reported over the last four decades [1-3]. Analytical solutions can be obtained only when simplifying assumptions are made regarding material properties. Bland [2] developed analytical solutions for materials with linear hardening properties. Recently, Megahed [3] considered a nonlinear hardening law $\sigma = Y + A \cdot \epsilon_p^n$ in uniaxial form and developed an approximate solution for any value of the strain-hardening exponent n. Closed-form analytical solutions for the plastic hoop strain can be obtained only for four particular values (n = 1, 1/2, 1/3, and 1/4), and the integral has to be evaluated numerically for n = 1/3 and 1/4.

The general hardening law employed in this investigation is a piecewise linear representation of actual stress-strain curves in uniaxial form. A finite number of straight lines can represent arbitrary curves with greater accuracy than other representations [4]. The problem is formulated in a manner similar to [2,3] by using Tresca's yield criterion and the associated flow rule. Closed-form analytical solutions are developed for the stresses, the elastic and plastic strains, and the displacements.

2. BASIC EQUATIONS. Consider a long thick-walled cylinder, internal radius a and external radius b, that is subjected to internal pressure p causing partial plastification. Assuming small strain and no body forces in the axisymmetric state of generalized plane-strain, the radial and tangential stress, σ_r and σ_θ, must satisfy the equilibrium equation

$$r(d\sigma_r/dr) = \sigma_\theta - \sigma_r \tag{1}$$

and the corresponding strains, ϵ_r and ϵ_θ, are given in terms of the radial displacement, u, by

$$\epsilon_r = du/dr \quad , \quad \epsilon_\theta = u/r \tag{2}$$

Total strains are decomposed into elastic and plastic components and the strain-stress relations are

$$\epsilon_r = [\sigma_r - \nu(\sigma_\theta+\sigma_z)]/E + \epsilon_r^p \tag{3a}$$

$$\epsilon_\theta = [\sigma_\theta - \nu(\sigma_r+\sigma_z)]/E + \epsilon_\theta^p \tag{3b}$$

$$\epsilon_z = [\sigma_z - \nu(\sigma_r+\sigma_\theta)]/E + \epsilon_z^p \tag{3c}$$

where E and ν are elastic constants. Subject to $\sigma_\theta \geq \sigma_z \geq \sigma_r$, Tresca's criterion states that yielding occurs when

$$\sigma_\theta - \sigma_r = \sigma(\bar{\epsilon}^p) \tag{4}$$

where the yield stress σ is a function of plastic strain ϵ^p. The associated flow rule states that

$$d\epsilon_\theta^p = -d\epsilon_r^p \geq 0 \quad \text{and} \quad d\epsilon_z^p = 0 \tag{5}$$

Hence, from Eq. (3c)

$$\sigma_z = \nu(\sigma_r+\sigma_\theta) + E\epsilon_z \tag{6a}$$

and the total axial force on any section is

$$F = 2\pi\nu a^2 p + \pi E(b^2-a^2)\epsilon_z \tag{6b}$$

There are three cases of importance: first, plane-strain, $\epsilon_z = 0$; second, a tube with open ends, $F = 0$; and third, a tube with closed ends, $F = \pi a^2 p$. In the latter two cases, substitution into Eq. (6b) determines ϵ_z. Since ϵ_z is now known, Eqs. (3a) and (3b) are inverted in order to express stresses in terms of total strains and plastic hoop strain as follows:

$$\sigma_r = \hat{E}[\nu\epsilon_\theta + (1-\nu)\epsilon_r + (1-2\nu)\epsilon_\theta^p + \nu\epsilon_z] \tag{7a}$$

$$\sigma_\theta = \hat{E}[\nu\epsilon_r + (1-\nu)\epsilon_\theta - (1-2\nu)\epsilon_\theta^p + \nu\epsilon_z] \tag{7b}$$

where $\hat{E} = E/[(1+\nu)(1-2\nu)]$. Substitution of Eqs. (7a) and (7b) into Eqs. (1) and (2) yields the following differential equation:

$$\frac{d^2u}{dr^2} + \frac{1}{r}\frac{du}{dr} - \frac{u}{r^2} = -\frac{1-2\nu}{1-\nu}\left[\frac{2\epsilon_\theta^p}{r} + \frac{d\epsilon_\theta^p}{dr}\right] \tag{8}$$

Integrating with respect to r leads to

$$du/dr + u/r = -\left(\frac{1-2\nu}{1-\nu}\right)(2J+\epsilon_\theta^p) + 2C \tag{9a}$$

where

$$J = \int_a^r \epsilon_\theta^p \, r^{-1} dr \tag{9b}$$

Integrating again yields the analytical solution

$$U = -(\frac{1-2\nu}{1-\nu})rJ + Cr + D/r \tag{10}$$

where C and D are integration constants to be determined from boundary conditions: $\sigma_r = -p$ at $r = a$ and $\sigma_r = 0$ at $r = b$.

Upon substitution of the resulting values of C and D into the expressions of displacement, radial and hoop stresses, the following distributions are obtained:

$$\sigma_r = -p - \bar{E}J + (P+\bar{E}J_o)(1-a^2/r^2)/(1-a^2/b^2) \tag{11a}$$

$$\sigma_\theta = -p - \bar{E}(J+\epsilon_\theta^p) + (p+\bar{E}J_o)(1+a^2/r^2)/(1-a^2/b^2) \tag{11b}$$

$$EU/r = (1+\nu)(1-2\nu)\sigma_r + 2(1-\nu^2)(a^2/r^2)(p+\bar{E}J_o)/(1 - \frac{a^2}{b^2}) - \nu E\epsilon_z \tag{11c}$$

where $\bar{E} = E/(1-\nu^2)$, and J_o is the value of the integral J at the plastic front, $r = \rho$, i.e.,

$$J_o = \int_a^\rho \epsilon_\theta^p \, r^{-1} \, dr = \int_1^{\rho/a} \epsilon_\theta^p \, \xi^{-1} d\xi \tag{11d}$$

Note that $\epsilon_\theta^p = 0$ and $J = J_o$ throughout the outer elastic zone defined by $\rho < r < b$. At the plastic front, the Tresca effective stress $\sigma = \sigma_\theta - \sigma_r = Y$, where Y is the initial yield stress and also $\epsilon_\theta^p = 0$. Therefore, using Eq. (11b) to provide σ, one readily obtains

$$P + \bar{E} \, J_o = \frac{\sigma_o}{2} \frac{\rho^2}{a^2} (1 - \frac{a^2}{b^2}) \tag{12}$$

Using Eq. (7), the distributions of σ_r, σ_θ, and u can be written in simpler forms as follows:

$$\sigma_r = -p - \bar{E}J + \frac{Y}{2} (\rho^2/a^2 - \rho^2/r^2) \tag{13a}$$

$$\sigma_\theta = \sigma_r + Y(\rho^2/r^2 - \frac{\bar{E}}{\sigma_o} \epsilon_\theta^p) \tag{13b}$$

$$Eu/r = (1+\nu)(1-2\nu)\sigma_r + Y(1-\nu^2)\rho^2/r^2 - \nu E\epsilon_z \tag{13c}$$

It is obvious from Eq. (13b) that Tresca effective stress $\bar{\sigma}$ is simply given by

$$\sigma = Y(\rho^2/r^2 - \frac{\bar{E}}{\sigma_o} \epsilon_\theta^p) \tag{14}$$

Therefore, if the radial variation of plastic hoop strain is known, the integral J and all field quantities can be determined.

3. GENERAL HARDENING. The general hardening law employed in this investigation is a piecewise linear representation. Arbitrary stress-strain curves in uniaxial form can be approximated by a finite number of straight lines [4]. The straight line through the origin is given by the relation

$$\bar{\sigma} = E\bar{\epsilon} \tag{15}$$

where E is Young's modulus. All of the other straight lines are given by the relation

$$\bar{\sigma} = (1-m_i)\sigma_{oi} + m_i E\bar{\epsilon} \tag{16}$$

where σ_{oi} is the stress at the intersection of the two straight lines given by Eqs. (15) and (16), and m_iE is the slope of the straight lines given by Eq. (16). Let σ_i, ϵ_i be the stress and strain at the intersection of two straight lines with slope $m_{i-1}E$ and m_iE as shown in Figure 1a. Then

$$\sigma_i = (1-m_{i-1})\sigma_{oi-1} + m_{i-1}E\epsilon_i = (1-m_i)\sigma_{oi} + m_iE\epsilon_i$$

which leads to ϵ_i and σ_i in terms of σ_{oi} and m_i

$$E\epsilon_i = [(1-m_i)\sigma_{oi} - (1-m_{i-1})\sigma_{oi-1}]/(m_{i-1}-m_i) \tag{17a}$$

and

$$\sigma_i = [m_{i-1}(1-m_i)\sigma_{oi} - m_i(1-m_{i-1})\sigma_{oi-1}]/(m_{i-1}-m_i) \tag{17b}$$

Eq. (16) can be written also as a function of effective plastic strain $\bar{\epsilon}^p$ as shown in Figure 1b.

$$\bar{\sigma} = \sigma_{oi} + h_iE\bar{\epsilon}^p \qquad \bar{\epsilon}_i^p \leq \bar{\epsilon}^p \leq \bar{\epsilon}_{i+1}^p \tag{18}$$

where $h_i = m_i/(1-m_i)$.

Since $\epsilon_\theta^p = -\epsilon_r^p$ and $\epsilon_z^p = 0$, the effective plastic strain $\bar{\epsilon}^p$ is determined as $2\epsilon_\theta^p/\sqrt{3}$, and hence, Eq. (14) is rewritten in terms of the plastic strain in the tube as

$$\sigma = Y(\rho^2/r^2 - \frac{\sqrt{3}}{2}\frac{E}{Y}\bar{\epsilon}^p) \tag{19}$$

A comparison between the expressions for effective stresses provided by Eqs. (18) and (19) yields the following explicit equation for $\bar{\epsilon}^p$:

$$\frac{E}{Y}\bar{\epsilon}^p = (\rho^2/r^2 - C_i)/b_i \tag{20}$$

which is valid in $\bar{\epsilon}_i^p \leq \bar{\epsilon}^p \leq \bar{\epsilon}_{i+1}^p$ and $r_i \geq r \geq r_{i+1}$ and

$$b_i = (\sqrt{3}/2)/(1-\nu^2) + h_i \quad , \quad C_i = \sigma_{oi}/\sigma_o \tag{21}$$

The values of r_i and r_{i+1} can be determined by

$$r_i = \rho[b_i \frac{E}{Y} \bar{\epsilon}_i^p + C_i]^{-\frac{1}{2}}$$

$$r_{i+1} = \rho[b_i \frac{E}{Y} \bar{\epsilon}_{i+1}^p + C_i]^{-\frac{1}{2}} \tag{22}$$

If $\bar{\epsilon}_1^p = 0$ and $\sigma_{01} = Y$, then $r_1 = \rho$. This is true for most materials. Since $r_i \geq a$, the calculation of r_i for $i = 1,2...m$ should stop when the above relation is violated, i.e., $r_{m+1} < a$. Let us define V_i for $i = 1,2,...m$ by the following integral:

$$V_i = \int_{r_{i+1}}^{r_i} \frac{E}{Y} \bar{\epsilon}^p \frac{dr}{r} = \frac{1}{b_i} \int_{r_{i+1}}^{r_i} (\frac{\rho^2}{r^2} - C_i) \frac{dr}{r}$$

$$= \frac{1}{2b_i} [(\frac{\rho}{r_{i+1}})^2 - (\frac{\rho}{r_i})^2 - 2C_i \ln(r_i/r_{i+1})] \tag{23}$$

Then

$$V = \int_a^r \frac{E}{Y} \bar{\epsilon}_p \frac{dr}{r} = \int_a^{r_m} \frac{E}{Y} \bar{\epsilon}^p \frac{dr}{r} + V_{m-1} + ... + V_{i+1} + \int_{r_{i+1}}^r \frac{E}{Y} \bar{\epsilon}^p \frac{dr}{r}$$

$$= \frac{1}{2b_m} [(\frac{\rho}{a})^2 - (\frac{\rho}{r_m})^2 - 2C_m \ln \frac{r_m}{a}] + V_{m-1} + ... + V_{i+1}$$

$$+ \frac{1}{2b_i} [(\frac{\rho}{r_{i+1}})^2 - (\frac{\rho}{r})^2 - 2C_i \ln (r/r_{i+1})] \tag{24}$$

and the maximum value of V is

$$V_o = \int_a^\rho \frac{E}{Y} \bar{\epsilon}_p \frac{dr}{r} = \frac{1}{2b_m} [(\frac{\rho}{a})^2 - (\frac{\rho}{r_m})^2 - 2C_m \ln \frac{r_m}{a}] + \sum_{i=1}^{m-1} V_i \tag{25}$$

The integrals V_i (i=1,2,...,m-1), V, and V_o given analytically by Eqs. (23), (24), and (25) can be easily evaluated. The integral J is related to the integral V by

$$J = \frac{\sqrt{3}}{2} \frac{Y}{E} V \tag{26}$$

All field quantities u, ϵ_r, ϵ_θ, σ_r, σ_θ, σ_z, and ϵ_θ^p can now be calculated.

4. MATERIAL PROPERTIES. Test members were made from three different metals as follows: SAE 1045 steel, OFHC copper, and aluminum alloy 1100 [4]. The values of the elastic constants (E and ν) for the three metals are shown in Table 1. The values of the constants (σ_{0i}, m, σ_i, ϵ_i) approximating the plastic portion of the stress-strain diagram for three metals are shown in Table 2.

TABLE 1. ELASTIC CONSTANTS FOR THREE METALS

Material	E, Ksi	ν
SAE 1045 steel	30,000	0.29
OFHC copper	16,000	0.35
aluminum alloy 1100	10,250	0.33

TABLE 2. PLASTIC CONSTANTS FOR THREE METALS

Straight Line	σ_{0i}, Ksi	m_i	σ_i, Ksi	ϵ_i, %
SAE 1045 Steel				
1	43.4	0.05083	43.4	0.145
2	54.0	0.02858	66.924	1.687
3	80.0	0.00847	90.638	4.453
4	95.0	0.00309	103.542	9.532
5	111.0	0.00128	122.280	29.745
OFHC Copper				
1	2.50	0.17125	2.50	0.016
2	3.25	0.07063	3.686	0.059
3	4.00	0.03125	4.553	0.136
4	5.37	0.01991	7.700	0.765
5	8.40	0.01313	14.151	2.790
6	21.0	0.00450	27.484	9.137
7	39.0	0.00078	42.757	30.350
Aluminum Alloy 1100				
1	8.0	0.67024	8.0	0.078
2	11.0	0.32683	11.942	0.135
3	13.0	0.09561	13.557	0.184
4	14.7	0.01590	15.007	0.332
5	16.1	0.00210	16.310	1.131

Each of the stress-strain diagrams can be approximated by a finite number of straight lines with extreme accuracy. The error introduced by the approximation is less than 1 percent for all cases.

5. NUMERICAL RESULTS. Typical results for the analytical solution are presented first by means of prescribing a plastic front and determining the corresponding plastic hoop strain and radial and hoop stresses in the tube. A tube with b/a = 2 is employed, and the plastic front is prescribed at ρ/a = 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0. Figure 2 illustrates the stresses and plastic hoop strains obtained using the material constants for SAE 1045 steel. Figures 3 and 4 present similar results for OFHC copper and aluminum alloy 1100, respectively. Figure 5 shows a comparison of stresses and plastic hoop strains for three partially-plastic tubes at ρ/a = 1.6. Figure 6 presents a similar comparison for three fully-plastic tubes at ρ/a = 2.0. Future work related to the results obtained here will look into the elastic-plastic behavior of the tube during pressure release. The influence of phenomena such as the Bauschinger effect on residual stresses should be modelled [5,6].

REFERENCES

1. Hill, R., The Mathematical Theory of Plasticity, Oxford University Press, London, 1950.

2. Bland, D.R., "Elastoplastic Thick-Walled Tubes of Work-Hardening Materials Subject to Internal and External Pressures and Temperature Gradients," Journal of Mechanics and Physics of Solids, Vol. 4, 1956, pp. 209-229.

3. Megahed, M.M., "Elastic-Plastic Behavior of a Thick-Walled Tube With General Nonlinear Hardening Properties," International Journal of Mechanical Sciences, Vol. 32, 1990, pp. 551-563.

4. Sidebottom, O.M. and Chu, S.C., "Bursting Pressure of Thick-Walled Cylinders Axial Load, and Torsion," Experimental Mechanics, Vol. 15, 1975, pp. 209-218.

5. Chen, P.C.T., "The Bauschinger and Hardening Effect on Residual Stresses in an Autofrettaged Thick-Walled Cylinder," Journal of Pressure Vessel Technology, Vol. 108, pp. 108-112.

6. Chen, P.C.T., "Stress and Deformation Analysis of Autofrettaged High Pressure Vessels," ASME PVP, Vol. 110, 1986, pp. 61-67.

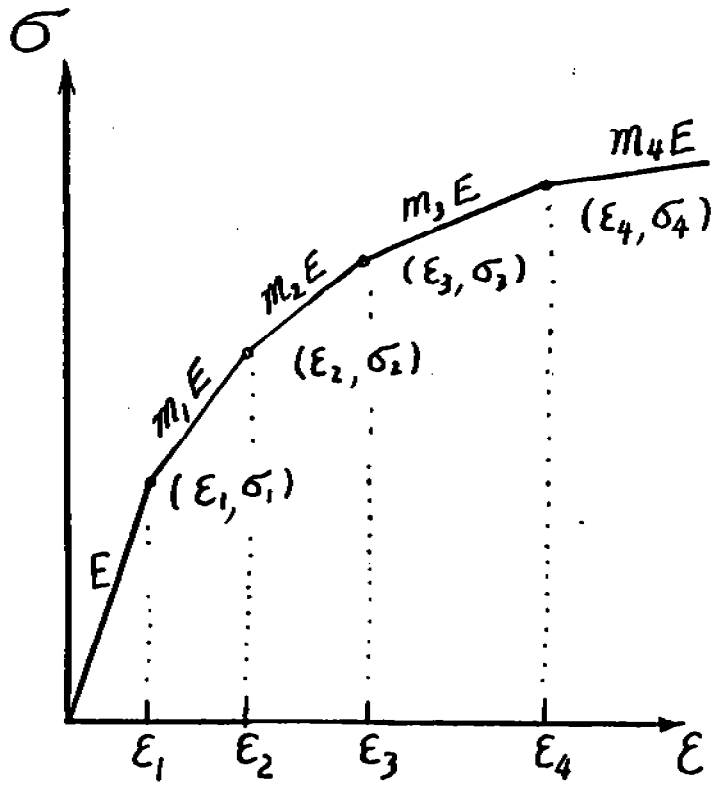


Fig. 1a. Effective stress-strain diagram.

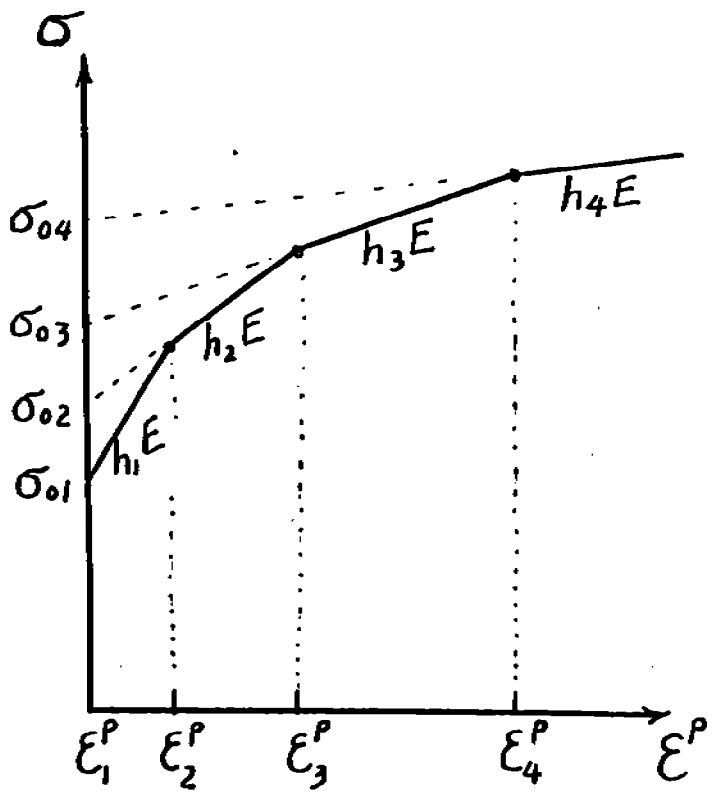


Fig. 1b. Effective stress-plastic strain diagram.

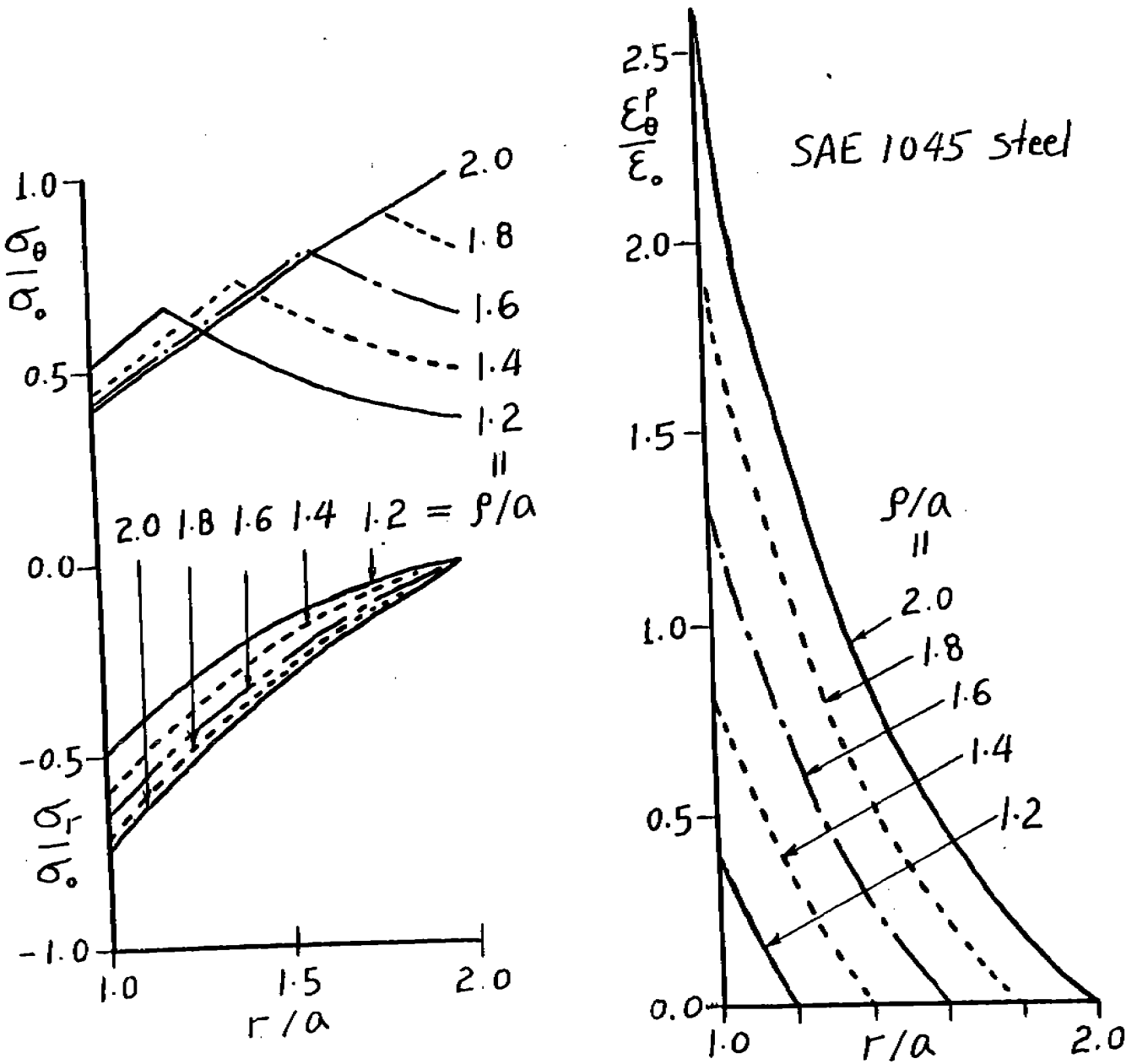


Fig. 2. Variations of stresses and plastic hoop strains in a steel tube.

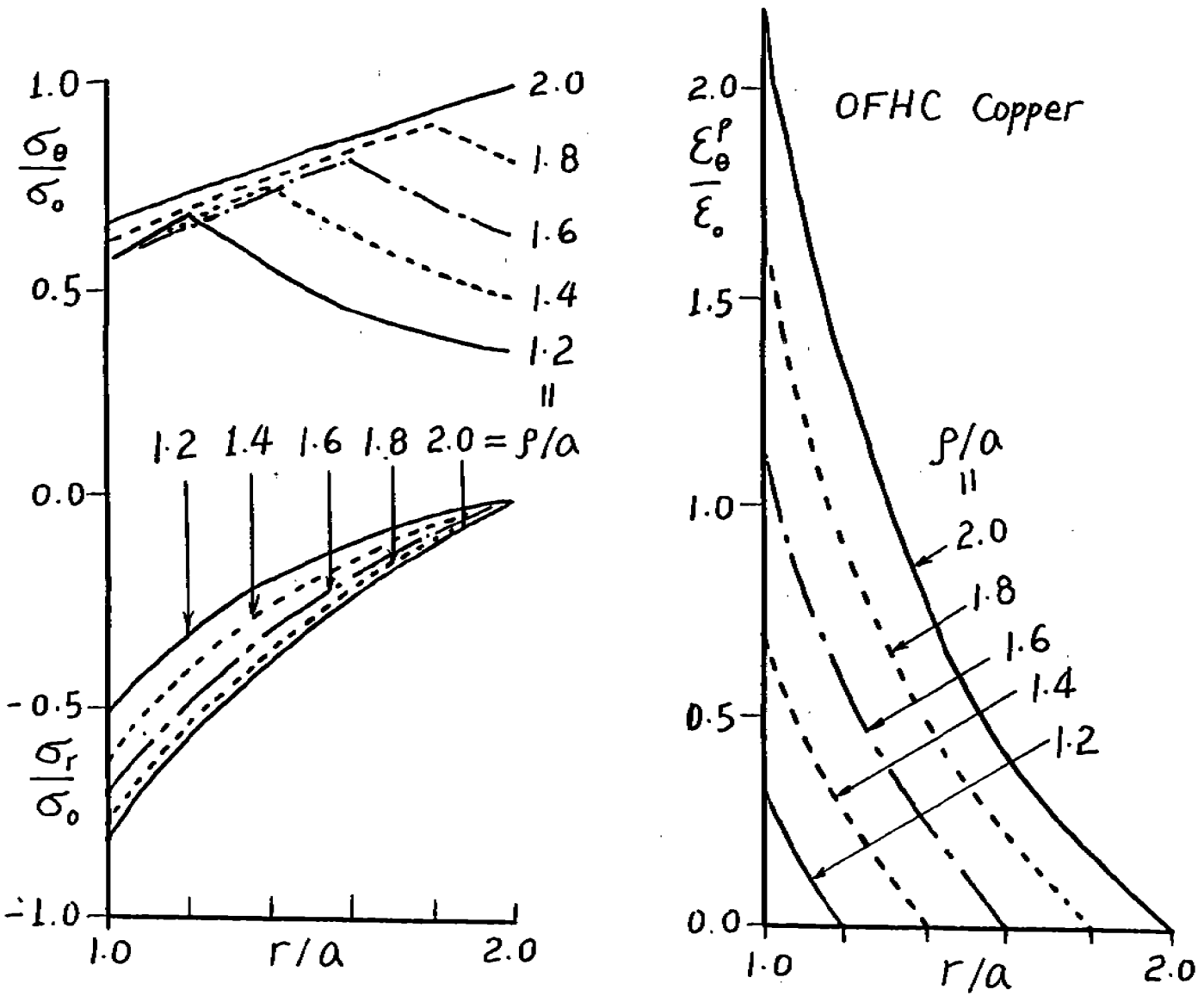


Fig. 3. Variations of stresses and plastic hoop strain in a copper tube.

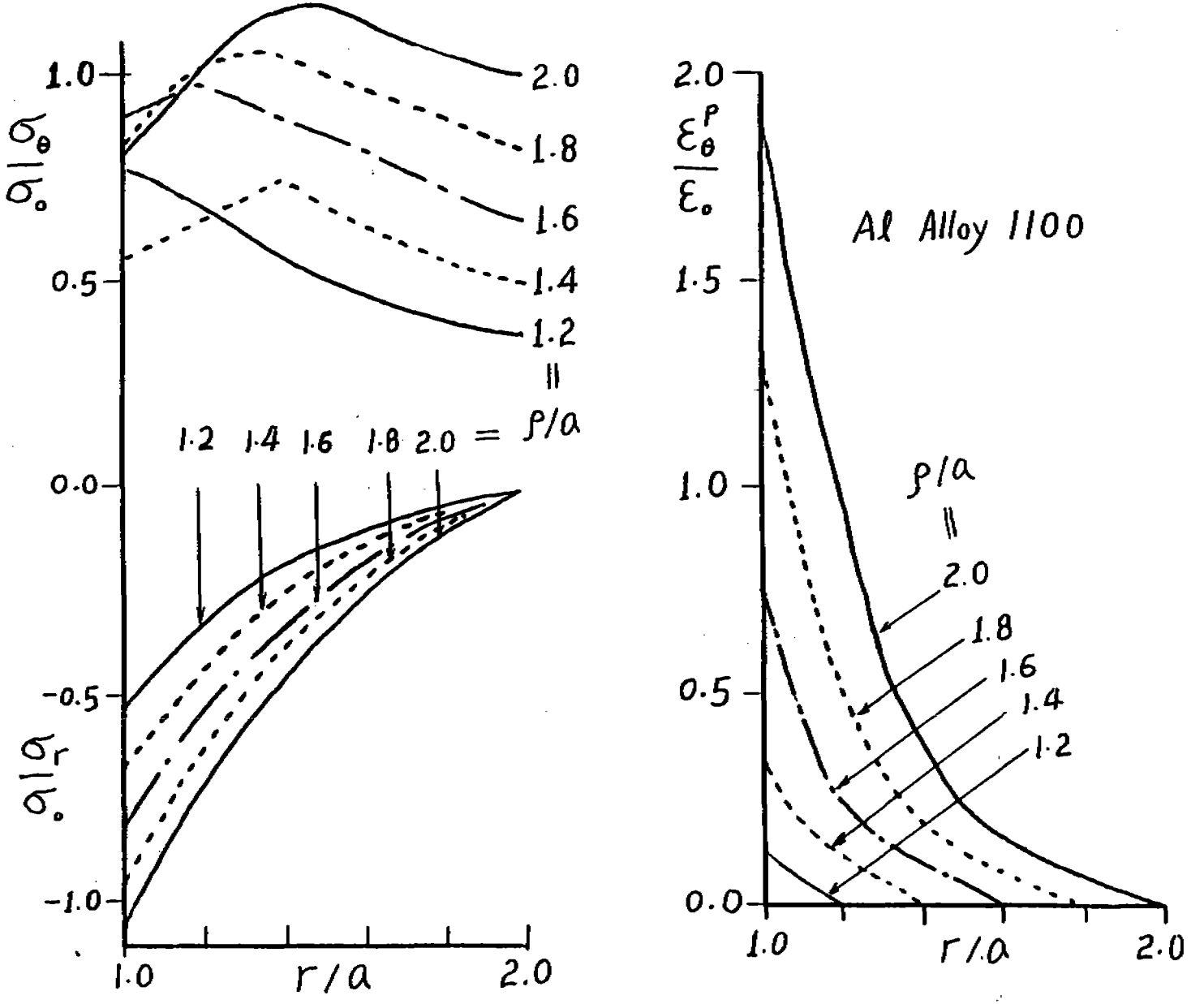


Fig. 4. Variations of stresses and plastic hoop strain in an aluminum alloy tube.

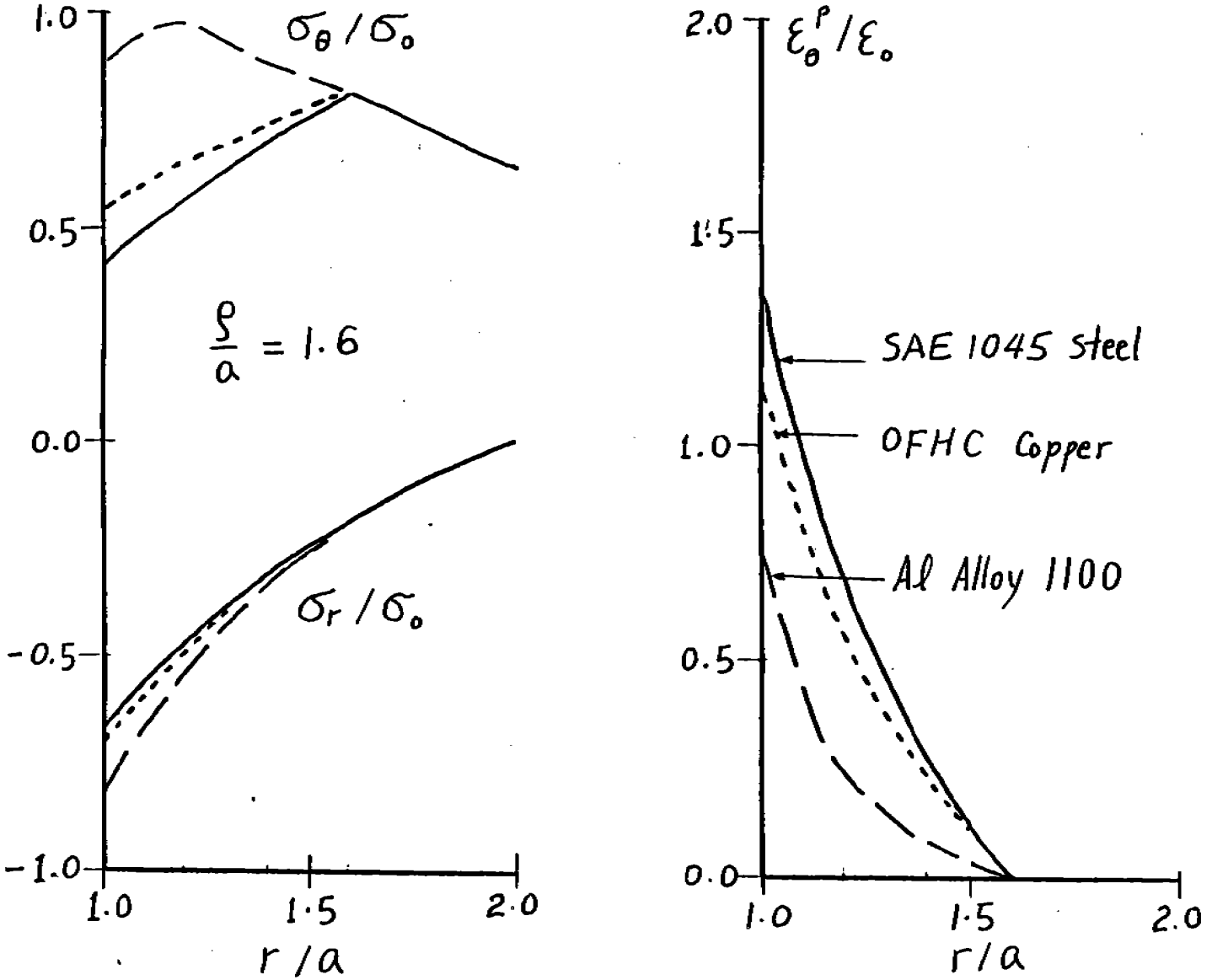


Fig. 5. Comparisons of stresses and plastic hoop strain in a partially-plastic tube.

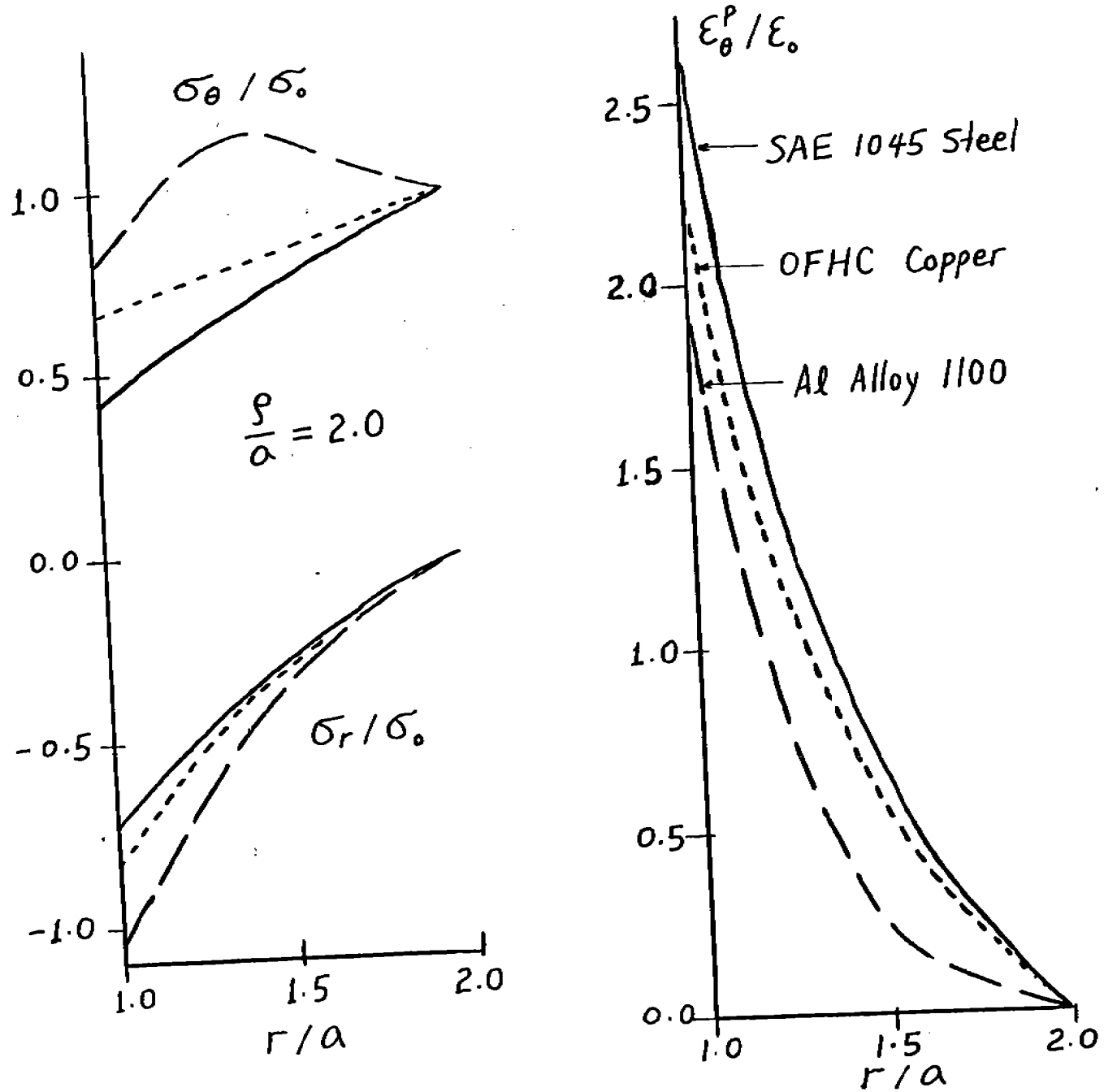


Fig. 6. Comparisons of stresses and plastic hoop strain in a fully-plastic tube.

ANALYSIS OF SHEAR BANDING IN ARMCO IF IRON, TUNGSTEN ALLOY AND DEPLETED URANIUM*

R. C. Batra and C. H. Kim
Department of Mechanical and Aerospace
Engineering and Engineering Mechanics
University of Missouri - Rolla
Rolla, MO 65401-0249

ABSTRACT. We study the problem of the initiation and growth of shear bands in three materials by analyzing the thermomechanical deformations of a block of nonuniform thickness undergoing overall simple shearing deformations. Each of these materials is assumed to obey the Johnson-Cook law. It is found that, for each material, the deformations of the block have become nonhomogeneous by the time the shear stress attains its maximum value. For Armco IF iron, a narrow band at the center develops when the shear stress there has dropped to 85% of its peak value, and the same occurs for the tungsten alloy when the shear stress at the specimen center equals 80% of the maximum value. For the depleted uranium satisfactory results could be computed only till the shear stress dropped to 99% of the peak value.

INTRODUCTION AND GOVERNING EQUATIONS. Tresca (1878) and Massey (1921) observed hot lines, now referred to as shear bands, during the hot forging of a metal. The research activity in this area has increased extensively during the last decade or so, possibly due to the realization that shear bands precede shear fractures, and once a shear band has formed subsequent deformations of the body occur in this narrow region and the strength of the rest of the body is not fully realized. We refer the reader to recent articles by Shawki and Clifton (1989) and Batra and Kim (1990) for references to the shear banding related works.

Even though a more realistic model of any of the experimental studies will involve analyzing at least a two-dimensional problem, we presume that useful information regarding the initiation and growth of a shear band in a material can be obtained by studying the simple shearing deformations of a block of non-uniform thickness and made of the material to be investigated. We assume that the thermomechanical response of each of the three materials studied herein can be modeled by the Johnson-Cook (1983) law with material parameters assigned values given by Johnson *et al.* (1983). Pertinent equations in terms of non-dimensional variables are:

---

* Supported by the U. S. Army Research Office Contract DAAL03-88-K-0184 to the University of Missouri - Rolla

$$\alpha\, w\, \dot{v} = (w\, s)_y \,, \tag{1}$$

$$\dot{s} = \mu(v_y - \dot{\gamma}_p) \,, \tag{2}$$

$$\dot{\gamma}_p = \dot{\gamma}_0 \exp\left[\left(\frac{s}{(A + B\,\gamma_p^n)\,(1 - \theta/\theta_m)} - 1.0\right) / D\right], \tag{3}$$

$$w\dot{\theta} = \beta\,(w\,\theta_y)_y + w\,s\,\dot{\gamma}_p \,, \tag{4}$$

where

$$\alpha = \rho\, v_0^2/\sigma_0,\ \beta = k/(\rho\, v_0\, c\, H), \text{ and } w(y) = w_0\left[1 + \frac{\delta}{2}\sin\left(\frac{1}{2} + 2y\right)\pi\right] \tag{5}$$

gives the thickness variation in the block. A schematic sketch of the problem studied is shown in Figure 1.

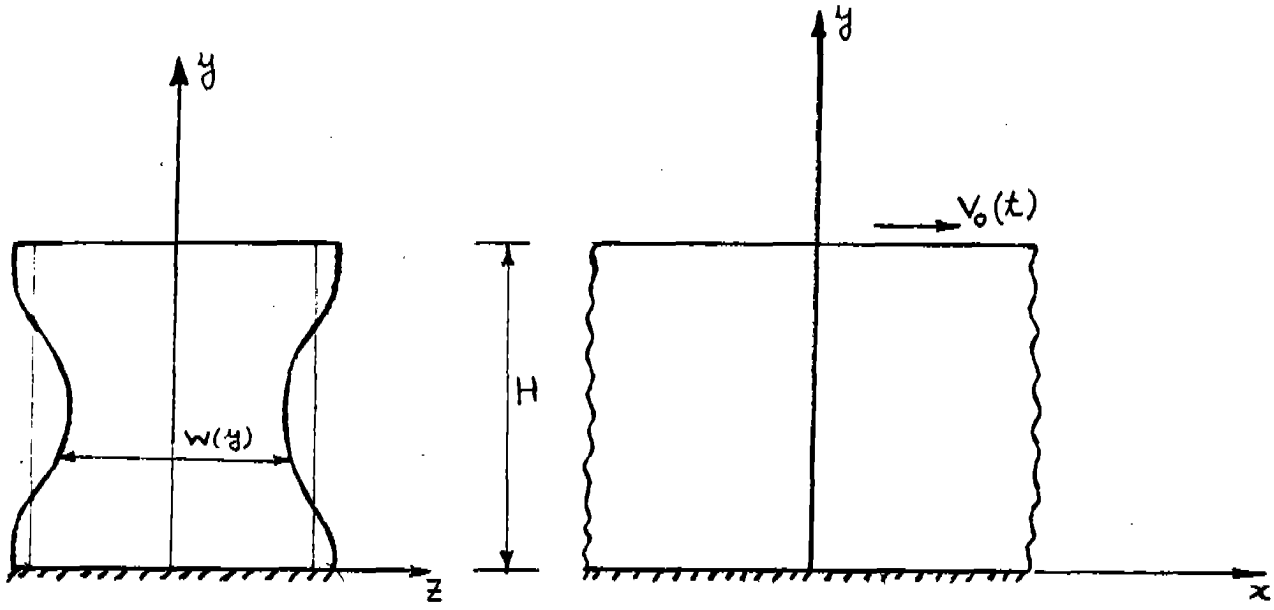


Figure 1. A schematic sketch of the problem studied.

In equations (1) - (4), v is the velocity of a material particle in the direction of shearing, s is the shear stress, a comma followed by y indicates partial differentiation with respect to y, a superimposed dot stands for the time derivative, $\dot{\gamma}_p$ is the plastic strain-rate, equation (3) is the flow-rule proposed by Johnson and Cook, A, B, n, θ_m, and D are material parameters, $\dot{\gamma}_0$ is the reference strain-rate, and in equation (4) we have assumed that all of the plastic working is converted into heating. The non-dimensional numbers α and β signify, respectively, the importance of inertia forces relative to the flow stress of the material, and the effect of heat conduction. In the expressions for α and β, ρ is the mass density, v_0 the final value of the speed imposed on the top surface of the block, σ_0 the yield stress in a quasistatic simple shearing test, k the thermal conductivity, c the specific heat, and H the height of the block. The variables have been non-dimensionalized as in Batra and Kim (1992).

For initial and boundary conditions, we take

$$v(y,0) = 0, \quad s(y,0) = 0, \quad \theta(y,0) = 0, \tag{6}$$

$$\theta(0,t) = 0, \quad \theta(1,t) = 0, \quad v(0,t) = 0, \tag{7}$$

$$\begin{aligned} v(1,t) &= t/0.01, \qquad 0 \le t \le 0.01, \\ &= 1, \qquad\quad t \ge 0.01. \end{aligned} \tag{8}$$

That is, the block is initially at rest, is stress free, and is at a uniform temperature. The lower and upper surfaces of the block are kept at a constant temperature by the grips which act as heat sinks, the lower surface is kept fixed while on the top surface, the prescribed shearing speed increases from 0 to 1.0 in a non-dimensional time of 0.01.

We note that the coupled partial differential equations (1) through (4) are highly nonlinear. Their approximate solution under the side conditions (6) - (8) has been obtained by the finite element method described in Batra and Kim (1990).

NUMERICAL RESULTS. In order to compute results, we took

$$H = 3.18\ mm, \quad v_0 = 4.77\ m/sec, \quad w_0 = 0.248, \quad \delta = 0.05,$$

and used a nonuniform finite element mesh with y-coordinate of the n^{th} node given by

$$y_n = 4\left(\frac{n-1}{100} - \frac{1}{2}\right)^2 + \frac{1}{2}, \qquad n = 1,2,3,\ldots,101.$$

Values of material parameters for the three materials studied herein are taken either from

the paper by Johnson *et al.* (1983) or from a handbook are listed in Table 1 of Batra and Kim's paper (1992). We note that the values of $10^3\ \alpha$ and $10^3\ \beta$ for Armco IF (interstitial free) iron, tungsten alloy, and depleted uranium are (2.362, 1.349), (0.449, 0.217), and (0.681, 0.848), respectively.

Figures 2 through 5 show solution surfaces for Armco IF iron depicting, respectively, the evolution of the velocity, shear stress, temperature rise, and the plastic strain within the specimen. The dark lines in these figures correspond to the region where a majority of nodes in the finite element mesh are concentrated. The three stages of the localization phenomenon, as reported by Marchand and Duffy (1988) based on their experimental observations of torsion tests on a HY-100 steel and by Wright and Walter (1987) who studied the simple shearing problem for a typical steel, but did not account for strain hardening and elastic effects, are most evident in the velocity plots of Fig. 2. The shear stress attains a maximum value of 3.65 at an average strain of 2.68, and the shear stress begins to drop noticeably when the average strain equals 3.37. The velocity field begins to show a sharp change in its slope within the central part of the specimen at an average strain of 3.4, and at an average strain of 4.0, the velocity variation within the block consists of three straight line segments. The parts of the block near the lower and upper surfaces move as a rigid body with the velocity of these surfaces, connected by a narrow thin central region within which the velocity changes sharply from nearly zero to almost one. The discontinuity in the velocity field across the shear band as asserted by Tresca (1878) and Massey (1921) corresponds in our computations to the severe increase in the speed of the material particles across the shear band because, in our work, the velocity field is forced to stay continuous throughout the region under study. From the velocity field plotted in Fig. 2, it is hard to decipher when it starts deviating from the linear variation. Recalling that the ends of the block are kept at a fixed environmental temperature, the solution surface for the temperature suggests that the temperature rise at the specimen center is somewhat discernible at an average strain of 2.7. Subsequently, the temperature difference between the central hotter region and the surrounding less hot material keeps on increasing, resulting in a very narrow central region of immensely heated material.

Figures 5, 6, and 7 show, respectively, the solution surfaces of velocity, shear stress, and the temperature rise for the tungsten alloy studied. The peak shear stress at the block center occurs at an average strain of 0.47%, and the shear stress does not become uniform through the thickness of the block. The shear stress at a point in the block gradually decreases till the average strain in the specimen equals 7%. Then the shear stress drops precipitously, and soon after that the computations become unstable in the sense that the spatial and temporal distributions of the shear stress show oscillations. The solution surface for the velocity shows that soon after the velocity on the top surface attains its steady value, the velocity distribution through the thickness of the specimen is no longer linear, implying thereby that the block is deforming nonhomogeneously. At an average strain of 0.08, the motion of the block can be divided into three parts, the nearly stationary lower portion, the narrow central region in which the velocity increases from essentially 0 to 1.0, and the top part moving virtually as a rigid block with the velocity prescribed on the top surface. The solution surface for the temperature rise suggests that the higher temperature at the specimen center relative to that at the surrounding points is noticeable at an average strain

of 3.5%. The difference between the temperature at the block center and the surrounding points continues to increase, and the slope of the curve depicting the temperature at the block center versus the average strain becomes quite steep at an average strain of 8%.

For the depleted uranium, we have plotted in Figures 8, 9, and 10, the solution surfaces for the velocity, the shear stress, and the temperature rise. The shear stress attains a maximum value at an average strain of 14.6%, and then decreases very slowly. The computations were stopped when the shear stress had dropped to 98.7% of its peak value, since the solution could not be computed to the same accuracy as for the other two materials. At an average strain of 18%, the velocity field begins to increase sharply in the central portion and the thickness of the central region starts decreasing. The temperature at the block center does not rise as rapidly as it did for the other two materials studied herein.

Figure 11 exhibits the deformed positions of a line element, initially straight in the unstressed configuration, when s/s_{max} = 1.0, 0.95, 0.85, 0.80, and 0.60 for the Armco IF iron, s/s_{max} = 1.0, 0.95, 0.90, 0.85, 0.80, and 0.75 for the tungsten alloy, and s/s_{max} = 1.0, 0.998, 0.996, 0.994, and 0.990 for the depleted uranium. Since the deformed position of the line even at s/s_{max} = 1.0 is not straight, the deformations of the block are nonhomogeneous at s/s_{max} = 1.0, probably due to the nonuniform thickness of the block. For Armco IF iron, the strain at the center of the block increases immensely once the shear stress there has dropped to 90 percent of its maximum value. For the tungsten alloy, the shear strain at the center when s/s_{max} = 0.85 is considerably lower than that for the Armco iron, and the strain at the center continues to increase as the shear stress there drops. For the depleted uranium, the deformation in the central region does not become excessive because the computations could not be carried far enough for the shear stress to drop to a large degree.

CONCLUSIONS. We have analyzed the initiation and growth of shear bands in Armco IF iron, tungsten alloy, and depleted uranium. The coupled nonlinear partial differential equations governing the thermomechanical deformations of a block of nonuniform thickness and undergoing overall simple shearing deformations were integrated by using the Gear method included in the package LSODE. Results for the depleted uranium could not be computed with the same accuracy as for the other two materials once the shear stress dropped to 99% of its maximum value. However, the deformation at the specimen center had begun to localize even at that instant. For the other two materials, sharp gradients of the deformation developed at the specimen center once the shear stress dropped to 90% of its peak value.

REFERENCES

1. Batra, R. C. and Kim, C. H., 1990, "Adiabatic Shear Banding in Elastic-Viscoplastic Nonpolar and Dipolar Materials", Int. J. Plasticity, Vol. 6, pp. 127-141.

2. Batra, R. C. and Kim, C. H., 1992, "Analysis of Shear Banding in Twelve Materials", Int. J. Plasticity, Vol. 8 (to appear).

3. Johnson, G. R. and Cook, W. H., 1983, "A Constitutive Model and Data for Metals Subjected to Large Strains, High Strain Rates and High Temperatures", Proc. 7th Int. Symp. Ballistics, The Hague, The Netherlands, pp. 1-7.

4. Johnson, G. R., Hoegfeldt, J. M., Lindholm, U. S., and Nagy, A., 1983, "Response of Various Metals to Large Torsional Strain Over a Large Range of Strain Rates, PartI: Ductile Metals, Part II: Less Ductile Metals", ASME J. Eng. Mat. Tech., Vol. 105, pp. 48-60.

5. Marchand, A. and Duffy, J., 1988, "An Experimental Study of the Formation Process of Adiabatic Shear Bands in a Structural Steel", J. Mechs. Phys. Solids, Vol. 36, pp. 251-283.

6. Massey, H. F., 1921, "The Flow of Metal During Forging", Proc. Manchester Assoc. Engineers, pp. 21-26.

7. Tresca, H., 1878, "On Further Application of the Flow of Solids", Proc. Inst. Mech. Engr., Vol. 30, 301.

8. Wright, T. W. and Walter, J. W., 1987, "On Stress Collapse in Adiabatic Shear Bands", J. Mechs. Phys. Solids, Vol. 35, pp. 701-720.

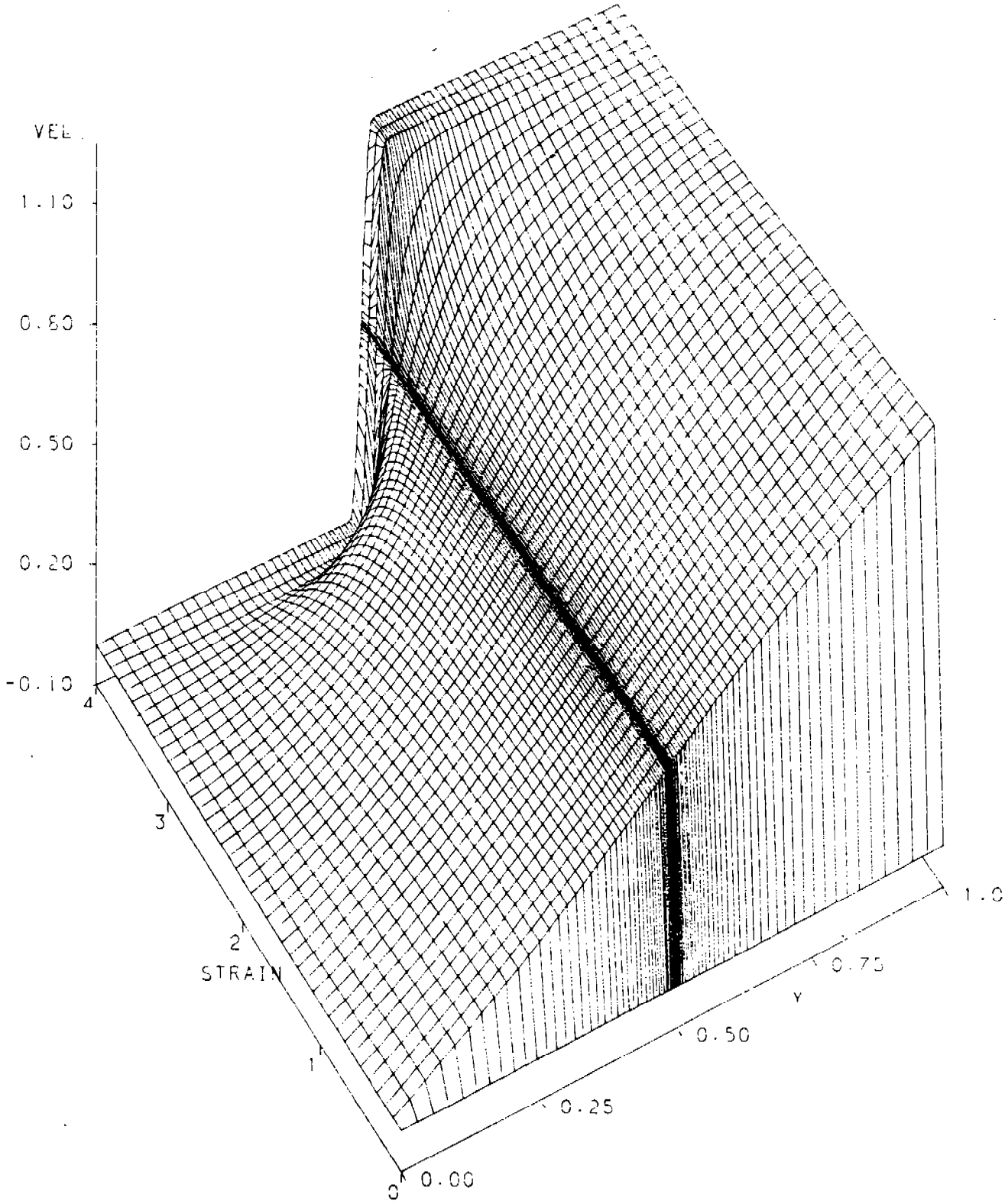


Figure 2. The solution surface for the velocity for the Armco IF iron block being sheared at a nominal strain-rate of 1500 $sec.^{-1}$

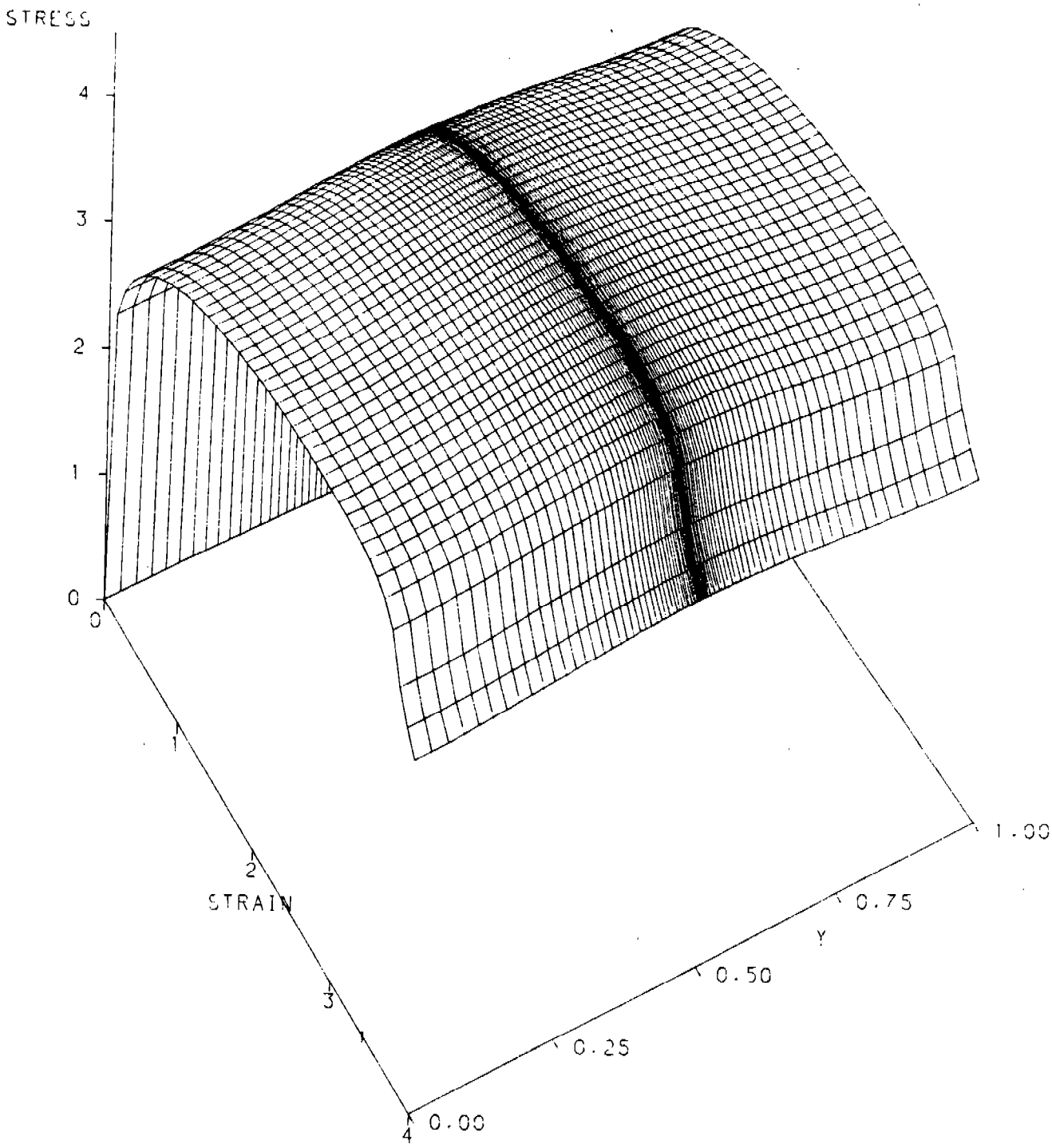


Figure 3. The solution surface for the shear stress for the Armco IF iron block being sheared at a nominal strain-rate of 1500 $sec.^{-1}$

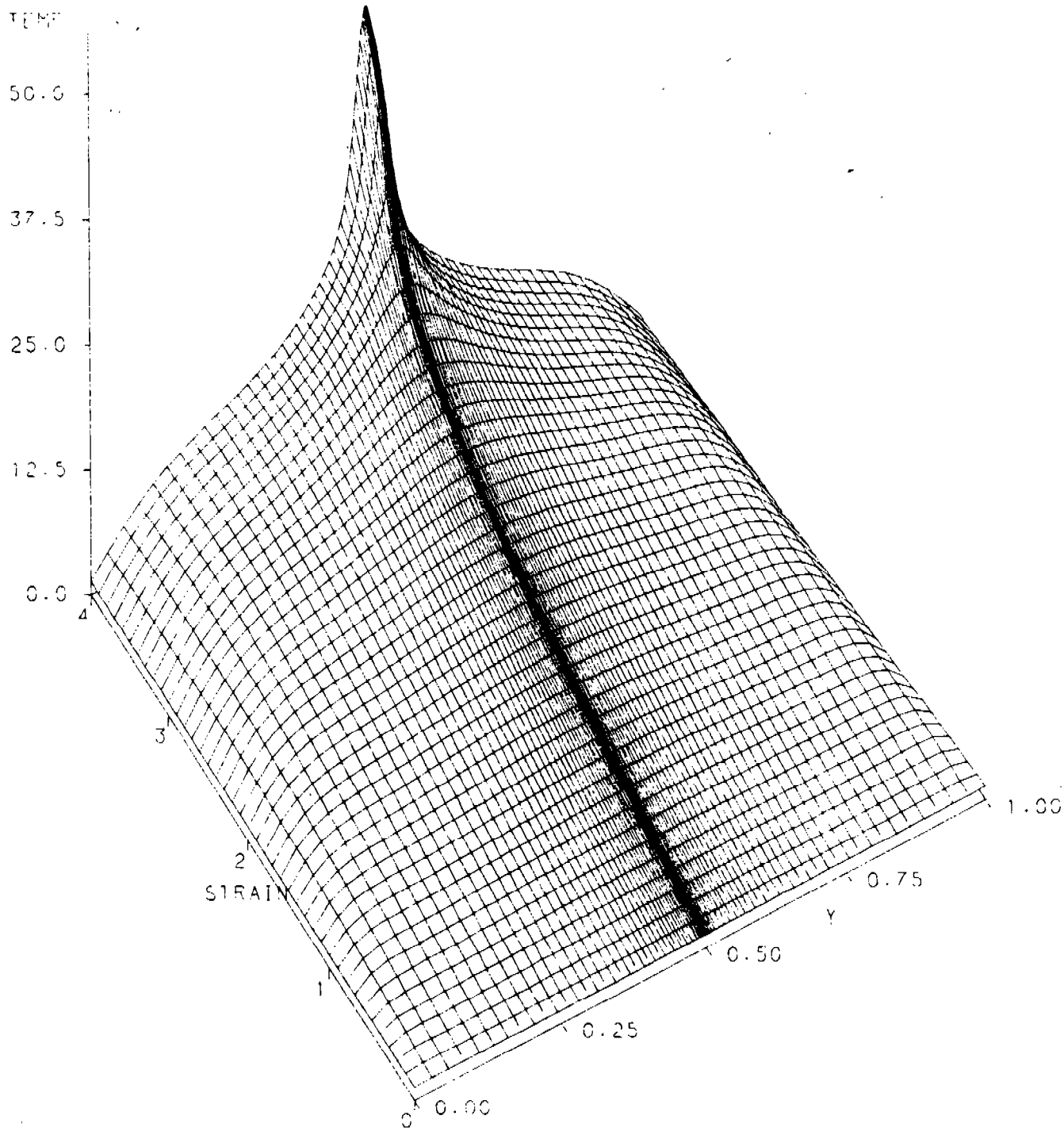


Figure 4. The solution surface for the temperature rise for the Armco IF iron block being sheared at a nominal strain-rate of 1500 sec.$^{-1}$

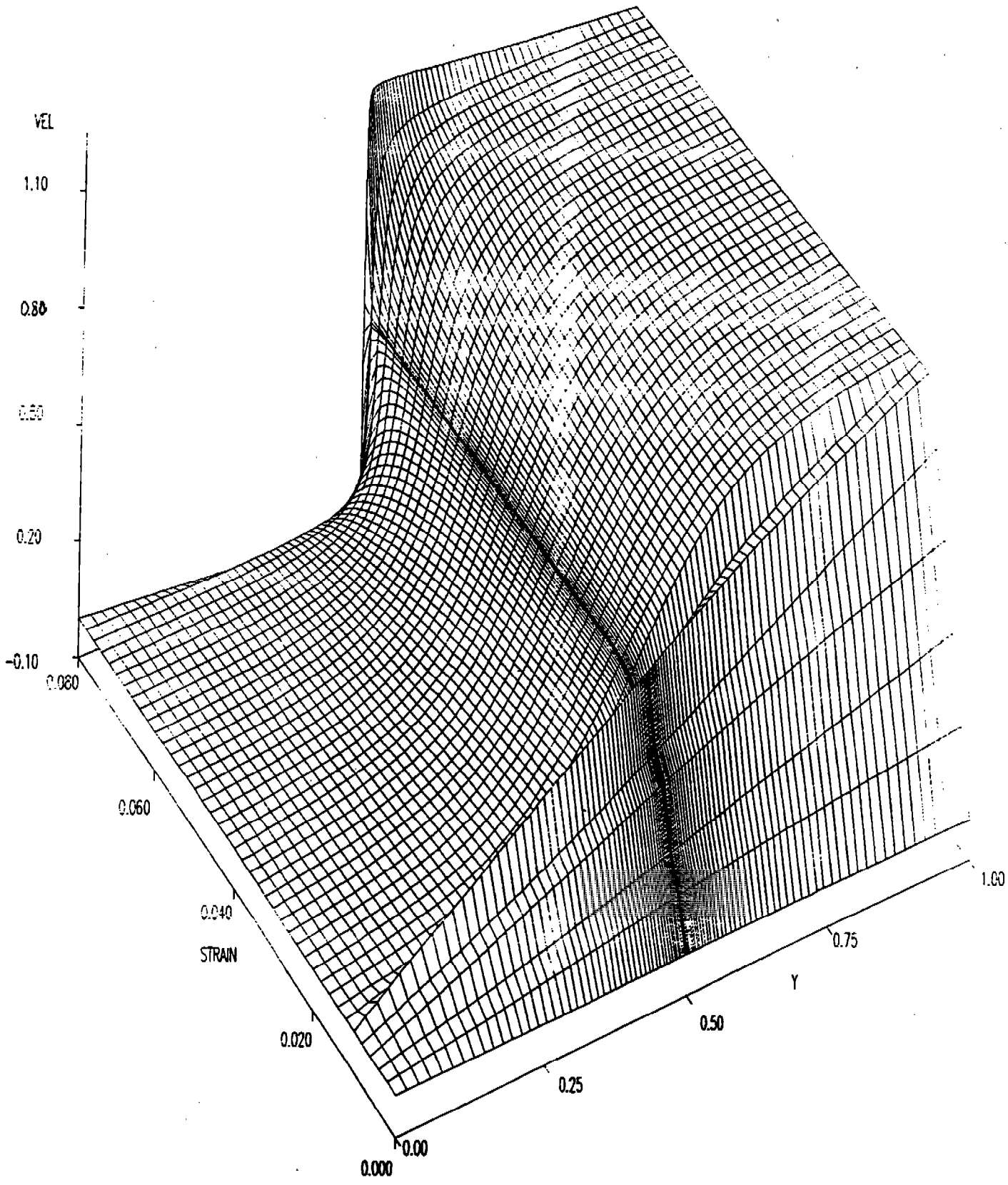


Figure 5. The solution surface for the velocity for the tungsten alloy block being sheared at a nominal strain-rate of 1500 sec.$^{-1}$

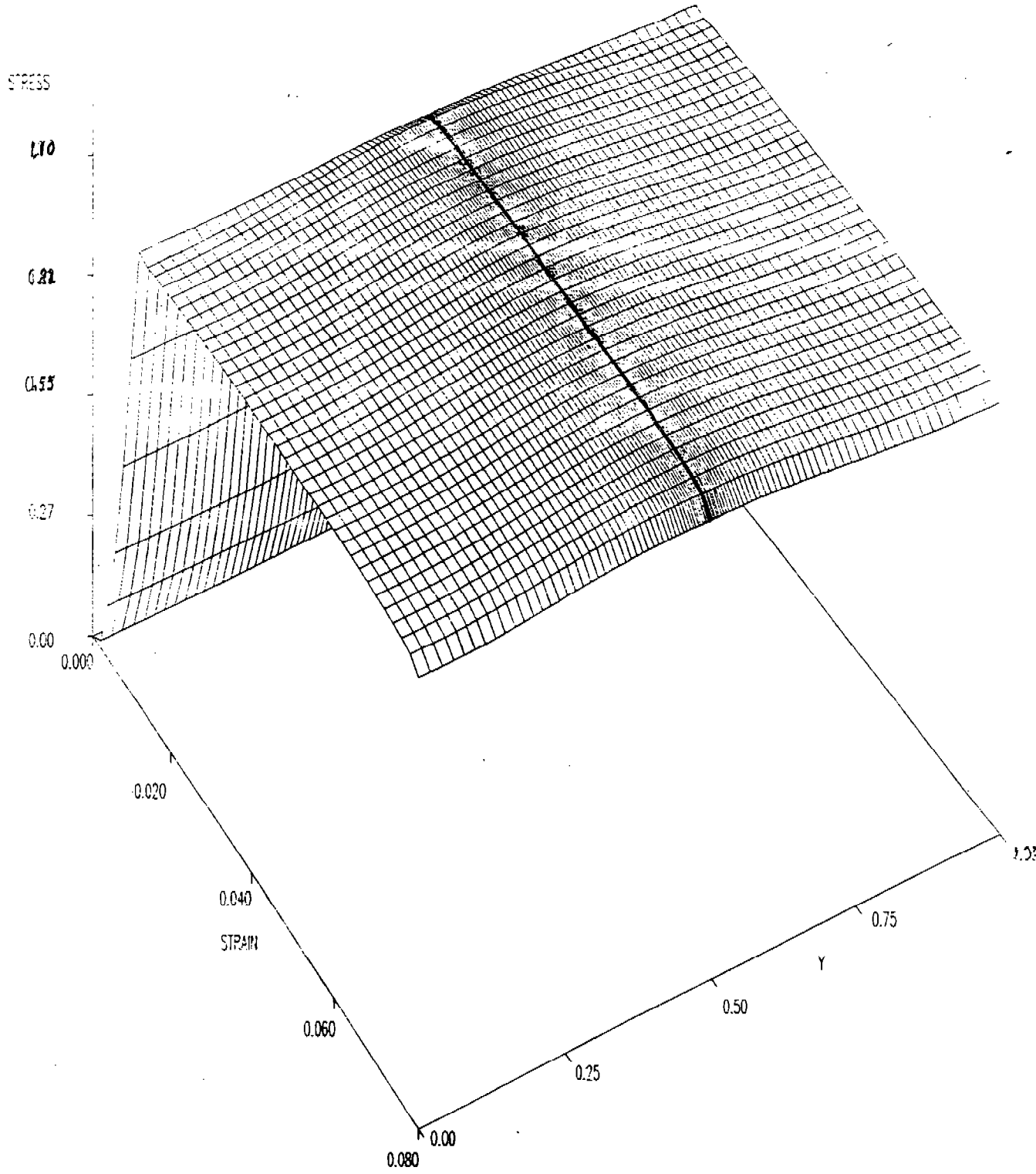


Figure 6. The solution surface for the shear stress for the tungsten alloy block being sheared at a nominal strain-rate of 1500 $sec.^{-1}$

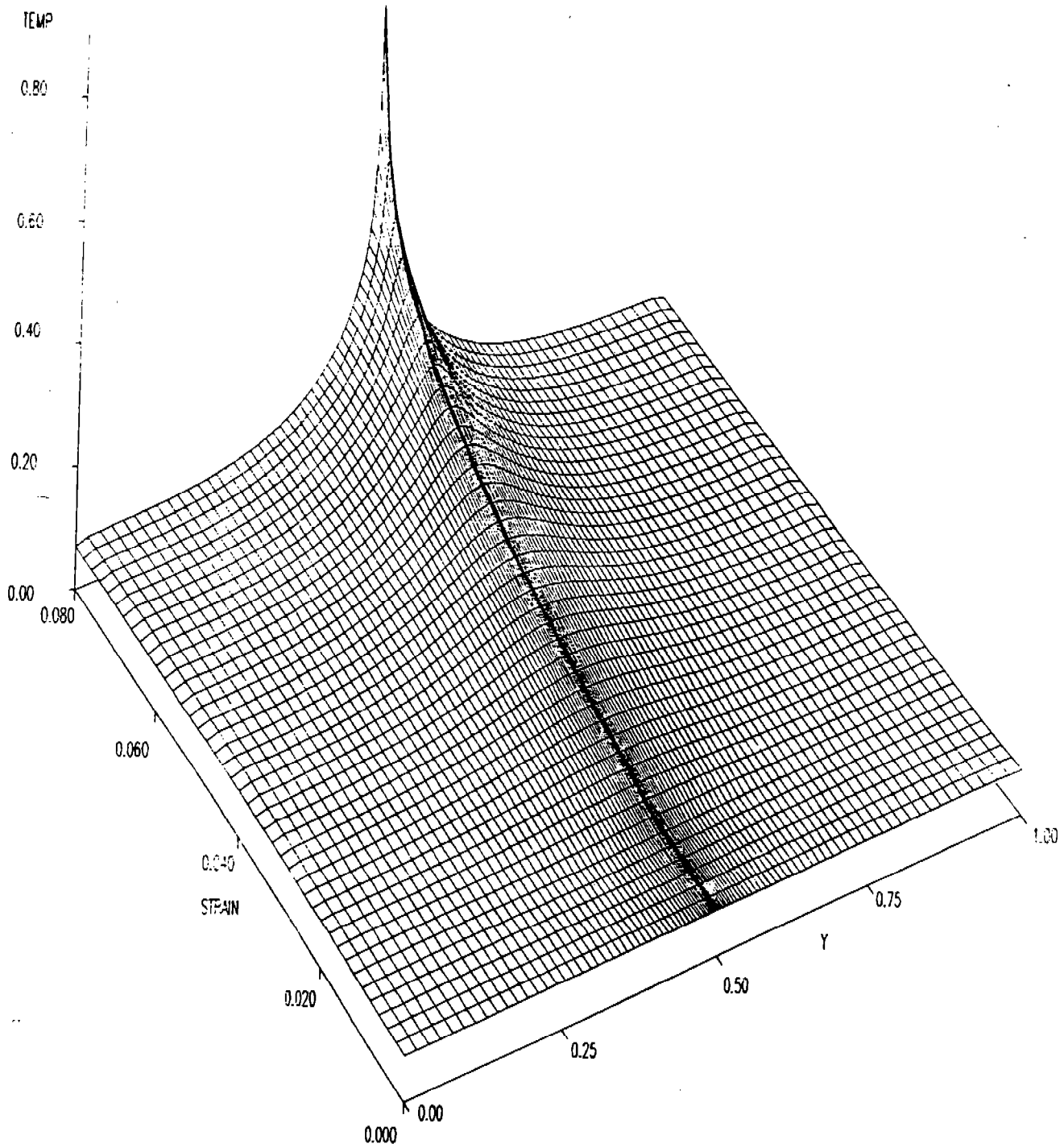


Figure 7. The solution surface for the temperature rise for the tungsten alloy block being sheared at a nominal strain-rate of 1500 sec.$^{-1}$

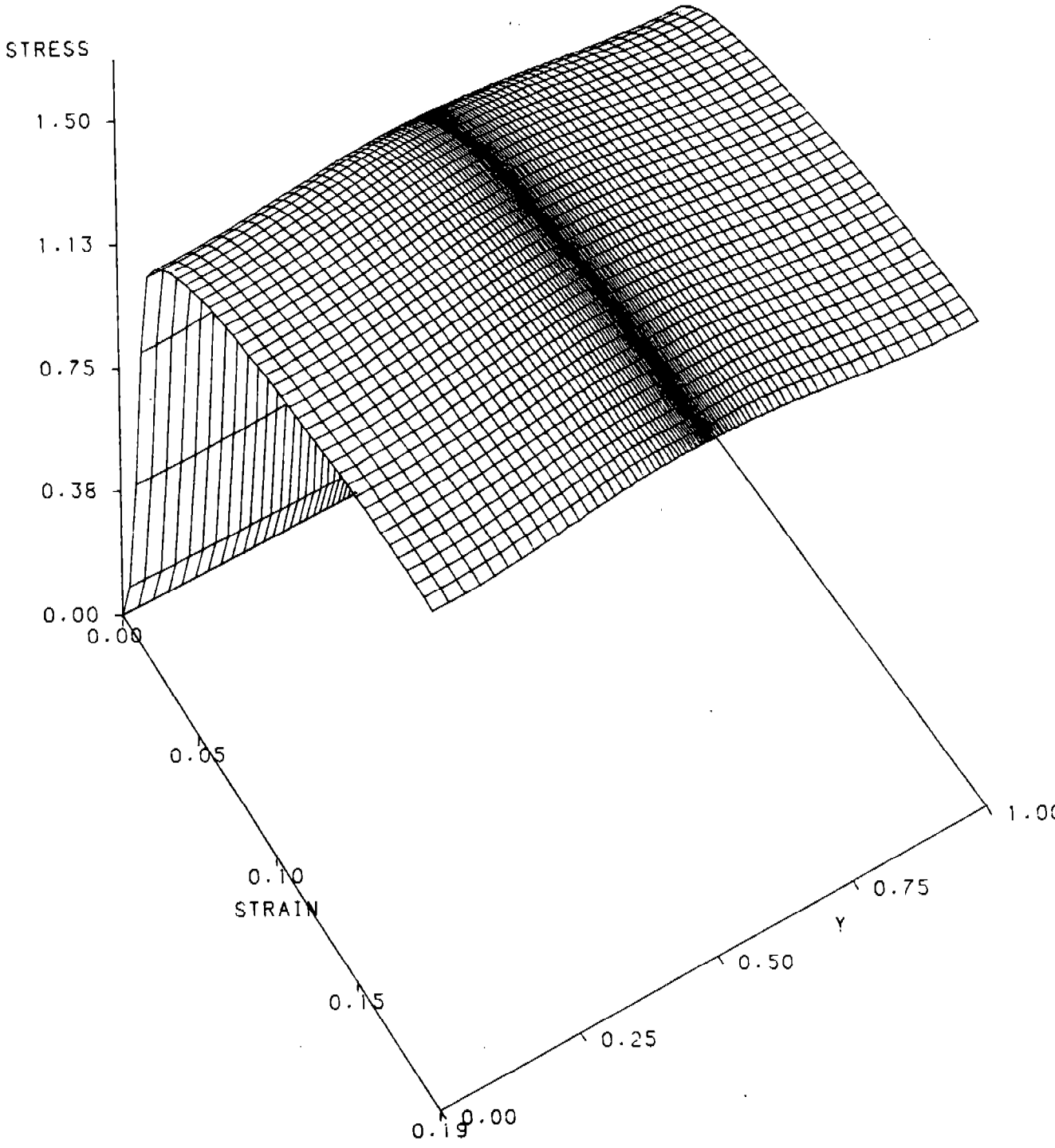


Figure 9. The solution surface for the shear stress for the depleted uranium block being sheared at a nominal strain-rate of 1500 $sec.^{-1}$

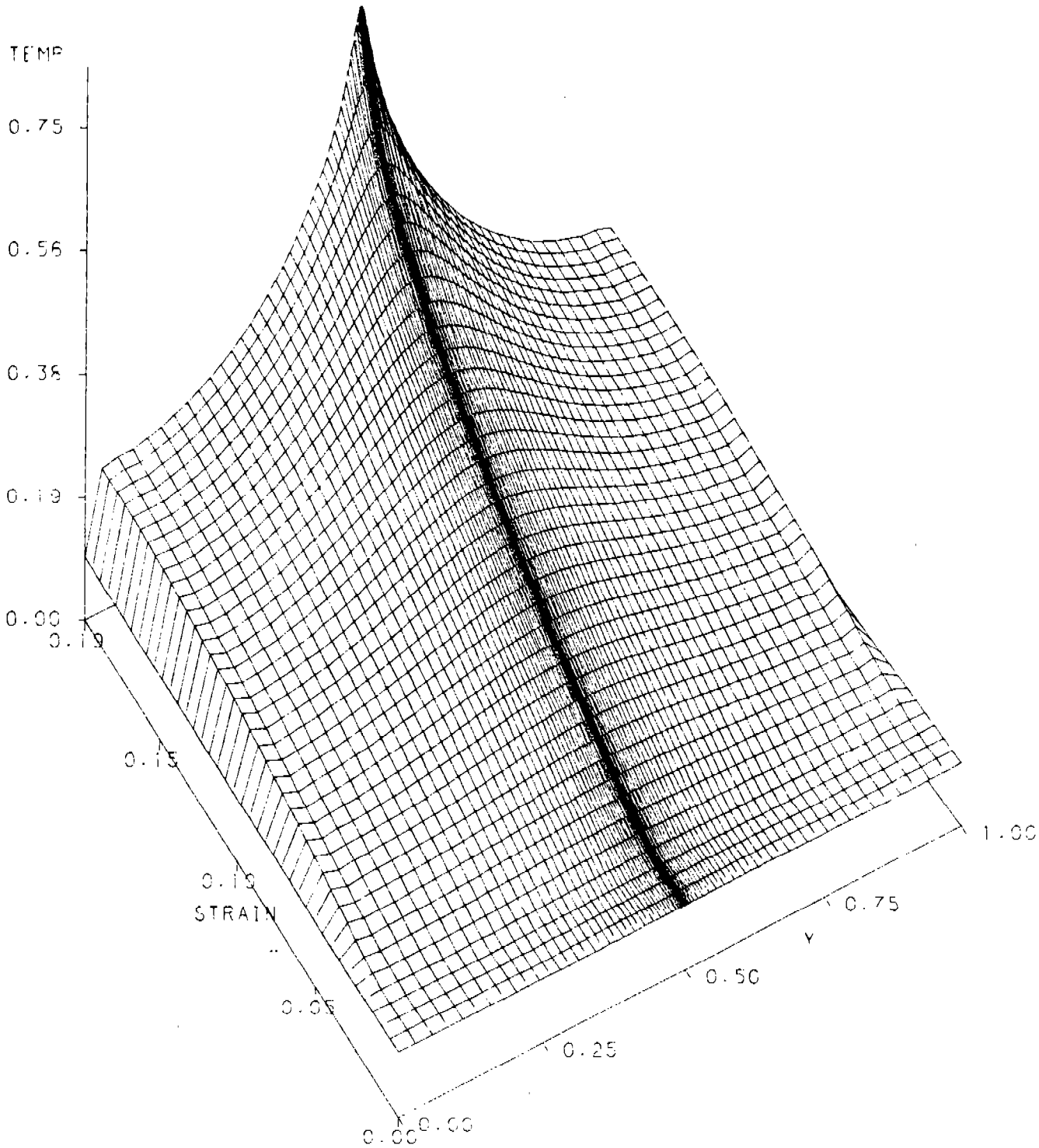


Figure 10. The solution surface for the temperature rise for the depleted uranium block being sheared at a nominal strain-rate of 1500 sec.$^{-1}$

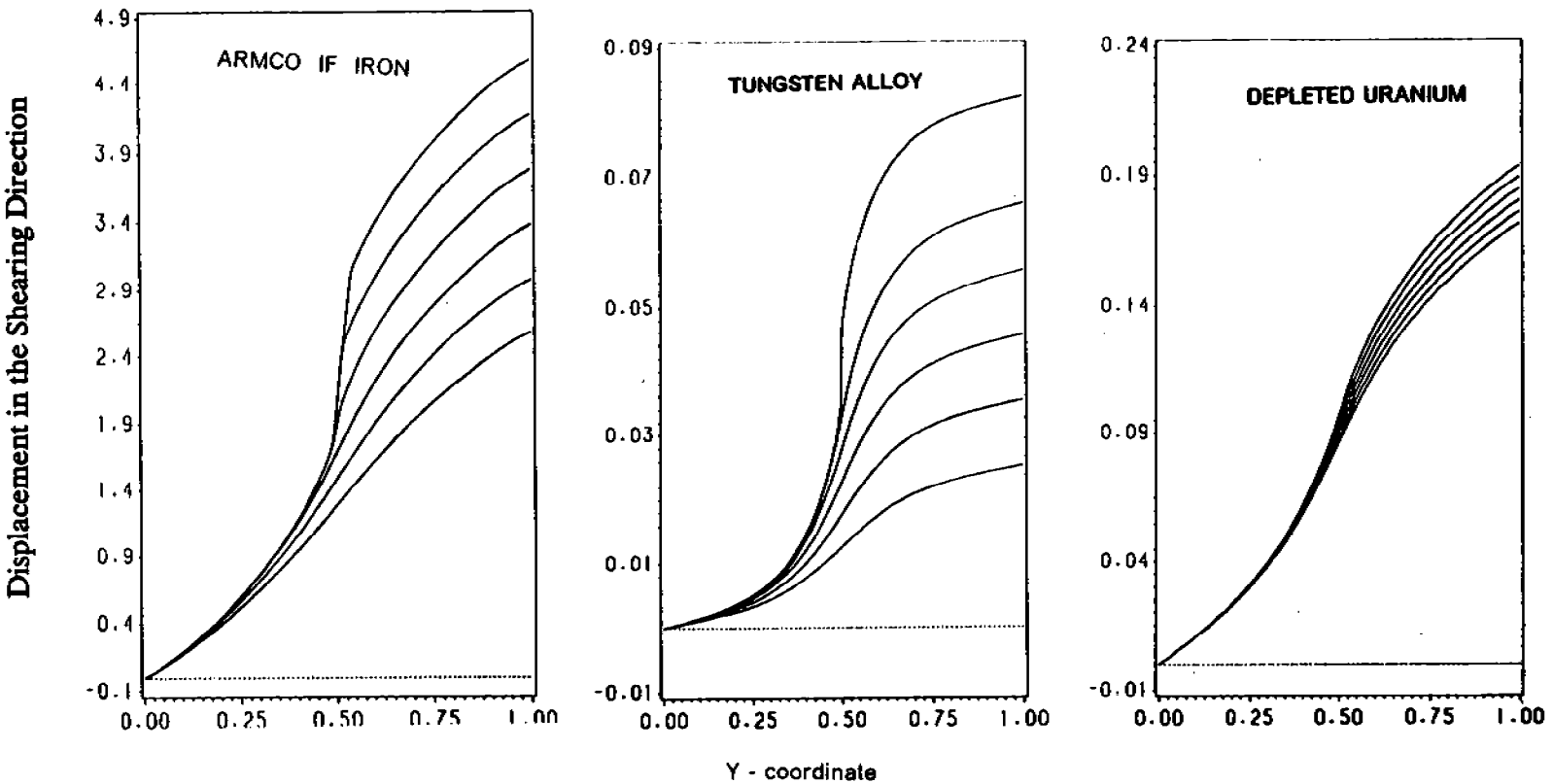


Figure 11. Deformed shapes of an initially straight line for (a) Armco IF iron when s/s_{max} = 1.0, 0.95, 0.90, 0.80, and 0.60; (b) Tungsten alloy at s/s_{max} = 1.0, 0.95, 0.90, 0.85, 0.80, and 0.75; (c) Depleted uranium at s/s_{max} = 1.0, 0.998, 0.996, 0.994, 0.992, and 0.990.

Analysis and Computation of Solutions to an Evolution Problem in Nonlinear Viscoelasticity

Donald A. French*
Department of Mathematical Sciences
University of Cincinnati
Cincinnati, OH 45221-0025

August 6, 1991

Abstract

Problems involving nonconvex energies where the equilibrium configuration may involve several phases have received alot of attention in recent years. We begin studying an evolution problem modeling the deformations of a simple viscoelastic material and a nonconvex energy. We show that the approximate solution given by a standard finite element method will converge at an optimal rate to the true solution. Through numerical computations we start to explore the long time behavior.

Introduction: We analyze and compute approximate solutions to the partial differential equation

$$u_{tt} - \Delta u_t - \nabla \cdot \left(\frac{\psi'(|\nabla u|)}{|\nabla u|} \nabla u\right) = f \text{ in } \Omega \times [0,T]$$

with

$$u = g \text{ on } \partial\Omega \times [0,T]$$

and

$$u(\cdot,0) = u_0, \quad u_t(\cdot,0) = v_0 \text{ on } \Omega.$$

This equation models antiplane shear deformations of an isotropic, homogeneous, incompressible, viscoelastic solid. The region Ω is the cross section

*Supported by the U.S. Army Research Office.

of a long tube; u represents the component of displacement in the direction perpendicular to Ω. Since the deformation is of antiplane shear type u depends only on $x \in \Omega$ and time t.

This short report summarizes results in French and Wahlbin [FW] where we assume that $g = 0$ and the initial data is smooth. We also make requirements on the growth of ψ. With these hypotheses the regularity theorems of Engler [E] apply as well as our approximation results.

Static Case: One hopes that as $u_t \to 0$ the solution will tend to a minimizer of

$$J(v) = \int_\Omega (\psi(|\nabla v|) - fv)dA \tag{1}$$

where $v = 0$ on $\partial\Omega$. The Euler-Lagrange equation for a critical point of J is

$$\nabla \cdot (\frac{\psi'(|\nabla v|)}{|\nabla v|} \nabla v) = f \text{ in } \Omega$$

which may not be well defined if v is not sufficiently smooth. We will consider ψ that are nonconvex (solid line) and relaxed or convexified (dashed line) (Figures 1 and 2).

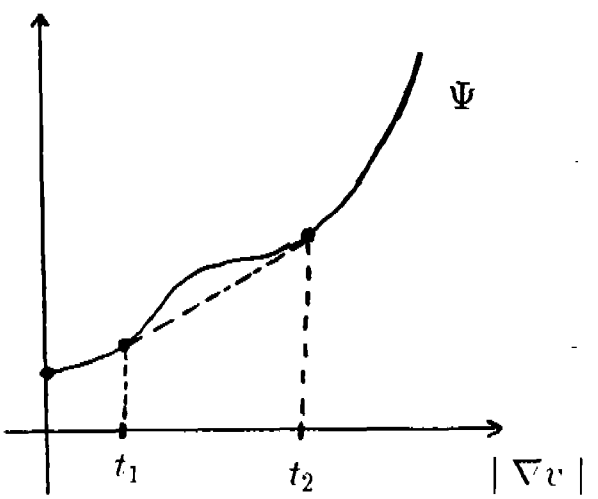


Figure 1

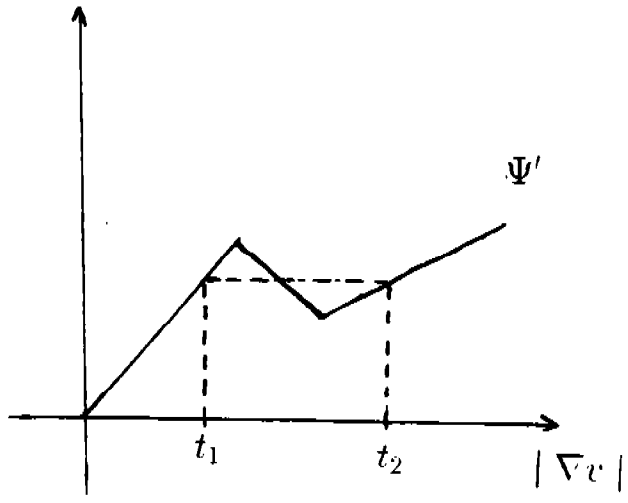


Figure 2

A theorem from Ekeland and Temam [ET] provides a connection between these two cases; if v is a solution of the nonconvex problem it is also a solution of the relaxed problem. If $\{v_n\}$ is a minimizing sequence for the nonconvex problem then for a subsequence $\{v_{n_j}\}$ we have weak convergence to a solution of the relaxed problem.

Bauman and Phillips [BP] show in a specific case that the nonconvex problem has no solution. This leads to an interesting question that we begin to explore in our computations: What does $u(\cdot, t)$ tend to as $t \to \infty$ in the nonconvex case?

We complete this section by noting some multigrid computations done on the convexified static problem by Goodman, Kohn, and Reyna [GKR]. Figure 3 is a surface plot of $|\nabla v_h|$ where v_h is their numerical solution. The '+' represent points where $t_1 \leq |\nabla v_h| \leq t_2$. Analysis of finite element approximations to the relaxed problem were done in [F].

Finite Element Method for the Evolution Problem: The approximate solution is sought in the finite dimensional space $S_h \subset H_0^1(\Omega)$ which we assume consists of piecewise polynomials of degree $\leq r-1$ on a quasi-uniform mesh where the diameter of the element domains is proportional to h.

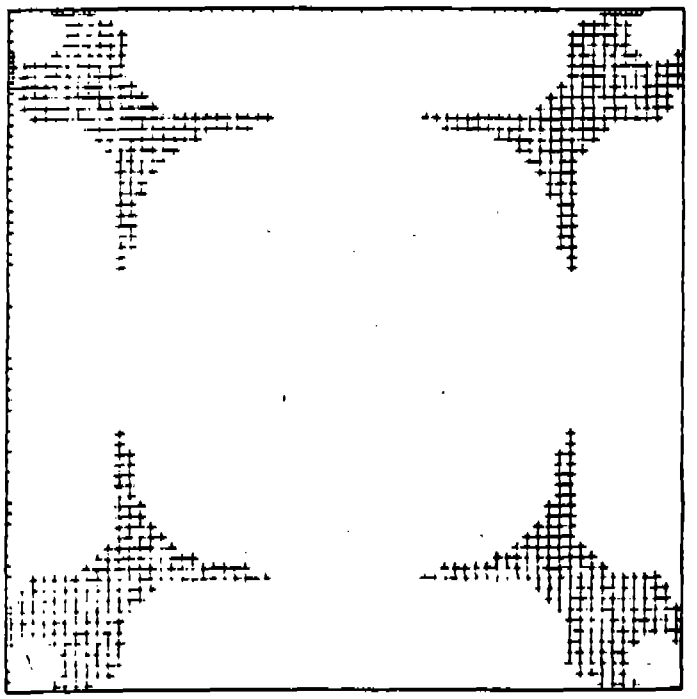


Figure 3

We analyzed the following semi-discrete finite element method: Find $u_h(\cdot,t) \in S_h$ for $t \in [0,T]$ such that

$$(u_{h,tt},\chi) + (\nabla u_{h,t},\nabla\chi) + (\frac{\psi'(|\nabla u_h|)}{|\nabla u_h|}\nabla u_h,\nabla\chi) = (f,\chi) \qquad (2)$$

for all $\chi \in S_h$ where $u_h(\cdot,0) \cong u_0$ and $u_{h,t}(\cdot,0) \cong u_1$.

The resulting system of ordinary differential equations has a unique global in time solution. We prove the following estimate concerning the accuracy of the method:

THEOREM: There exists a constant C independent of h such that

$$\| (u-u_h)(\cdot,t) \|_{L^2(\Omega)} \leq Ch^r$$

for $t \in [0,T]$.

REMARK: The constant C depends on norms of u and u_t as well as the final time T. We also assume that ψ is C^2 and therefore doesn't have the sharp corners displayed in figure 2.

Computations: We performed two numerical experiments using the method (2) which we discretized in time by a second order energy preserving scheme (See [FS] and [CHMM]) which has been labeled a generalized Crank-Nicolson scheme. We used fixed point iteration to solve the nonlinear systems on each time step and preconditioned conjugate gradients to solve the linear systems. In the first experiment the nonlinear function ψ is chosen to be the same as the one used in [GKR] (Dashed line figures 1 and 2.). In the second we formed a nonconvex energy from the [GKR] energy (Solid line figures 1 and 2). In each computation we tracked $\| u_h(\cdot, t) \|_{L^2(\Omega)}$. In both cases it first increased quickly then decreased, tending to zero asymptotically. We stopped the computation when the norm was small. Figures 4 and 5 have the steady state surface plots of $| \nabla u_h |$ for the first and second experiments.

Notice the plot from the relaxed case is very similar to the plot from [GKR].

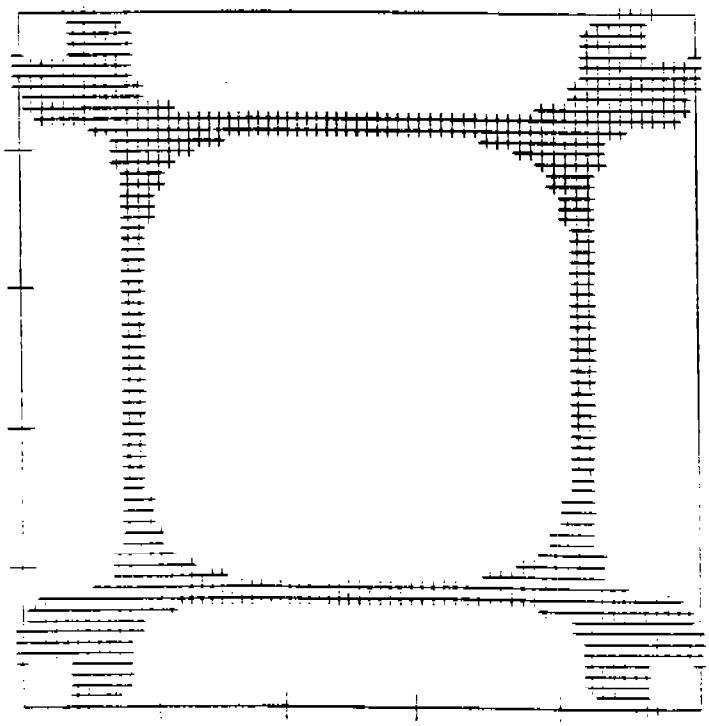


Figure 4: Relaxed Case

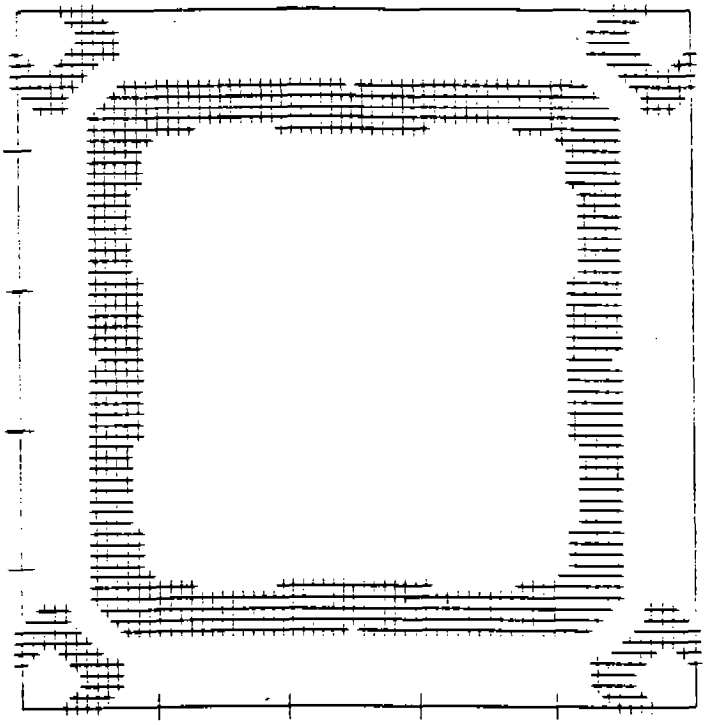


Figure 5: Nonconvex Case

References

[BP] P. Bauman and D. Phillips, A nonconvex variational problem related to change of phase, Appl. Math. Opt. 21 (1990), 113-138.

[CHMM] A.J. Chorin, J.J.R. Hughes, M.E. McCracken, and J.E. Marsden, Product formulas and numerical algorithms, Comm. Pure Appl. Math. 31 (1978), 205-256.

[ET] J. Ekeland and R. Temam, Convex Analysis and Variational Problems, North Holland, 1976.

[E] H. Engler, Global regular solutions of the dynamic anitplane shear problem in nonlinear viscoelasticity, Math Zeitschriff. 20 (1989), 251-259.

[F] D.A. French, On the convergence of finite element approximations of a relaxed variational problem, SIAM J. Num. Anal., 27 (1990), 419-436.

[FS] D.A. French and J.W. Schaeffer, Continuous finite element methods which preserve energy properties for nonlinear problems, Appl. Math. Comp., 39 (1990), 271-295.

[FW] D.A. French and L.B. Wahlbin, Analysis and computation of solutions to an evolution problem in nonlinear viscoelasticity (In preparation).

[GKR] J. Goodman, R.V. Kohn, and L. Reyna, Numerical study of a relaxed variational problem from optimal design, Comp. Meth. Appl. Mech. Eng., 57 (1986), 107-127.

NUMERICAL MODELLING OF MODE I LINEAR VISCOELASTIC FRACTURE

M.K. Warby*, J.R. Walton† and J.R. Whiteman*

* BICOM, Institute of Computational Mathematics,
Brunel University, Uxbridge, Middlesex, England.
† Department of Mathematics, Texas A & M University,
College Station, Texas, USA

ABSTRACT

Numerical schemes, based on finite elements in space and finite differences in time, are outlined for modelling stationary and moving cracks in two-dimensional linear viscoelastic materials. The mathematical formulation involves a linear single time integral constitutive model for the viscoelastic behaviour, together with the inclusion of a Barenblatt failure zone in the viscinity of the crack tip. Crack propagation and subsequent motion is based upon a crack opening displacement criterion (COD).

1. INTRODUCTION

We give here an outline of a model for predicting the onset of fracture and for following a propagating crack for Mode I planar fracture problems involving viscoelastic solids. The mathematical model of the deformation assumes a quasi-static linear viscoelastic response, constant Poisson's ratio and small deformation strains. The discretisation of this mathematical model is achieved using finite elements in space and finite differences in time, and with this approximations to the deformation resulting from given external loadings are calculated. The fracture of the viscoelastic material is modelled by incorporating a failure zone of the Barenblatt type about the crack tip, see Barenblatt [1], Knauss [3] and Schapery [5]. The purpose of the failure zone is to include in the model some representation of the cohesive forces and the local damage which occurs in the neighbourhood of the crack tip. It is assumed that there is small scale yielding at the crack tip, and constant stress in the failure zone. A further assumption, necessary to the validity of the Barenblatt concept, is that the material is free of voids. The state of crack propagation is determined using a Crack Opening Displacement (COD) criterion.

The model and finite element/finite difference scheme have been applied to various Mode I viscoelastic fracture problems. Space limitations here determine that only an outline of the scheme and a synopsis of the numerical results can be presented; more complete details and more extensive results can be found in Warby et al. [7]. A theoretical error analysis, together with error estimates, for the approximations to the deformation of the linear viscoelastic solid, can be found in Shaw et al. [6].

The work of Walton was supported in part by the United States Air Force Office of Scientific Research and the National Science Foundation through NSF Grant No. DMS-8903672.

The work of Whiteman was similarly supported in part by the United States Army Research, Development and Standardization Group, London under Contract No. DAJA45-89-C-003. All this support is gratefully acknowledged.

In the present work consideration has been limited to linear single integral constitutive relations, appropriate to **standard linear solids.** This type of constitutive equation may be regarded as a first approximation to the constitutive equations required for modelling realistically the behaviour of more general isotropic viscoelastic solids under isothermal conditions. Clearly the choice of the linear single integral relation restricts the range of materials to which the model may be applied, see e.g. Christensen [2]. However, for the relevant class of materials the algorithm is able to predict the critical states prior to crack propagation and the form of crack growth subsequent to this. Our intention in subsequent work is to track the growth of cracks for a range of materials and to incorporate nonlinear single integral, or multi-integral constitutive relations into the model.

2. MODEL OF VISCOELASTIC DEFORMATION

2.1 Equilibrium Problem and Weak Formulation

We consider the deformation of a solid body defined in a region $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega$, under the action of external forces. The displacement at the point $\mathbf{x} \equiv (x_1,x_2)^T \in \Omega$ (the reference configuration) for time $t \in (0,T] \equiv J$ is $\mathbf{u} \equiv \mathbf{u}(\mathbf{x},t) = (u_1(\mathbf{x},t),u_2(\mathbf{x},t))^T$. The stress and strain tensor components are denoted respectively by σ_{ij}, ε_{ij}, $1 \le i,j \le 2$.

From the law of conservation of momentum (for the case where there is no acceleration) the deformation $\mathbf{u}$ of the body under the action of external forces $\mathbf{f} \equiv (f_1, f_2)^T$ and boundary tractions $\mathbf{g} \equiv (g_1, g_2)^T$, at time t satisfies

$$\sum_{i=1}^{2} \frac{\partial \sigma_{ij}(\mathbf{x},t)}{\partial x_j} + f_i(\mathbf{x},t) = 0 \ , \quad \mathbf{x} \in \Omega, \ t \in J, \ i = 1,2 \ , \tag{2.1}$$

together with boundary conditions

$$\mathbf{u}(\mathbf{x},t) = 0 \ , \quad \mathbf{x} \in \partial\Omega_c \ , \ t \in J \ , \tag{2.2}$$

$$\sum_{j=1}^{2} \sigma_{ij}(\mathbf{x},t) \cdot n_j = g_i(\mathbf{x}) \ , \ \mathbf{x} \in \partial\Omega_T \ , \ t \in J \ , \tag{2.3}$$

where $\partial\Omega \equiv \partial\Omega_c \cup \partial\Omega_T$ and $\mathbf{n} \equiv (n_1, n_2)^T$ is the outward normal unit vector to $\partial\Omega$ at any point.

With the usual Sobolev space notation we specify the space V of functions defined over Ω as

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

If $\mathbf{u}(\mathbf{x},t) \in V \times H^1(J)$ then the weak form of (2.1) - (2.3) is obtained by multiplying (2.1) by $\mathbf{v}(\mathbf{x}) \in V$ and integrating over Ω. Thus in the weak problem we seek $\mathbf{u}(\mathbf{x},t) \in V \times H^1(J)$ such that

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\int_{\Omega}\sigma_{ij}(u(x,t))\varepsilon_{ij}(v(x))dx = \sum_{i=1}^{2}\int_{\Omega}f_i(x,t)v_i(x)dx + \sum_{i=1}^{2}\int_{\partial\Omega_T}g_i(x)v_i(x)ds\ ,$$

(2.4) $\qquad \forall\quad v(x)\in V\ ,\ t\in J\ ,$

where the tensor components ε_{ij} are defined by

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

and the vectors ε and σ are given by $\varepsilon \equiv (\varepsilon_{11},\varepsilon_{22},2\varepsilon_{12})$ and $\sigma \equiv (\sigma_{11},\sigma_{22},\sigma_{12})$.

The involvement of the displacement $u(x,t)$ in (2.4) requires a constitutive relation. For the case of linear viscoelasticity considered here, under the assumption that there is no deformation for time $\tau < 0$, we take the constitutive relation to be

(2.5) $$\sigma(t) = \int_0^t D(t-\tau)\frac{\partial\varepsilon(\tau)}{\partial\tau}\,d\tau\ .$$

where

(2.6) $$D(t-\tau) \equiv D_0\phi(t-\tau)$$

is the 3×3 stress relaxation matrix of the viscoelastic material, $\phi(t-\tau)$ is the stress relaxation function and D_0 is a 3×3 matrix involving the Lamé coefficients associated with the instantaneous elastic response of the material.

Using the vectors σ and ε as defined above, we may write (2.4) as

$$\int_{\Omega}\sigma(u(x,\tau)\ ;\ \tau\le t)\ \cdot\varepsilon(v)dx = \int_{\Omega}f(t)\ \cdot v\,dx + \int_{\partial\Omega_T}g(t)\ \cdot v\ ds\ ,$$

(2.7) $\qquad \forall\cdot\quad v\in V\ ,\ t\in J\ ,$

and substitution of (2.5) for σ in (2.7) gives

$$\int_0^t\int_{\Omega}\varepsilon(v)\ \cdot D(t-\tau)\dot{\varepsilon}(\tau)dx\,d\tau = \int_{\Omega}f(t)\ \cdot v\,dx + \int_{\partial\Omega_T}g(t)\ \cdot v\ ds\ ,$$

(2.8) $\qquad \forall\quad v\in V\ ,\ t\in J\ .$

2.2 Numerical Discretisation

The numerical algorithm to be applied to (2.8) is based on finite elements in space and finite differences in time. Thus for any time $t \in J$ the region Ω is partitioned into elements Ω^e such that $\Omega = \cup \Omega^e$ and a finite dimensional space $S^h \subset V$ consisting of piecewise polynomial functions defined over the partition is set up. We first produce the semi-discrete form of (2.8) by defining the approximation to $\mathbf{u}(x,t)$

$$\tilde{\mathbf{u}}_h(x,t) = \mathbf{N}(x)\tilde{\mathbf{U}}(t) \tag{2.9}$$

where $\mathbf{N}(x)$ is the $2\times 2n$ matrix involving the usual spatial finite element basis functions for the n nodes of the piecewise polynomial approximation over Ω and $\tilde{\mathbf{U}}(t)$ is the vector of (nodal) functions associated with each node. In the usual way, see e.g. Zienkiewicz [8], we may define

$$\tilde{\boldsymbol{\varepsilon}}_h = \boldsymbol{\varepsilon}(\tilde{\mathbf{u}}_h(x,t)) = \mathbf{B}(x)\tilde{\mathbf{U}}(t)$$

and the corresponding approximation to the stress vector using (2.5). The semi-discrete form of (2.8) is thus the system

$$\int_0^t \left(\int_\Omega \mathbf{B}^T \mathbf{D}(t-\tau)\mathbf{B}\, dx \right) \dot{\tilde{\mathbf{U}}}(\tau) d\tau = \int_\Omega \mathbf{f}(t)^T \cdot \mathbf{N} dx + \int_{\partial\Omega^T} \mathbf{g}(t)^T \cdot \mathbf{N} ds \ . \tag{2.10}$$

The final discrete form of (2.8) is obtained by taking time levels t_j, $j = 0,1,2,\ldots$ and for $t_{j-1} < \tau < t_j$ approximating $\dot{\tilde{\mathbf{U}}}(\tau)$ by $(\mathbf{U}(t_j) - \mathbf{U}(t_{j-1}))/(t_j - t_{j-1})$ so that

$$\left(\int_\Omega \mathbf{B}^T \int_{t_{j-1}}^{t_j} \frac{\mathbf{D}(t_j-\tau)d\tau}{(t_j-t_{j-1})} \mathbf{B} dx \right) \left(\mathbf{U}(t_j) - \mathbf{U}(t_{j-1}) \right)$$

$$= \int_\Omega \mathbf{f}(t)^T \cdot \mathbf{N} dx + \int_{\partial\Omega_T} \mathbf{g}(t)^T \cdot \mathbf{N} ds - \sum_{q=1}^{j-1} \int_\Omega \mathbf{B}^T \left(\int_{t_{q-1}}^{t_q} \frac{\mathbf{D}(t_j-\tau)d\tau}{t_q - t_{q-1}} \mathbf{B}\, dx \right) \left(\mathbf{U}(t_q) - \mathbf{U}(t_{q-1}) \right) \ . \tag{2.11}$$

The system (2.11) is solved for $\mathbf{U}(t_j)$ which gives the nodal approximation to $\mathbf{u}(x,t_j)$.

It can be seen that the history of the deformation is retained throughout the time stepping process by the summation $q = 1,2,\ldots,j-1$. The effect of this depends on the form of the stress relaxation function $\phi(s)$ in (2.6).

3. MODE I VISCOELASTIC FRACTURE

3.1 Linear Elastic Fracture

Let us consider the Mode I linear elastic fracture problem as in Fig. 1, where the external loadings L_e are applied to the crack faces at distances remote from the crack tip. For problems of this type the strength of the stress singularity local to the crack tip is given by the stress intensity factory K. This factor can be calculated from the path independent J-integral, see Rice [4], which is defined as

$$J \equiv \int_{\Gamma} \left(W \, dx_2 - \sum_{i=1}^{2} T_i \frac{\partial u_i}{\partial x_1} ds \right), \tag{3.1}$$

where Γ is a contour running anticlockwise from the lower to the upper crack faces enclosing the crack tip, W is the strain energy density and the T_i are the components of the outward normal traction to Γ. For plane stress and plain strain problems of this type $J \sim K^2$, so that approximations to K may be obtained from approximations to J.

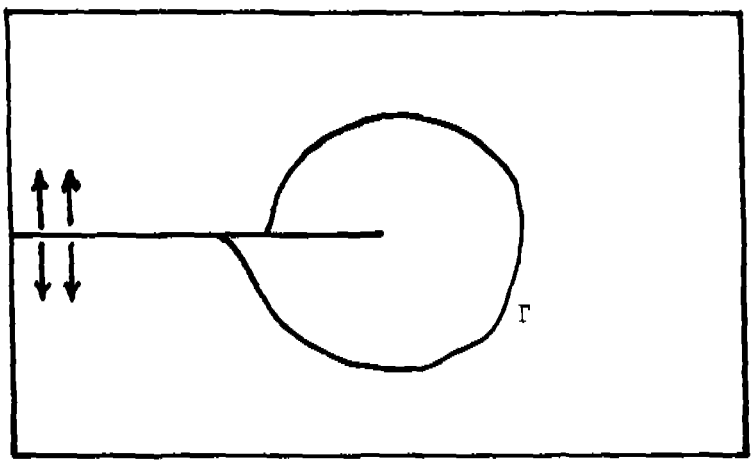


Fig. 1.

3.2 Correspondence Principle

For viscoelastic materials of the type described in Section 2 it is possible at time t to relate the stresses and strains of problem (2.1) - (2.3) to the stresses σ^R and strains ε^R of a related reference elastic problem, see Schapery [5]. In particular, if for the Mode I fracture problem the elastic body is subject to the same boundary conditions as the viscoelastic problem at time t, then $\sigma(x,t) = \sigma^R(x)$ and hence $K = K^R$, where K^R is the stress intensity factor of the elastic problem. Thus K can be obtained from J^R, the J-integral for the reference elastic problem.

3.3 Failure Zone

In order to give a more physically realistic model of the fracture of the viscoelastic material, we introduce a Barenblatt failure zone, see [1]. The mathematical concept of this is that in a small zone of length a_f behind the crack tip, see Fig. 2a, cohesive stresses L_f are applied normal to the crack faces in order to cancel the stress singularity at the crack tip produced by the external loads L_e applied to the body, see Fig. 2b. In the model the cohesive stresses L_f are assumed to be constant and the length a_f is determined as follows.

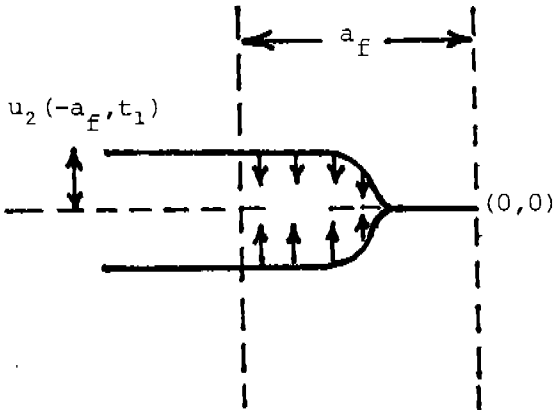


Fig. 2a

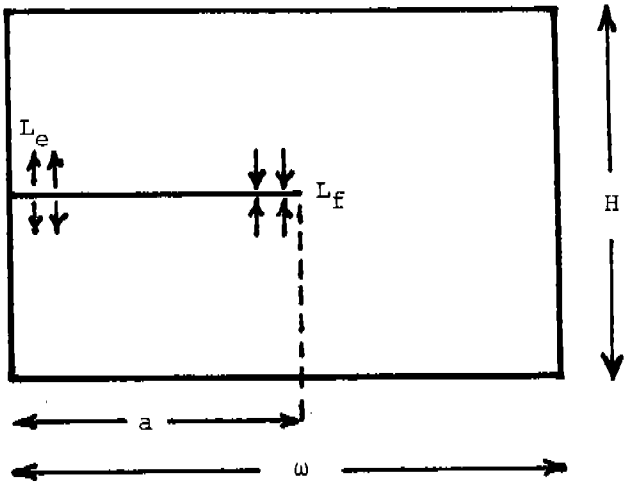


Fig. 2b

At time t let $K^{Re}(t)$ and $K^{Rf}(a_f,t)$ denote the stress intensity factors associated with the two reference elastic problems in which the external load L_e and the cohesive stresses L_f are applied separately. The length a_f is then determined from

$$K^{Re}(t) + K^{Rf}(a_f,t) = 0 \ ,$$

or equivalently from

$$(K^{Re}(t))^2 - (K^{Rf}(a_f,t))^2 = 0 \ ,$$

or equivalently by solving

(3.2) $$g_1(a_f,t) = 0 \ ,$$

where

(3.3) $$g_1(a_f,t) \equiv \left\{1 - \left(\frac{J^{Rf} - 2L_f\, U_2^{Rf}(-a_f,t)}{J^{Re}}\right)\right\} .$$

In (3.3) J^{Re} and J^{Rf} are the J-integrals for the reference elastic problems respectively with external load L_e and failure load L_f, and $U_2^{Rf}(-a_f,t)$ is the vertical displacement at the end of the failure zone due to the failure load: Equation (3.2) is used to determine a_f. (It should be observed that the displacements due to the failure load are non-physical as they correspond to a situation in which material near the top and bottom crack faces of the crack occupies the same physical position. A physical solution is obtained when the displacement fields due to the external and failure loads are combined.)

3.4 Algorithm for Viscoelastic Fracture

The following additional notation is adopted.

$U^e(t)$, $U^f(t,a_f)$ = calculated nodal viscoelastic displacements due respectively to external load L_e and to failure load L_f applied on an interval $(-a_f,0)$ at time t.

$U^{Re}(t)$, $U^{Rf}(t,a_f)$ = corresponding nodal reference elastic displacements,

$U_2(-a_f,t)$ = calculated vertical viscoelastic displacement at the end of the failure zone,

U_2^{cr} = critical crack opening displacement (COD) at $(-a_f,0)$,

(3.4) $$g_2(a_f,t) \equiv U_2(-a_f,t) - U_2^{cr}.$$

For the given geometry and crack length a value of a_f is chosen and cohesive forces L_f are applied to the corresponding failure zone.

Step 1 : Stationary Crack

Time steps $t_1, t_2, \ldots$

For time t_1 solve two viscoelastic problems.

(a) with external load L_e, to obtain $U^e(t_1,a_f)$

(b) with failure load L_f, to obtain $U^f(t_1,a_f)$.

In order to adjust the a_f use the numerical solutions U^{Re} and U^{Rf} of the corresponding reference elastic problems to calculate

$$J^{Re}\ ,\ J^{Rf}$$

Solve : $g_1(a_f,t_1) = 0$ for a_f

With this value of a_f

Test : $g_2(a_f,t) \lessgtr 0$.

If $g_2(a_f,t) < 0$, then increment the time and repeat the step. Otherwise go to Step 2.

Step 2 : Crack Initiation

From Step 1 we have that $g_2(a_f,t_{i-1}) < 0 \leq g_2(a_f,t_i)$.

Solve :

$$\begin{cases} g_1(a_f,t_{cr}) = 0 \\ \\ g_2(a_f,t_{cr}) = 0 \end{cases} \tag{3.5}$$

for the time t_{cr} of crack initiation and the failure zone length at that time. Again we must point out that the evaluation of g_1 and g_2 involves the determination of relevant viscoelastic and elastic displacement fields. When a solution of (3.5) is obtained, we reset $t_i = t_{cr}$ and proceed to Step 3.

Step 3 : Crack Propagation

Increment the crack length to a_i

Attempt to solve the system

$$\begin{cases} g_1(a_f,t_i) = 0 \\ \\ g_2(a_f,t_i) = 0 \end{cases} \tag{3.6}$$

for the time t_i at which the crack length a_i is attained (assuming that it is), and the failure zone length at that time. The details of the manner in which the finite element mesh is moved to correspond to the new crack tip positions are given in [7].

Step 3 is repeated until a specified time is reached or until crack arrest or the onset of unstable crack growth is detected; this also is described in [7].

3.5 Results

Numerical results have been obtained for a range of problems based on the above model and the geometry of Fig. 2b, see [7]. A finite element method in space based upon a mesh of eight-noded quadrilateral elements was used. In order to model the failure load and the displacement field in the failure zone adequately it was necessary to use local mesh refinement near the crack tip.

Example problem

$E = 2.98$, $\nu = 0.49$, $\omega = 8$, $a = 4$, $H = 2$

$\phi(t) = C + (1-C)e^{-t}$, $C = 10^{-2}$

$L_f = 10^{-4}$, $L_e = L_e(t) = 10^{-5}t$, $t < 0.1$

$= 10^{-6}$, $t \geq 0.1$

with L_f applied over the unit lengths (–4,–3) of the faces of the crack

$U_2^{cr} = 3 \times 10^{-7}$.

As a result of the computations, we obtain at time $t_1 = 0.1$secs

$$a_f(t_1) = 0.7827 \times 10^{-2}\ ,\ U_2(-a_f,t_1) = 0.274 \times 10^{-7}\ .$$

For this particular problem, with a constant load for time $t > t_1$, the failure zone length a_f does not change whilst the crack is stationary but, due to the creep effect, $U_2(-a_f,t)$ increases with t. Whether or not the crack moves depends on whether or not $U_2(-a_f,t)$ ever reaches the critical COD of 3×10^{-7}. With the particular parameters chosen we determine that the critical value is reached at time $t=t_{cr}=11.48$secs. We then successively increase the length of the crack in increments of equal length and find that the time taken between each crack length becomes progressively smaller and eventually negative. The "negative" increment does not correspond to a valid physical solution but it is an indication that unstable crack growth has begun.

Experiments with other values of C and a are described in [7]. The results of these experiments show, as we would expect, that the smaller the value of C, $0 < C < 1$, the larger the crack length can be before the onset of unstable crack growth. ($C = 0$ corresponds to a purely viscous material, whilst $C = 1$ corresponds to a purely elastic material.)

REFERENCES

1. Barenblatt, G.I., The mathematical theory of equilibrium cracks in brittle fracture. pp.55-129 of Advances in Applied Mechanics, Vol. VII. Academic Press, New York, 1962.
2. Christensen, R.M., Theory of Viscoelasticity. Academic Press, New York, 1971.
3. Knauss, W.G., On the steady propagation of a crack in a viscoelastic sheet: Experiments and analysis. pp.501-541 of H.H. Kausch and R. Jaffe (eds.) Deformation and Fracture of High Polymers. Plenum Press, London, 1973.
4. Rice, J.R., A path independent integral and the approximate analysis of strain concentration by notches and cracks. J. Appl. Mech. **34**, 379-386, 1968.
5. Schapery, R.A., Correspondence principles and a generalised J integral for large deformation and fracture analysis of viscoelastic media. Int. J. Fracture **25**, 195-223, 1984.
6. Shaw, S., Dawson, C., Warby, M.K., Wheeler, M.F. and Whiteman, J.R., Error estimates for finite elements in space/finite differences in time approximations to problems of linear viscoelasticity. (to appear)
7. Warby, M.K., Walton, J.R. and Whiteman, J.R., A finite element model of crack growth in a finite body in the context of Mode I linear viscoelastic fracture. Technical Report BICOM 90/8, Institute of Computational Mathematics, Brunel University, 1990. (to appear in Comp. Meth. Appl. Mech. Eng.)
8. Zienkiewicz, O.C., The Finite Element Method (3rd ed.). McGraw-Hill, New York, 1977.

NONLINEAR STATIC AND DYNAMIC ANALYSES OF A GENERIC ENCLOSURE SUBJECTED TO AN INTERNAL PRESSURE

Aaron Das Gupta
Research Mechanical Engineer
US Army Ballistic Research Laboratory, US Army Laboratory Command
Aberdeen Proving Ground, MD 21005-5066

ABSTRACT

Nonlinear elasto-plastic static and equivalent dynamic analyses of a box shaped generic enclosure subjected to an internal quasi-steady residual overpressure have been conducted using the ADINA finite element analysis code. The results indicate substantial deflection of the sidewalls and stress concentration effects at the corner joints between walls of the generic enclosure. Peak static deflection in excess of twice the wall thickness was predicted at both 38 and 54 MPa overpressures which are in satisfactory agreement with large deflection theory of plates.

INTRODUCTION

The deformation and stresses in a 3-D generic box shaped enclosure subjected to a residual quasi-static overpressure due to an internal rapid pressurization has been investigated in order to assess integrity and structural containment capability of the enclosure. Although the box shaped structure may be subjected to transient loads due to an internal explosive blast, only static and quasi-static residual overpressure have been considered due to long duration of this type of load after the transient phase resulting from lack of venting in containment structures. For an optimum design, the entire structure would experience large elastic-plastic deformation, thus providing a sink for the chemical energy of the explosive.

The impetus for this study is based on earlier work on suppressive and containment structures by Huffington et al [1] and Gupta et al [2-6]. In addition to an optimized hemispherical shell protective structure, it was demonstrated by Gupta et al [2] that other designs such as the rectangular parallelipiped configuration consisting of plates and bar elements can provide adequate containment capability, provided the structure is reinforced in critical areas. Since detailed modeling of enclosures with reinforcement members and fasteners is rather cumbersome, it was decided to model a generic box configuration which retains the basic features of the geometrical and constraint characteristics of realistic enclosure structures.

ESTIMATION OF RESIDUAL OVERPRESSURE

Estimation of residual overpressure in a vented enclosure due to detonation of an equivalent charge weight of TNT [7] is given as a relationship for the resultant increase in peak pressure, ΔP, from Reference [8] as

$$\Delta P = [0.4 h_c W_E]/V, kPa \tag{1}$$

where, $V = 0.18\ m^3$, the internal volume of the enclosure, $W_E = 1.36$ kg or 2.04 kg weight of the explosive charges, and $h_c = 13.5$ KJ/g, the heat of combustion of TNT.

PRESSURE DECAY AND BLOW-DOWN TIME

An internally pressurized structure vents the overpressure to the surroundings through openings in its walls and corners, causing leakage and a slow decay to ambient conditions. A relationship between overpressure and venting time is given by Kinney and Sewell [9] based on a modified Friedlander exponential decay of overpressure in order that

$$lnP = lnP_o - .315(A_v/V)t_s \tag{2}$$

where, t_s = venting time in ms, P = absolute pressure at t_s, and A_v = available vent area.

The long term duration of the decay is essentially due to the relatively small vent area available for blow-down to ambient conditions, resulting in a very slow pressure decay to the external atmosphere. The blow-down time, t_g, required to reduce the residual overpressure to ambient conditions obtained by Keenan et al [10], based on the test firing of explosives in chambers with known vent areas and internal volumes, is given as

$$t_g = 6.4(A_v/V)^{-.95} \tag{3}$$

The above equation is valid for $A_v/V^{2/3} \leq 0.21$. In the current design the ratio, $A_v/V^{2/3}$ equals 0.031 and the duration time for the quasi-steady pressure is approximately 100 ms.

Due to the large duration time and a slow rate of decay, the pressure is assumed to be uniformly distributed and is statically applied internally. From a conservative viewpoint the uniformly distributed static pressure was assumed to remain constant at the peak residual overpressure which was calculated from Equation (1) to be approximately 38 MPa due to detonation of 1.3 kg TNT. Transient effects of the detonation wave arriving at and reflecting from the sidewalls were part of a seperate study and as such was not included in the current investigation. However, the transient effects upon the response of the sidewall arising from application of the quasi-steady reflected overpressure in the form of a ramp function was deemed to be of some interest and was included in this study.

NUMERICAL MODELING

The analysis was performed using the ADINA [11,12] nonlinear finite element analysis code. The finite element mesh for the generic box was generated with the aid of the GEN3D mesh generator program. Due to considerable reinforcement and relatively large thickness used in the design of the rear wall, it was assumed to be a rigid wall to which the other sidewalls were ideally clamped at the rear while the front wall was relatively unrestrained.

The finite element mesh for the structure was generated with the aid of the GEN3D mesh generator program. The enclosure was modeled as an assembly 3-D brick elements uniformly spaced with the exception of the corner region where a refined mesh pattern consisting of two rows of brick elements were employed. A total of 144 eight-noded elements with 350 corner nodes were used to model the entire structure. A 2x2x2 integration points layout scheme was selected in each element for stress and strain computation. Corner radii have been simulated using an assemblage of piecewise linear segments along the inner surface which can be altered to represnt any desired corner radius.

The static overpressure load was applied unformly at each node point on the inner surface in a direction normal to the wall surface. A 3-D finite element mesh of the generic box configuration with element numbers is shown in isometric view in Figure 1.

MATERIAL MODEL

The primary construction material for the enclosure structure is a high strength steel alloy. Only the quasi-static material properties of the steel were considered and strain-rate effects were neglected, because these effects increase the structural resistance and thus reduce the total deformation.

The unixial static stress-strain curve and its bilinear approximation for use in ADINA are shown in Figure 2. The initial portion of the loading curve is linear with the proportional limit at .82 GPa stress and 0.4 percent strain. Young's modulus and Poisson's ratio are assumed to be approximately 205 GPa and 0.3 respectively, and the shear modulus is computed as 79 GPa.

The constitutive model selected in ADINA is the elastic-plastic material model with isotropic hardening. The input material parameters were obtained by approximating the experimental behavior as a bilinear elasto-plastic loading curve indicated by the intermittent curve in Figure 2. This curve is followed by linear elastic-plastic unloading resulting in a polygonal approximation of the experimental data.

TRANSIENT RESPONSE ANALYSIS

In addition to static analysis, a transient response analysis of the generic enclosure structure was conducted using a ramp loading function with gradual loading and subsequent unloading. Inertia effects due to a high rate of loading were thus kept to a minimum. A forcing function with a peak load of the same magnitude as the quasi-steady residual overpressure was uniformly distributed over the entire inner surface of the box. The forcing function was thus tailored to be equivalent to the static load applied gradually in a linear manner during the loading as well as unloading phases. Once

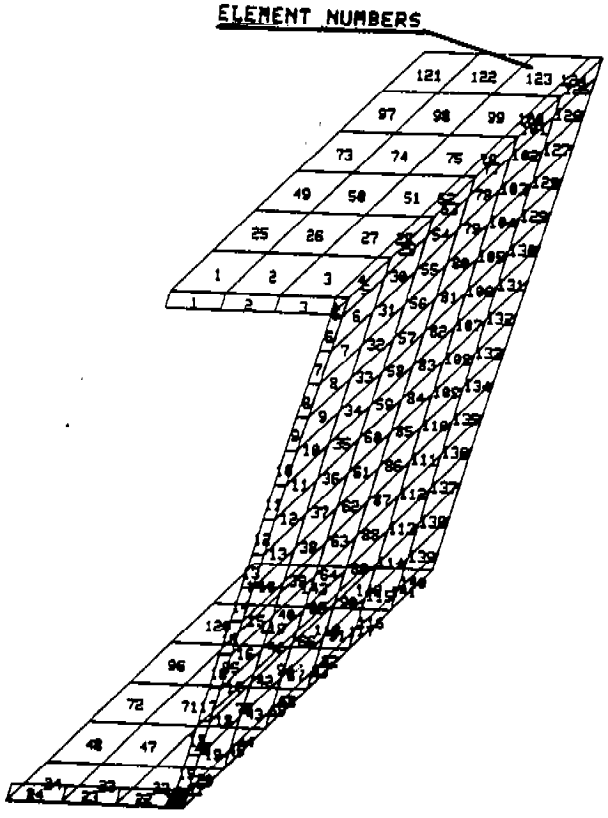


Fig. 1. A three-dimensional finite element model of a section of the internally pressurized box.

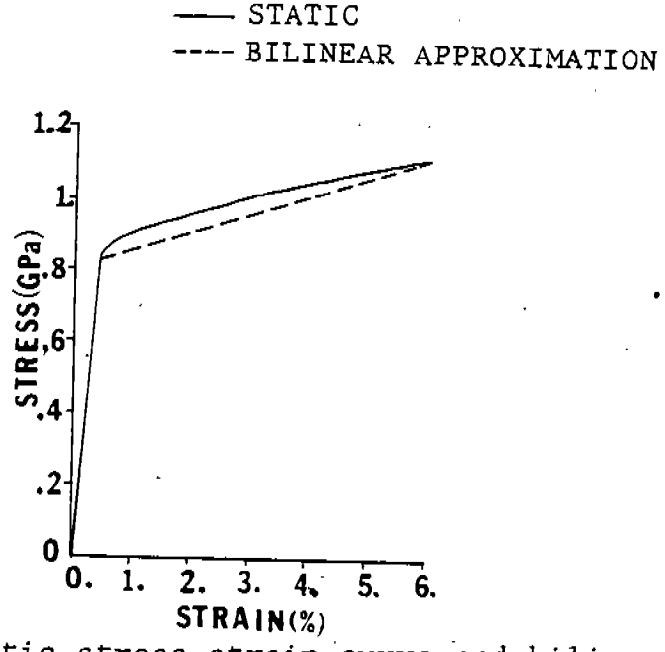


Fig. 2. Static stress-strain curve and bilinear approximation in ADINA for rolled homogenous armor(RHA) material.

the structure is fully unloaded, any residual permanent plastic deformation could be easily identified and compared to experimental measurements from the unloaded box. A plot of the applied force-time history is shown in Figure 3.

A central-difference explicit time integration scheme and a lumped mass formulation in the ADINA code were selected for the dynamic analysis. The time step used for the computational cycle was determined from the Courant stability criteria given as

$$\Delta t = \Delta t_{crit.}/n = \Delta l/[n\sqrt{E/\rho}] \tag{4}$$

where, $\Delta t_{crit.}$ is the minimum Courant stability time step, Δl is the distance between the two closest nodes in the system, E is the modulus of elasticity, ρ is the density of the structural material and n is the number of time steps with which we wish to represent the stress wave in passing through the distance, Δl. The value of $\Delta t_{crit.}$ was calculated to be approximately 4 microseconds. A value of 4 was selected for n which resulted in an initial time step for the explicit time integration scheme of 1.0 microsecond.

RESULTS AND DISCUSSION

Both nonlinear static and dynamic analyses of the internally pressurized generic box model and a comparison of responses from the two cases will be described in this section.

Nonlinear static analysis

Although displacements along all three major cartesian coordinate directions are observed throughout the generic enclosure away from the ideally clamped edges, the resultant peak deformation is predominantly associated with the sidewall region along a transverse thickness direction normal to the plate surface as shown in Figure 4. A magnification factor of .77 was selected for the deflection due to static pressure in the isometric configuration along the X-coordinate direction. The continuous lines represent isometric end view of the deformed box on which the initial configuration is superimposed as shown by intermittent lines. Computations using the ADINA Code indicate peak static deflections of approximately 6.0 and 8.6 cm at the inner surface at specified locations near the center of the sidewall for a generic box configuration with a uniform wall thickness of 2.54 cm corresponding to uniformly distributed applied pressure levels of approximately 38 MPa and 54 MPa, respectively.

A three-dimensional surface plot of the resultant static deflection of the sidewall with respect to the original configuration is shown in Figure 5. Peak deflection seems to occur at the center of the unclamped front edge of the sidewall. The opposite edge does not deflect due to the clamped boundary condition at the junction with the rear wall. However, high stress levels in excess of yield strength develop near the corners, possibly due to the existence of sharp corner radii at the junction. The other two edges are connected to the top and bottom walls and exhibit some displacement near the front edge. A peak effective stress of approximately 940 MPa and an effective plastic strain of .0231 at an integration point near the inner corner radius are observed in element No. 4 at a junction with the top wall of the box. The corresponding effective stress in the corner region near the bottom wall is nearly 900 MPa with an effective plastic strain of .015. These stresses may increase somewhat with further

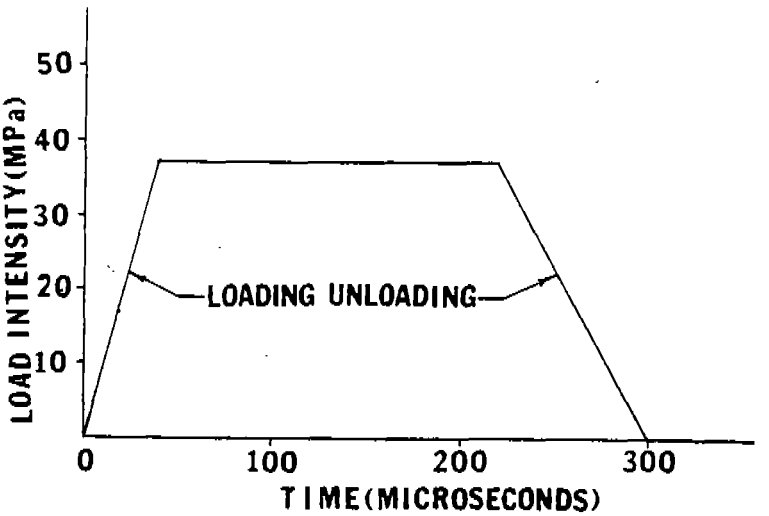


Fig.3. Applied pressure-time history assumed for internal loading for the equivalent dynamic problem.

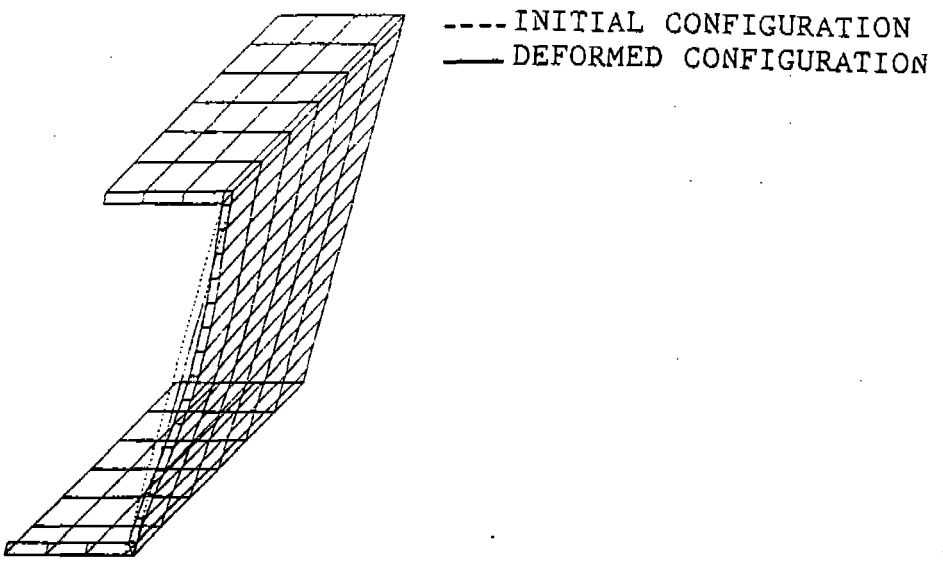


Fig. 4. Isometric view of resultant deflection at .78 magnification due to static pressure.

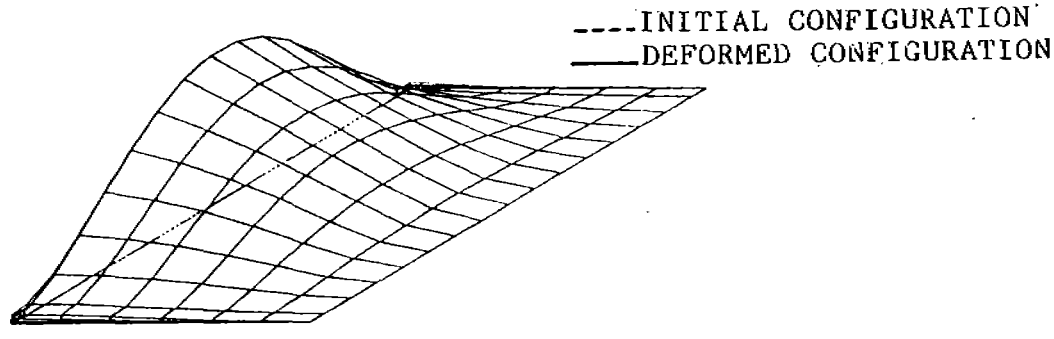


Fig. 5. Three-dimensional surface plot of the resultant sidewall deflection at 5.82 magnification due to static pressure.

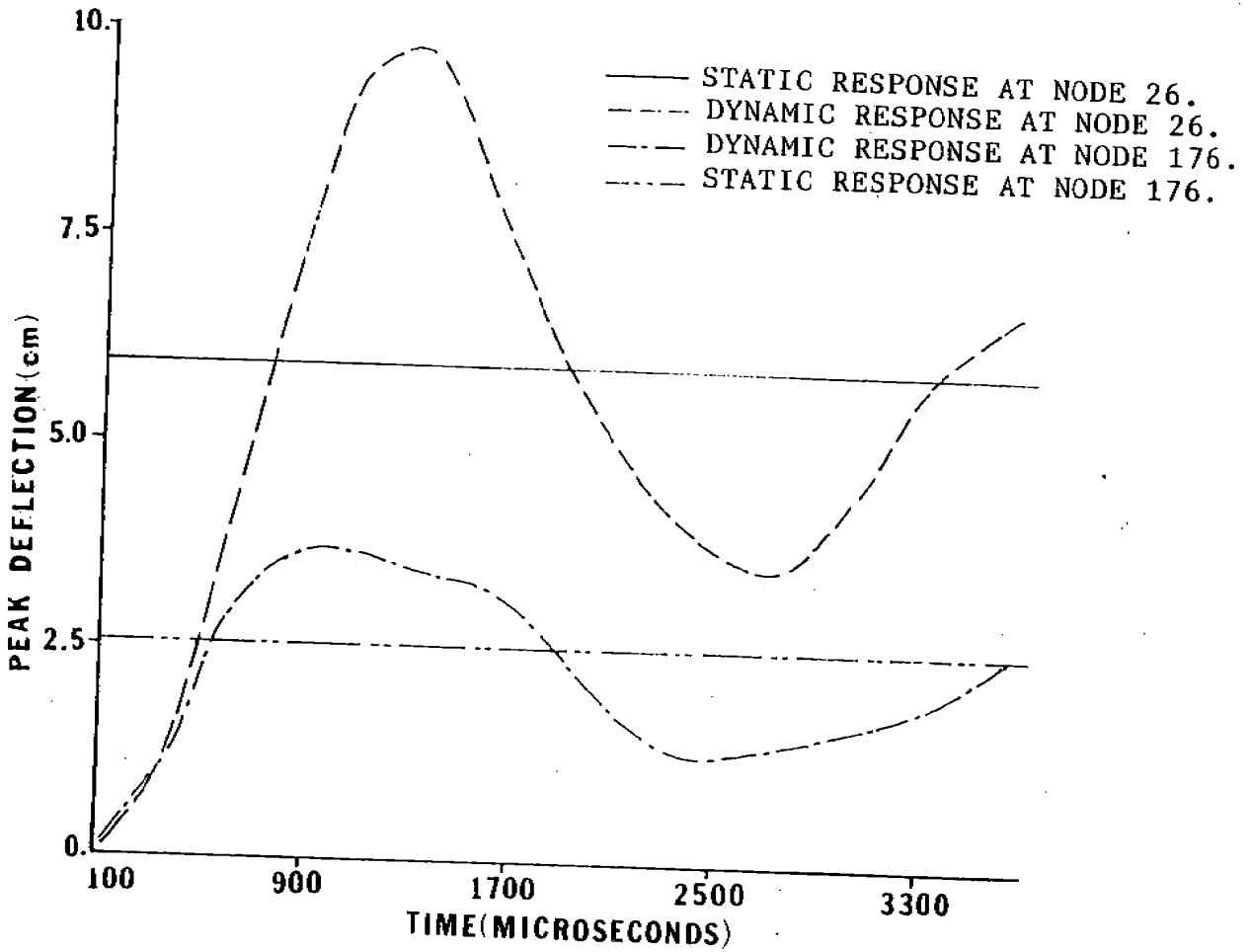


Fig. 6. Comparison of peak static and dynamic deflections at specified locations of the sidewall.

refinement of the model and are still the result of high stress concentration effects near the sharp corner radius. These levels could be reduced considerably by increasing the radius and consequent smoothening of the corner. Since both peak stress levels and displacements occur at the front edge, additional stiffeners and reinforcements in the corner region and in the sidewall near the front face are strongly recommended.

Nonlinear dynamic analysis

When the internal pressure is applied gradually in a time-dependent manner and is retained upon the structure for a sufficiently long time, oscillations in predicted deflection levels are observed. This results in overshooting followed by underprediction when compared with static predictions as shown in Figure 6. In this figure a delay of 1.3 ms in attaining the peak response is observed due to system inertia. However, at increased response times, oscillation peaks appear to diminish gradually until the structure is fully unloaded.

The transient response behavior of the enclosure structure as depicted in the isometric plots of resultant deflection in Figures 7-10 was monitored at response times between 1.1 and 3.5 ms at 800 microseconds interval. Deflection of the wall in a transverse thickness direction appears to diminish in magnitude beyond 1.5 ms and attains a minimum at 2.7 ms after which the deflection starts to increase at a rather slow rate. Figures 11-14 are a set of 3-D surface plots of the sidewall. These figures describe the deformation of the entire sidewall from 500 microseconds to 2.9 ms at an interval of 800 microseconds. Location of peak deflection and stress levels for the dynamic case coincide with the static data as expected due to identical geometry and equivalent loading data assumed for the dynamic model of the generic enclosure structure. The computation was terminated 3.5 ms because the response behavior appeared to be dominated by elastic oscillations without additional plastic deformation or plastic strain accumulation indicating the onset of steady state conditions.

CONCLUDING REMARKS

A static nonlinear large deflection analysis using ADINA is capable of determining the structural effects due to an internal blast inside a suppressive enclosure or containment structure. However, if the loading is applied dynamically at a high rate as for an explosively loaded hollow structure of a generic box configuration, it is evident from this investigation that the peak transient deflection can easily exceed the peak static deflection by a factor of 1.6 or more for a ramp loading function. For a step loading function the ratio of peak observed deflection between the static and the dynamic case is expected to be even higher due to large initial oscillations of the deflection curve. To ensure safe containment with an allowable margin of safety, the factor should be increased to 2.0 which is equivalent to doubling the quasi-static residual overpressure load. The resulting stress and deflection levels are large enough to require additional stiffening of the enclosure near the front as well as rear end of the sidewall accompanied by large increase in radius at the corner junctions of the box.

Implementation of structural damping in ADINA during the unloading phase could facilitate determination of residual plastic strains and deflections at critical locations within the enclosure. These results could then be effectively compared with available experimental measurements of residual plastic deformation of the sidewall after the occurrence of an internal blast due to detonation of an explosive inside a containment structure.

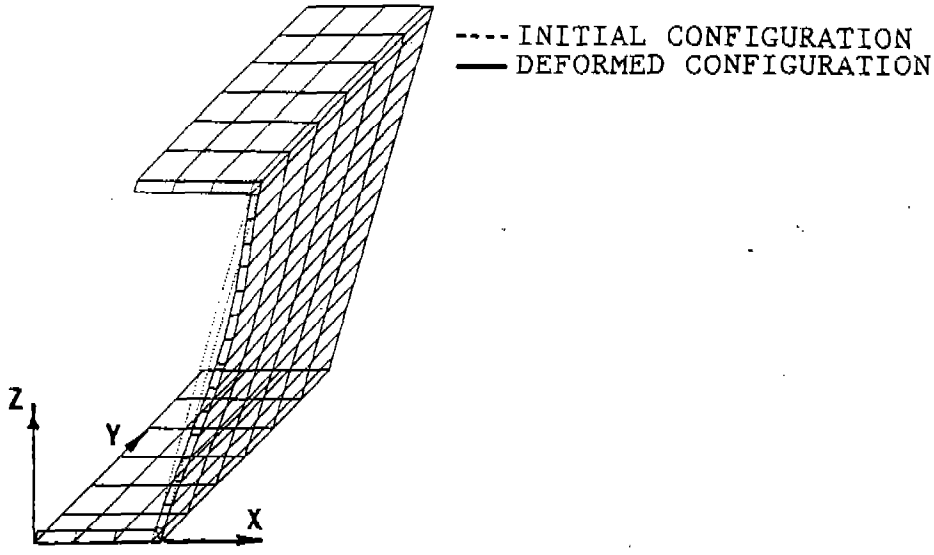


Fig. 7. Isometric view of the resultant deflection at .46 magnification at 1.1 ms. due to transient load.

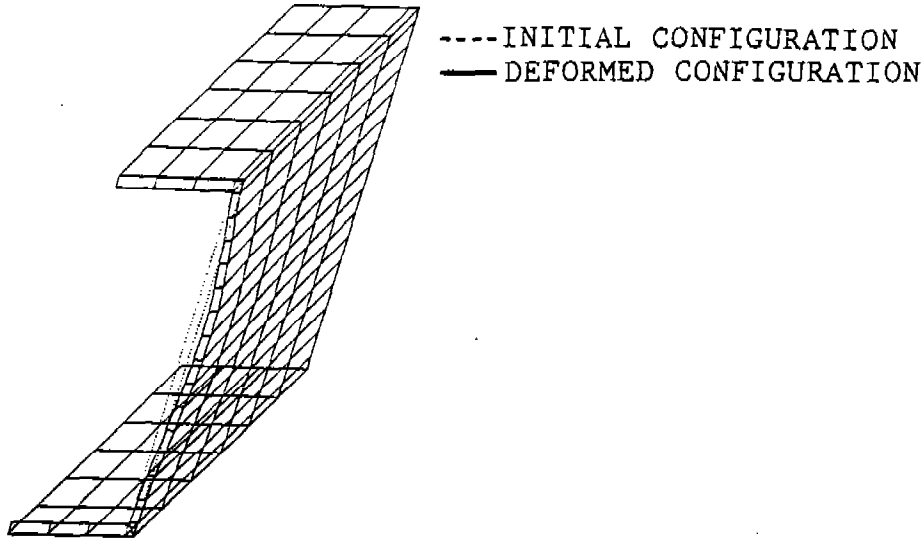


Fig. 8. Isometric view of the resultant deflection at .6 magnification at 1.9 ms. due to transient load.

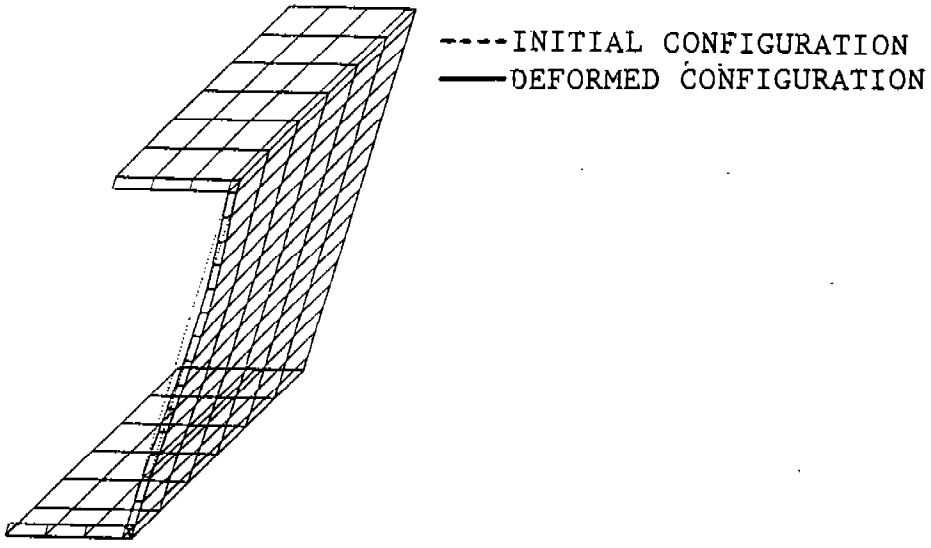


Fig. 9. Isometric view of the resultant deflection at .6 magnification at 2.7 ms. due to transient load.

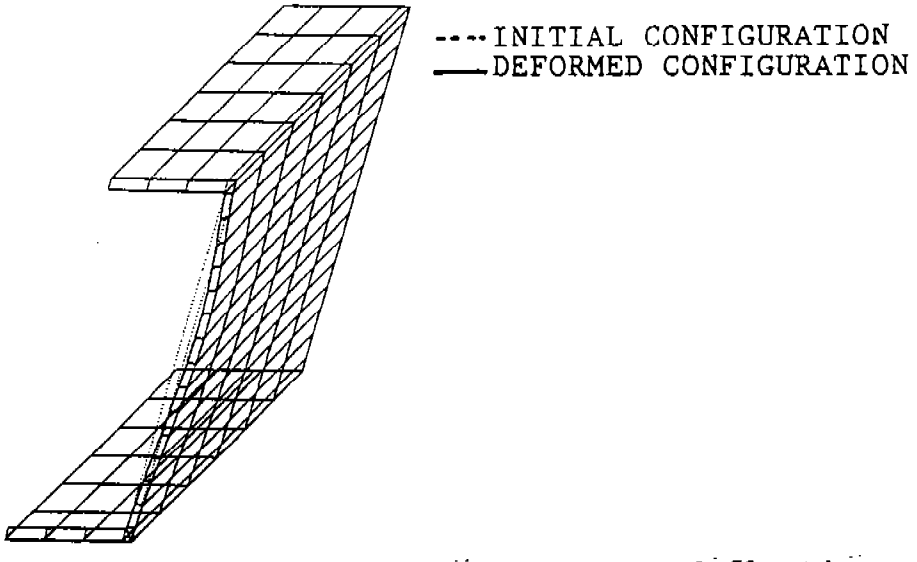


Fig. 10. Isometric view of the resultant deflection at .6 magnification at 3.5 ms. due to transient load.

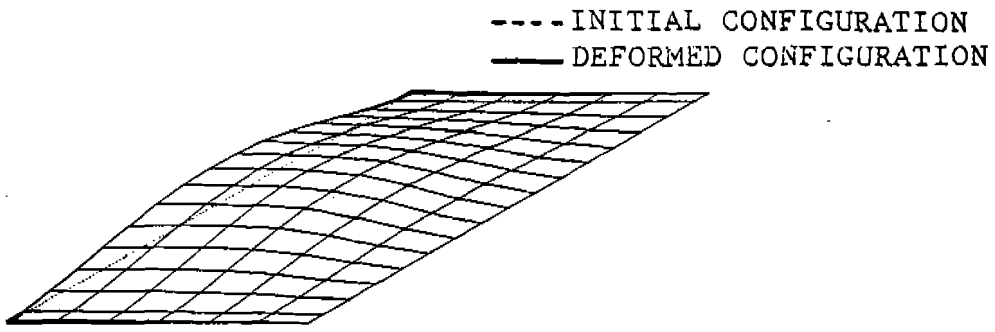


Fig. 11. Three-dimensional surface plot of the resultant sidewall deflection at 3.45 magnification at .5 ms.

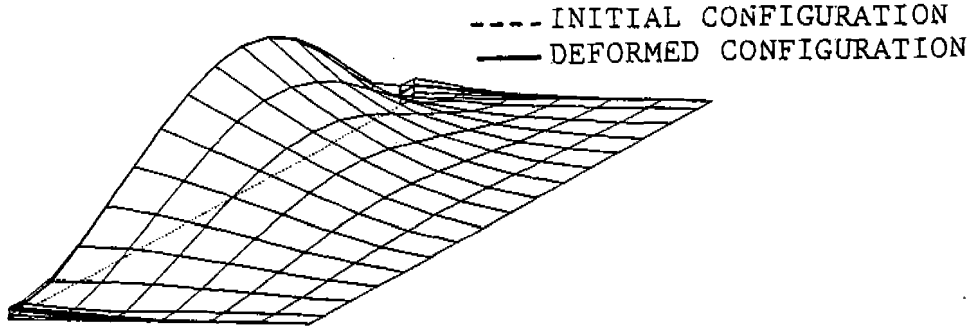


Fig. 12. Three-dimensional surface plot of the sidewall deflection at 3.45 magnification at 1.3 ms. due to transient load.

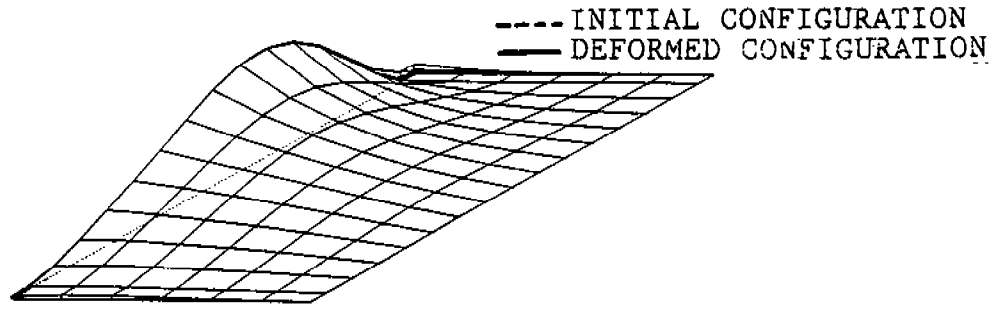


Fig. 13. Three-dimensional surface plot of the sidewall deflection at 4.55 magnification at 2.1 ms. due to transient load

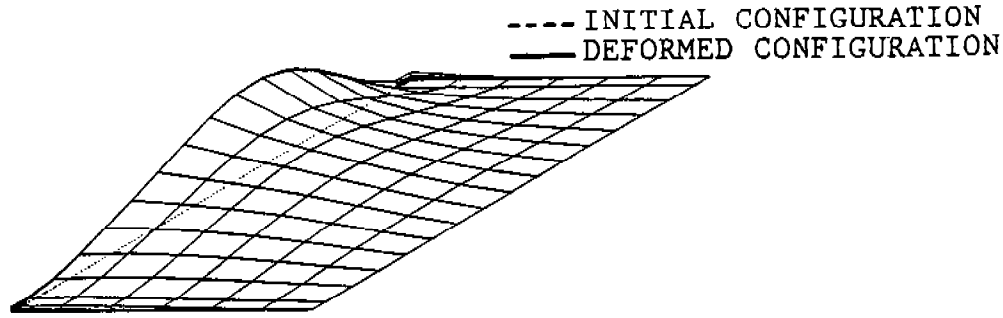


Fig. 14. Three-dimensional surface plot of the sidewall deflection at 4.55 magnification at 2.9 ms. due to transient load.

REFERENCES

1. N.J. Huffington and S.F. Robertson, "Containment Structures versus Suppressive Structures," BRL Memorandum Report No. 2596, February 1976.

2. A.D. Gupta and H.L. Wisniewski, "Stress Analysis of Enclosure Structures Subjected to Explosive Loads," Proceedings of the joint ASME/SESA Conference on Experimental Mechanics, Honolulu, Hawaii, May 28-30, 1982.

3. A.D. Gupta and H.L. Wisniewski, "An Optimized Configuration of an Enclosure Structure for Safe Containment of Internal Blasts," Proceedings of the International Symposium on Optimum Structural Design, The University of Arizona, Tucson, Arizona, December 1981.

4. A.D. Gupta and H.L. Wisniewski, "Dynamic Analysis of a Hemispherical Containment Structure Subjected to Transient Loads," AMD. Vol. 49, "Computer Analysis of Large Scale Structures," Part II, ASME, New York, November 1981.

5. A.D. Gupta, "Dynamic Behavior of a Discontinous Hemispherical Shell Subjected to an Internal Blast," PVP. Vol. 76, "Application of Nonlinear Analysis to Structural Problems," ASME, New York, N.Y., June 1983.

6. A.D. Gupta and H.L. Wisniewski, "Dynamic Response of the Hemispherical Containment Structure Subjected to Transient Loads at the R-9 Firing Range," BRL Memorandum Report ARBRL-MR-03249, March 1983.

7. B.M. Dobratz, "Properties of Chemical Explosives and Explosive Simulants," UCRL-51319 (Rev 1), July 1974.

8. Edward M. Weyer, Editor-in-chief, Annals of the New York Academy of Sciences, Vol. 152, "Prevention of and Protection Against Accidental Explosion of Munitions, Fuels and other Hazardous Mixtures," Published by the Academy, 2 East Sixty-Third Street, New York, N.Y. 10021, p.317

9. G.F. Kinney and R.G.S. Sewell, "Venting of Explosives," NWC Technical Memorandum No. 2448, July 1974.

10. W.A. Keenan and J.A. Tamareto, "Blast Environment from Fully and Partially Vented Explosions in Cubicles," U.S. Naval Civil Engineering Laboratory, Technical Report No. 51-027, February 1974.

11. K.J. Bathe, "ADINA (Automatic Dynamic Incremental Nonlinear Analysis) User's Manual, ADINA Engineering Inc., Cambridge, Massachusetts, December 1981.

12. K.J. Bathe, "Finite Element Procedures in Engineering Analysis," Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1980.

13. S.P. Timoshenko and S. Woinowsky-Krieger, "Theory of Plates and Shells," Second Edition, McGraw-Hill Book Company, Inc., New York, N.Y., 1959, pp. 420-424.

14. Y.C. Fung, "Foundations of Solid Mechanics," Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1965, pp. 466-469.

Calculation of Elastic-Plastic Wave Propagation on the Connection Machine

Mark A. Olson* and Kent D. Kimseyt

US Army Ballistic Research Laboratory
Aberdeen Proving Ground, Maryland 21005-5006

Abstract. This paper describes the parallel algorithms and data structures for implementing a 2-D multi-material kernel of the wave-propagation code HULL on a Connection Machine. Computational performance is illustrated for a rod-plate impact problem with material strength described through an elastic-perfectly plastic formulation. The hydrodynamic behavior of materials is modeled via the gamma law and Mie-Gruneisen equations of state.

1. Introduction. The emergence of massively parallel computers, such as the present generation of hypercube machines, is having a significant influence on the development and implementation of computational models for describing physical phenomena. A pressing concern in the construction of parallel applications is the mapping of algorithms onto scalable multiprocessors which can be scaled to the teraflop performance range.

An important class of problems where the principal limitation is CPU performance is the large-scale numerical solution of partial differential equations applied to shock physics modeling in two and three dimensions. The successful utilization of parallel computers for these problems requires the adaptation of existing sequential algorithms into reliable and robust parallel algorithms.

This paper presents a brief overview of the parallel algorithms and data structures for implementing a two-dimensional multi-material *kernel* of the wave-propagation code HULL on a Connection Machine. Computational performance is illustrated for a prototypical rod-plate impact problem. Particular detail is given to computational methodology, performance characteristics, and algorithm scalability. Complementary parallel computing efforts for recently developed wave-propagation codes are being conducted by Sandia National Laboratories[1] and Los Alamos National Laboratory.[2]

2. The Connection Machine. The Connection Machine CM-2[3] is a massively data-parallel computer configured with a maximum of 64K (2^{16}) bit-serial processors interconnected in a boolean n-cube topology. Each processor is equipped with 128

*Army High Performance Computing Research Center/Computer Sciences Corporation supported by US Army Research Office contract DAAL03-89-C-0088.
†Terminal Ballistics Division.

Kbytes of memory giving a total memory capacity of 8 Gbytes. The processors are arranged in hardware with 16 processors to a chip, and each pair of chips (referred to as a node) shares a Weitek floating-point accelerator each having 64-bit precision arithmetic.

Floating point computations on the CM-2 are implemented via two models: *fieldwise* and *slicewise*. In the fieldwise model, the atomic unit is the processing element and the storage of a 64-bit word is allocated in 64 sequential bits of a physical processor's memory. In the slicewise model, the atomic unit is the processing node and a word is stored in a 64-bit slice across the memories of the 64 processors in two nodes. The advantage of the slicewise model, is that a 64K-processor CM-2 becomes 1024 double-precision floating point nodes networked in a 12-cube topology with two communication channels between connected nodes.

The granularity of the CM-2 is reflected in the application of *virtual sets*. For the fieldwise model this refers to the formation of *virtual processors* (VPs) and for the slicewise model the abstraction of *virtual grids*. A virtual processor is the segmentation of the local memory of each processor, thus enabling the CM-2 to simulate a system with more physical processors. In contrast to VPs, a virtual grid does not exist as a formal object in CM memory, but provides a useful way for describing the allocated memory across processing nodes. The run-time system determines allocated memory within the processing elements and maps declared array dimensions onto the virtual grids. The execution of instructions by the virtual sets is performed by time-slicing the physical processing units.

The CM-2 processing units operate in a SIMD (Single-Instruction Multiple-Data) mode, meaning all processors receive the same instruction stream on each cycle. Conditional operations, *i.e.* masks, permit any subset of the processors to be deselected such that the instruction will only be performed by those processors in the selected set. The instruction stream is broadcast by sequencers which are controlled by a conventional front-end machine. The front-end machine supports the operating and programming environment. Current languages supported include CM-Fortran, C*, *Lisp, and Paris.

Interprocessor communication is carried out using two mechanisms referred to as the NEWS (North-East-West-South) grid and router. The addressing of a virtual processor is based on a Gray coded grid which provides an n-bit cube address, where $n \leq 16$, for specifying the location of the processor on an n-dimensional hypercube. The NEWS addressing scheme allows processors to pass data according to a structured rectangular grid. The router on the other hand, is the more general mechanism which allows any virtual processor to communicate with any other virtual processor on the hypercube. In addition, the router allows the local memories of the processors to be treated as a single large shared memory. The application of the NEWS grid and router for a given problem depends on the data pattern which may vary as a function of time.

3. The HULL Eulerian Hydrocode. The HULL code[4] is a multi-dimensional and multi-material Eulerian wave-propagation code that numerically solves the partial differential equations of continuum mechanics. Explicit terms for heat conduction and viscous effects are not included. The equations solved in axisymmetric cylindrical coordinates for 2-D are:

$$\dot{\rho} + \rho\left[\frac{\partial(xu)}{\partial x} + \frac{\partial v}{\partial y}\right] = 0, \tag{3.1}$$

$$\rho\dot{u} - \frac{\partial T_{xx}}{x\partial x} - \frac{\partial T_{xy}}{\partial y} + \frac{T_{\theta\theta}}{x} = 0, \tag{3.2}$$

$$\rho\dot{v} - \frac{\partial T_{xy}}{x\partial x} - \frac{\partial T_{yy}}{\partial y} = -\rho g, \tag{3.3}$$

$$\rho\dot{E} - \frac{\partial}{x\partial x}[x(uT_{xx} + vT_{xy})] - \frac{\partial}{\partial y}(uT_{xy} + vT_{yy}) = -\rho v g, \tag{3.4}$$

where ρ is the material density, x and y are the radial and axial coordinates, respectively, u and v are the corresponding radial and axial velocity components, T is the stress tensor, E is the total specific energy, and g is gravitational body force.

Equations (3.1) through (3.4) are solved on a finite-difference rectangular mesh composed of discrete spatial intervals Δx_i, Δy_j in the radial and axial coordinates. The solution is advanced explicitly from the initial conditions by discrete time steps, Δt^n, and is defined on the mesh (x_i, y_j, t^n) where each of the state variables $\xi(x,y,t)$ in the solution space is defined by $\xi^n_{i,j} = \xi(x_i, y_j, t^n)$.

State variables are defined at the geometric center of each cell. Cell boundary values are interpolated through one computational cycle via cell-centered values from nearest-neighbor cells. These boundary values are then advanced through one-half time step using cell-center to cell-center gradients. This step is then followed by a full time step using half-time advanced cell-boundary gradients. Lagrangian conservation Eqs. (3.1) - (3.4) are utilized in this time update. To maintain the original Eulerian mesh, material is advected from one cell to another via a first-order donor cell algorithm with a heuristic multi-material diffusion limiter to preserve material interfaces.

Material models in HULL include elastic-perfectly plastic with von Mises yield criterion as well as temperature and work hardening effects. The Mie-Gruneisen equation-of-state is used to model solids and liquids, and the gamma law is used to model gases. Explosives are modeled via the Jones-Wilkins-Lee equation-of-state. Material failure models include maximum principal stress, maximum principal strain, and the Hancock-Mackenzie triaxial failure model.

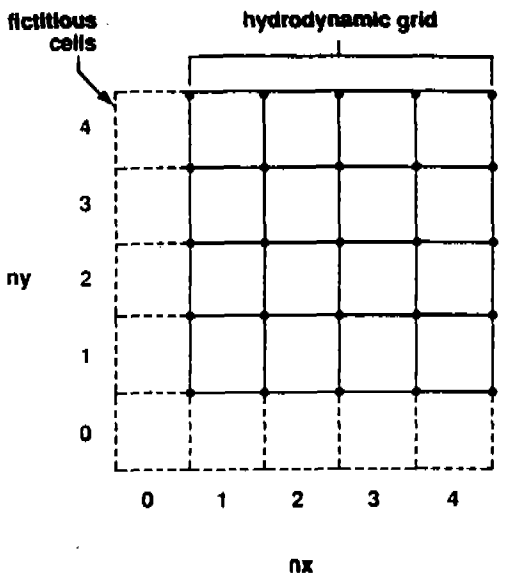


Figure 1: CM-2 computational grid.

4. Parallel implementation of HULL. Implementation complexity of adapting the HULL code to a parallel platform depends on several factors—namely, the degree of parallelism, granularity and scalability, interprocessor communication, and I/O demands. To achieve high performance, efficient data parallelism must be constructed which maximizes processor load and streamlines interprocessor communication.

4.1 CM-2 data structure. The algorithmic framework for mapping the HULL data structure onto the CM-2 architecture lies in the utilization of both the canonical layout of arrays and the use of the compiler array directive LAYOUT.[5]

Hydrodynamic variable arrays for pressure, velocity, stresses, and strains are canonically allocated one element per virtual processor[6] with each conformable array being placed in the same virtual set. Array dimensions are defined in 2-D as $(0{:}nx,0{:}ny)$, where nx and ny are the number of hydrodynamic computational cells in the x and y spatial directions, respectively. Each array is buffered with fictitious cells (see Figure 1) containing the appropriate boundary conditions. Boundary conditions accounted for include both transmissive and reflective.

Fictitious cells are incorporated into the mesh to perform uniform computations on all active cells at all times independent of whether the cells are internal or boundary cells. This approach maximizes processor utilization during a clock cycle for the Lagrangian and advection computations, thereby decreasing the overall computational grind time. The boundary conditions for the top and right are carried out in parallel while the densities of the fictitious cells are being numerically updated.

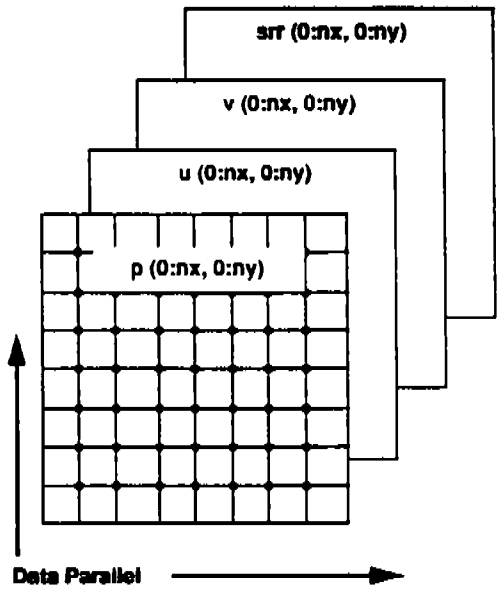


Figure 2: Data-parallel hydrodynamic variable arrays.

All grid axes for the hydrodynamic variable arrays are NEWS-ordered (see Figure 2). Elemental operations between the arrays in a virtual set require no interprocessor communication and dimensional shifts on cells, as required in finite-difference schemes, are performed with NEWS communication.

The compiler directive LAYOUT allows the programmer to specify the axis ordering and weights of the virtual set in which an array is allocated. An important application of LAYOUT is for arrays with mixed data-parallel (NEWS-ordered) dimensions and serial dimensions. An example is the mass array shown in Figure 3. Elements are given by xm(:SERIAL, :NEWS, :NEWS), where the SERIAL dimensions span the number of materials (denoted by *nm*) and NEWS the mesh space. Computations over the serial dimensions are performed via the front-end and data-parallel dimensions are performed on the CM-2. Similar mixed arrays are constructed for material volumes and energies. Each mixed array can be viewed as an indexed collection, *i.e.* a material slice, of data-parallel arrays.

4.2 Lagrangian computations. The cornerstone in reprogramming the Lagrangian step for SIMD operations lies in the functionality of the NEWS communication. Finite-difference schemes are implemented via the application of intrinsic shift functions performed on data-parallel arrays.

As an example, the finite-difference representation for the u-component velocity computed at the cell boundary $i+1/2$ at time $t = t^n$ is given by:

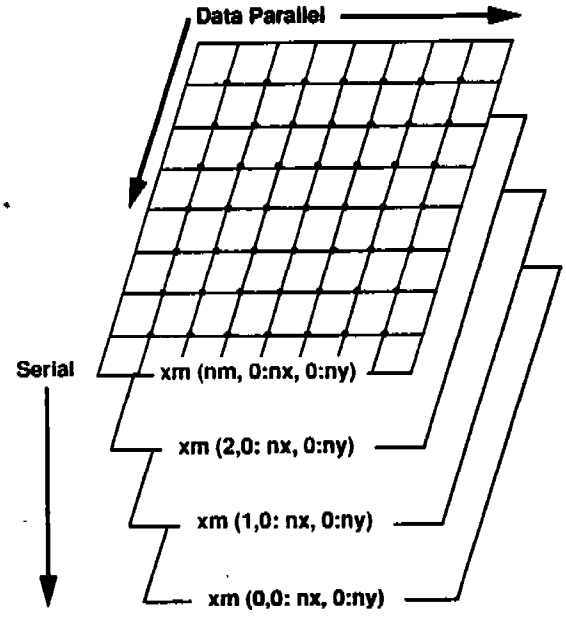


Figure 3: Data-parallel material-indexed hydrodynamic variable arrays.

Serial Lagrangian

$$u^n_{i+1/2,j} = \frac{\rho^n_{i,j} u^n_{i,j} + \rho^n_{i+1,j} u^n_{i+1,j}}{\rho^n_{i,j} + \rho^n_{i+1,j}}; \tag{4.2.1}$$

Data-Parallel Lagrangian

$$u^n_{1/2} = \frac{\rho^n u^n + \text{cshift}(\rho^n u^n, 2, 1)}{\rho^n + \text{cshift}(\rho^n, 2, 1)}. \tag{4.2.2}$$

The key point is the replacement of sequential operations on array elements $\rho^n_{i,j}$, $u^n_{i,j}$ with the global uniform operation on data-parallel arrays ρ^n, u^n. The circular shift, cshift(ρ^n,2,1), has the effect of shifting the data-parallel array ρ^n to the left by one position. These operations are one of the most efficient CM-Fortran operations due to the direct mapping onto the NEWS communication grid. (A caveat is that the grid dimensions must be a power of two for fieldwise and multiples of four for slicewise.)

The data-parallel solution for the Lagrangian Eqs. (3.2) - (3.4) with the assumption $\Delta t = \Delta t^n$ is given by:

$$u^{n+1} = u^n - \frac{\Delta t}{\rho^n}\left[\delta^x P^{n+1/2} - \frac{1}{x}\delta^x x S^n_{xx} - \delta^y S^n_{xy} + \frac{S^n_{xx} + S^n_{yy}}{x}\right], \tag{4.2.3}$$

$$v^{n+1} = v^n - \frac{\Delta t}{\rho^n}\left[\delta^y (S^n_{yy} - P^{n+1/2}) \frac{1}{x}\delta^x x S^n_{xy}\right], \tag{4.2.4}$$

$$E^{n+1} = E^n + \frac{\Delta t}{\rho^n}\left\{\frac{1}{x}\delta^x\left[x u^{n+1/2}\left(S^n_{xx} - P^{n+1/2}\right) + x v^{n+1/2} S^n_{xy}\right]\right\}$$

$$+ \frac{\Delta t}{\rho^n}\delta^y\left[v^{n+1/2}\left(S^n_{yy} - P^n\right) + u^{n+1/2} S^n_{xy}\right], \tag{4.2.5}$$

where P^n and $S^n_{\mu\lambda}$ for (μ=x,y;λ=x,y) are data-parallel arrays for pressure and stress deviator, respectively, δ^λ is the spatial derivative

$$\delta^\lambda \xi^n = \left(\xi^n_{1/2} - \text{cshift}(\xi^n_{1/2},\text{dim},-1)\right)/\Delta\lambda$$

with $\xi^n_{1/2}$ defined as the spatial-centered term, dim = 1,2 depending on if $\lambda = x$ or y, and $\Delta\lambda = \text{cshift}(\lambda,\text{dim},1) - \lambda$.

Data-parallel expressions for $P^{n+1/2}_{1/2}$ are given by

$$P^{n+1/2}_{1/2} = P^n_{1/2} - \frac{\Delta t}{2}\left(\rho C_S^2\right)^n_{1/2} \frac{1}{x}\,\delta^x (xu)^n_{1/2} \tag{4.2.6}$$

for the radial direction and

$$P^{n+1/2}_{1/2} = P^n_{1/2} - \frac{\Delta t}{2}\left(\rho C_S^2\right)^n_{1/2} \delta^y (v^n_{1/2}) \tag{4.2.7}$$

for the axial direction. The spatial-centered pressures of Eqs. (4.2.6) and (4.2.7) are defined by

$$P^n_{1/2} = \frac{\text{cshift}(P^n,\text{dim},1)\rho^n + P^n \text{cshift}(\rho^n,\text{dim},1)}{\rho^n + \text{cshift}(\rho^n,\text{dim},1)}$$

with dim depending on either the radial or axial direction. The $\left(\rho C_S^2\right)^n_{1/2}$ term in Eqs. (4.2.6) and (4.2.7), where C_S is the isentropic sound speed, is given by

$$\left(\rho C_S^2\right)^n_{1/2} = \min\left[\left(\rho C_S^2\right)^n, \text{cshift}\left(\left(\rho C_S^2\right)^n,\text{dim},1\right)\right]$$

with $\left(\rho C_S^2\right)^n$ computed via the equation-of-state.

Data-parallel time advanced velocities in Eq. (4.2.5) are computed via the following expressions:

$$u_{1/2}^{n+1/2} = u_{1/2}^{n} - \left[\frac{\Delta t}{2\max\,(\rho^n,\text{cshift}(\rho^n,2,1))}\right](\text{cshift}(P^n,2,1) - P^n)/\Delta x,$$

$$v_{1/2}^{n+1/2} = v_{1/2}^{n} - \left[\frac{\Delta t}{2\max\,(\rho^n,\text{cshift}(\rho^n,1,1))}\right](\text{cshift}(P^n,1,1) - P^n)/\Delta y$$

$$- \frac{\Delta t}{4}(g + \text{cshift}(g,1,1)),$$

where $u_{1/2}^{n}$ is given by Eq. (4.2.2) and $v_{1/2}^{n}$ has an analogous form.

Similar computations are carried out for the stress deviators. The numerical solution in a data-parallel format is obtained explicitly by

$$S_{\mu\lambda_{1/2}}^{n} = \Phi_{1/2}\frac{S_{\mu\lambda}^{n}\text{cshift}(\rho^n,\text{dim},1) + \text{cshift}(S_{\mu\lambda}^{n},\text{dim},1)\rho^n}{\rho^n + \text{cshift}(\rho^n,\text{dim},1)},$$

where

$$\Phi_{1/2} = \min(VF,\text{cshift}(VF,\text{dim},1))$$

with VF defined as a data-parallel array describing the fractional volume of solid in a given computational cell. The stress deviators are numerically updated and are subject to the Von Mises yield criterion.

The application of the boundary conditions for the Lagrangian and advection computations is implemented through the use of data-parallel selector arrays containing values of 1.0 for selecting computational cells and values of 0.0 for deselecting cells. For example, the left reflective boundary condition for $u_{1/2}^{n}$ given by Eq. (4.2.2) requires the left fictitious cells to hold the temporary value of $u_{1/2}^{n} = 0.0$. This is accomplished by multiply the data-parallel expression for $u_{1/2}^{n}$ by an array containing 1.0 for all active cells and 0.0 for the left fictitious cells. Similar selector arrays are employed for implementing analogous boundary conditions.

4.3 Equation-of-state computations. Equation-of-state (EOS) calculations are in general good candidates for the SIMD data parallelism of the CM-2. They are characterized as being free of both interprocessor communication and grid boundary conditions. However, for multi-material problems EOS calculations are inherently MIMD (Multiple-Instructions Multiple-Data) type operations. The MIMD nature is due to the nonhomogeneity of the computations derived from materials with different EOS formulations (*e.g.*, gamma law and Mie-Gruneisen) and different material parameters

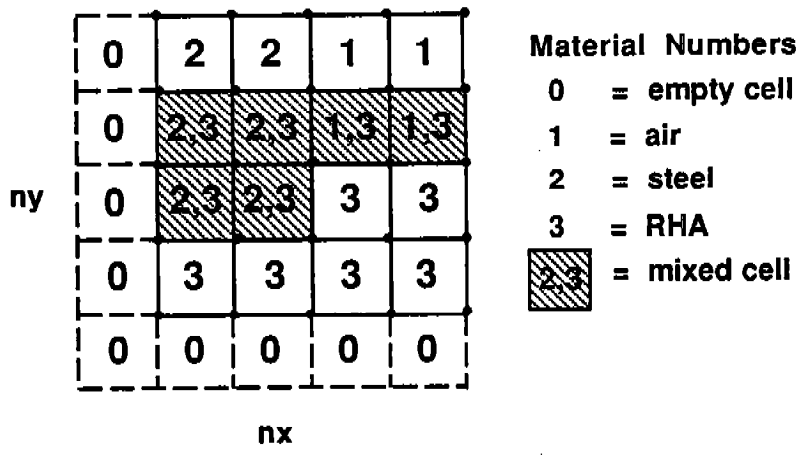


Figure 4: HULL EOS computations.

characterizing the same EOS (*e.g.*, steel and RHA). Moreover, mixed material cells, which require an iterative procedure to equilibrate the pressure for each material, induce a MIMD style of programming. Figure 4 depicts schematically the general condition for computing the EOS for a three-material simulation.

The most direct method for computing pressures employing analytic EOS expressions of the form $p = p\{\rho,I\}$, where I is the internal energy, is one which calculates in parallel cell pressures (partial pressures for mixed material cells) as part of a sequential loop over all materials. The calculated result is placed in a data-parallel scratch array *pp(im,:,:)*, where *im* is the material index. A logical mask is then use to segregate pure and mixed cells, with mixed cells requiring further calculations.

The problem with this method is twofold. First, there is a *nm*-factor increase in the set of required computations due to the sequential loop over the materials rather than one data-parallel SIMD computation. This can be somewhat relaxed for materials with identical EOS formulations by introducing data-parallel material property arrays for each material at each VP (or virtual grid). For virtual sets with identical materials one array would be required. Unfortunately this determination is dynamic and not static.

The second problem deals with mixed material cells. Each mixed cell under goes a volume iteration in an effort to compute an equilibrium pressure. During this iteration, the VPs (or virtual grids), which hold pure cells, are conditionally masked such that they are inactive. As the number of iterations and mixed cells grows, the relative cycle throughput of SIMD operations decreases. Similar problems occur during the advection phase. The elimination of these problems require asynchronous constructs and are not supported in a SIMD platform.

The SIMD methodology for computing material strength is similar to that for computing pressures. Scratch data-parallel arrays are employed to store temporary values of the shear modulus, yield strength, stress deviators, etc. for both pure and

mixed material cells during volume iterations. Upon convergence all cell values are reloaded into their respective hydrodynamic variable arrays.

4.4 Advection computations. As mentioned above, HULL advects materials based on a first-order donor cell method. The calculation of the relative transport weights for apportioning the volume flux is carried out using the intrinsic cshift function for computing the fractional volumes in the receiver and upstream cells. A diffusion limiter algorithm is employed in an attempt to unmix mixed material cells.

The material slices for computing transport terms are stored in a data-parallel array hs(:SERIAL, :SERIAL, :NEWS, :NEWS), where the SERIAL dimensions cover the number of materials and spatial flux directions (4 in 2-D), respectively. The NEWS-ordered dimensions span the mesh space and are conformable with the advected hydrodynamic variable arrays. Volume iterations are required to reduce the flux of overemptied materials. Convergence is checked by monitoring a data-parallel array consisting of ones and zeros.

The final remapping step is transparent in its implementation using simple grid finite-difference quantities computed via cshift operations. For example, the volume of material n, denoted by the data-parallel array V_n, is advected to the original fixed Eulerian mesh

$$
\begin{aligned}
{}^{e}V_n &= V_n + \delta V_{n,2}(left) + \delta V_{n,3}(bottom) - \delta V_{n,4}(right) - \delta V_{n,1}(above) \\
&= V_n + \text{cshift}(\delta V_{n,4},2,-1) + \text{cshift}(\delta V_{n,1},1,-1) - \delta V_{n,4} - \delta V_{n,1}
\end{aligned}
$$

where ${}^{e}V_n$ is the Eulerian volume and the transporting volume is

$$\delta V_{n,l} = \left(\Delta_{n,l} V_n\right)_{donor}$$

with $\Delta_{n,l}$ defined as the transport fraction for each material in particular direction l. Active cells are advected while fictitious cells along with inactive cells are masked.

5. Application and performance results. The application we report here as an illustration of the computational performance is a 2-D multi-material computation of a steel rod impacting rolled-homogeneous armor (RHA) at a striking velocity of 3 km/sec (see Figure 5). The computational geometry is such that the length-to-diameter ratio of the steel rod was set to five. Material strength was implemented via an elastic-perfectly plastic formulation with the hydrodynamic behavior of materials modeled using the gamma law and Mie-Gruneisen equations of state.

Calculations were performed on a 16K segment of a 32K-processor CM-2 located at the University of Minnesota. The total memory capacity is 4 Gbytes with a DataVault of 10 Gbytes. The front-end is a VAX 6420 with 64 Mbytes of memory running the

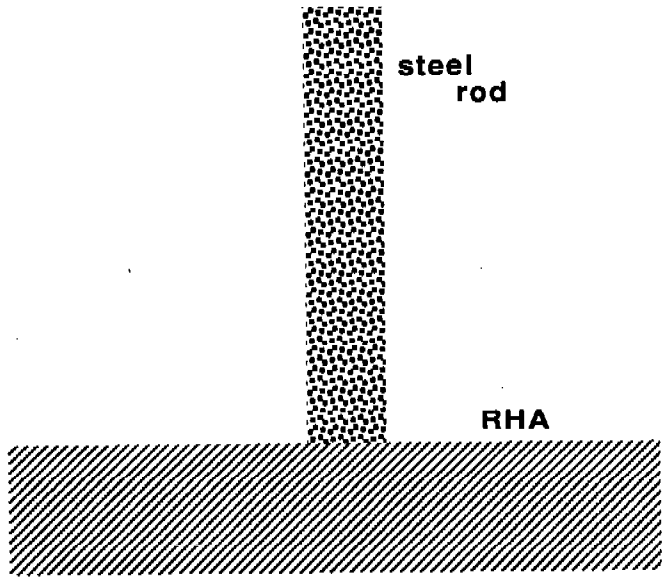


Figure 5: HULL application on the CM-2.

ULTRIX operating system. Reprogramming of the HULL code was carried out using CM-Fortran with double-precision arithmetic implemented via the slicewise compiler.

Results for the grind times (microsec/cell/cycle) computed on the CM-2 for various mesh sizes along with the corresponding CRAY-2 single processor results are presented in Table I.

TABLE I. HULL hydrocode performance results on the CM-2.[a]

	CM-2			CRAY-2
Grid Size	VG length[b]	efficiency[c]	grind time[d]	grind time[d]
128 x 128	32	0.87	39	-
256 x 256	128	0.93	22	-
512 x 512	512	0.97	16	196

[a]CM-Fortran with double precision using slicewise compiler on a 16K CM-2.
[b]VG (virtual grid) length = number of grid points/number of FPUs.
[c]efficiency = CM-2 execution time/CM-2 elapsed time.
[d]grind time = μ-sec/cell/cycle.

A comparison of the computed grind times shows the 16K-processor CM-2 performance is faster than a CRAY-2 processor. For a 512x512 mesh the CM-2 is twelve times faster. Note that the grind times for a fixed CM-2 scale inversely and nonlinearly with the virtual grid length.

The observed improvement in efficiency as a function of data set size is due to the amortization of the start-up overhead over large blocks of computations and to some of the communication occurring on the same chip. The overall SIMD parallelism performance of the HULL code is limited by the equation-of-state solution procedure employed in solving for mixed cells. Recently developed EOS methods[7] appear to be more amenable to the data parallelism of the CM-2.

6. Conclusions. In this paper, we have presented the initial step toward the adaptation of the HULL code for the Connection Machine. Results for a parallel implementation of a prototypical rod-plate impact calculation have been shown to be faster than the CRAY-2 results. Extrapolating the CM-2 grind times to a full 64K-processor machine, suggests that this machine is capable of *fifty* times the performance of the CRAY-2 for executing the HULL code. However, performance is limited by the EOS calculation for the multi-material mixed cells.

Acknowledgements. The computations reported here were made possible by the University of Minnesota AHPCRC Supercomputer Resources.

REFERENCES.

1. A. C. Robinson, *et al.*, Sandia National Laboratories Report SAND90-0589 (1990).

2. J. W. Hopson, Los Alamos National Laboratory, private communication (1990).

3. *Connection Machine Model CM-2 Technical Summary*, Thinking Machines Corporation, Cambridge, MA (1990).

4. D. A. Matuska and J. J. Osborne, *HULL Documentation: Technical Discussion VOL I*, Orlando Technology Incorporated (1987).

5. *CM Fortran Reference Manual, Ver. 5.2-0.6*, Thinking Machines Corporation, Cambridge, MA (1989).

6. Formally a distinction should be made between fieldwise and slicewise mapping of arrays. See the report: *CM Fortran Programming Guide, Ver. 1.0*, Thinking Machines Corporation, Cambridge, MA (1991).

7. J. M McGlaun, S. L. Thompson and M. G. Elrick, Int. J. Impact Engng Vol 10, pp. 351-360, 1990.

FINITE ELEMENT SOLUTION OF TRANSIENT IN-BORE RESPONSE PROBLEMS

Kenneth A. Bannister and
Stephen A. Wilkerson
US Army Ballistic Research Laboratory (BRL)
Aberdeen Proving Ground, Maryland 21005-5066

and

Donald A. Rabern
Los Alamos National Laboratory (LANL)
Los Alamos, New Mexico 87545

ABSTRACT. Many interesting aspects of mathematical and numerical modeling come to bear in solving dynamic structural response problems concerning gun launches of projectiles. A cooperative effort between BRL and LANL has been underway for several years now on 3D transient modeling of tank gun-launched saboted-rod kinetic energy (KE) projectiles. The focus of this work has been on numerical/experimental investigations to better understand the transient behavior of KE rounds during in-bore travel. We believe that improved knowledge of this behavior will lead to improved designs of KE projectiles and improve accuracy on targets. Because formal mathematical solutions of such complex large-scale modeling problems are impractical, approximate but accurate solutions are being sought by means of the Finite Element (FE) Method. Numerical simulations are being carried out with the DYNA2D, DYNA3D, and PRONTO3D nonlinear transient FE codes, together with their respective pre- and post-processor software. Calculations have been successfully carried out with a network in which engineering work stations are used at the local level, and Cray XMP and YMP supercomputers are used for heavy-duty computational work. To the authors' knowledge, this is the first such comprehensive use of full transient 2D and 3D FE simulation techniques to model the interior ballistic phase of KE round launches.

INTRODUCTION. In this paper we illustrate the use of advanced state-of-the-art structural analysis tools such as the explicit nonlinear 2D and 3D finite element (FE) codes DYNA2D (Hallquist 1984), DYNA3D (Hallquist and Benson 1986), and PRONTO3D (Taylor and Flanagan 1989) to simulate the transient in-bore structural responses of saboted-rod kinetic energy (KE) projectiles. In light of the many finite element tools which are available for quasistatic analysis of projectiles, one may well ask: Why transient analyses? In years past, of course, the main obstacle to performing transient FE analyses on the problem at hand was the lack of supercomputing resources; this situation has been resolved with the ready availability of Cray-class machines. We can now truly concentrate on the physical reasons which justify fully transient modeling:

o Quasistatic analyses of in-bore problems yield at best approximate results: This has long been recognized but the lack of transient analysis resources prevented progress;

o Wave propagation effects: In real guns, pressure waves due to combustion processes are entirely possible; solid phase impacts of propellant grains on projectile surfaces can occur;

o Material strain rate and large strain effects: Rotating bands and obturators undergo strain rates as high as 1000/second; strains of order 200% can occur;

o Response to an imperfect world: Tube wear and erosion effects can lead to torsional impulse (in artillery shells, this is a sudden torque applied to the projectile at instant of barrel engagement); for KE projectiles variations in bore profile straightness cause balloting and tube vibration; asymmetries in KE sabot designs are important;

o Shot exit: Rapid unloading effects can occur at muzzle exit as the gun gas pressure suddenly drops off.

We have established a hierarchy of computer models to analyze the problems just described. For example, the RASCAL (Erline, et al. 1990)and SHOGUN (Hopkins 1990) gun dynamics models are beam element-based quasi-2D and quasi-3D codes which we use for preliminary studies or "quick looks" at KE projectile/barrel interaction problems. DYNA2D and NIKE2D are explicit and implicit continuum FE codes, respectively, useful for analyzing the structural response and integrity of projectiles when subjected to axisymmetric loadings. DYNA2D is by far the more useful since it has good transient analysis capabilities. NIKE2D, because of its implicit formulation, runs much faster than DYNA2D, but lacks the capability to handle highly transient loads. We use NIKE2D primarily for problem check-out and certain specialized calculations. BRL has coupled DYNA2D with in-house interior ballistic burn codes for more accurate modeling of the actual interaction of combustion physics with projectile motion. This coupling of these models has recently allowed study of more difficult interior problems such as the dynamic strain amplification problem in gun barrels, a resonance condition in the barrel muzzle region attributable to a pressure front moving rapidly downbore. For 3D transient work, we are using the explicit DYNA3D (at both BRL and LANL) and PRONTO3D (at LANL only) codes.

Figure 1 shows schematically the mechanical components of a KE projectile configuration used for defeat of tank armor, in this case the 120-mm M829 design. The lower portion of the Figure shows the DYNA2D FE grid we use at BRL for 2D calculations. Figure 2 shows the corresponding DYNA3D grid used at LANL for 3D calculations. Only since 1988 has there been a concerted effort to model the entire 3D in-bore travel phase, including barrel/projectile interactions. This has been made possible by

the availability of supercomputers such as the CRAY XMP and CRAY YMP. We hasten to add, however, that even with CRAY-class machines, some aspects of the 3D transient interior ballistic problem will severely tax current computing capabilities. We also point out that reliance on transient analyses in the present context is not necessarily an absolute must, particularly in early design studies of a new projectile concept. Although the five to ten millisecond time span of the loadings is short compared to everyday experience, a quasi-static stress analysis conducted at peak pressure conditions will often suffice for "first-cut" design purposes.

The ultimate goal of the present 3D modeling effort is to reduce the dispersion of KE rounds on targets. We believe that the key to achieving this goal is to understand, and thus be in a position to control, the perturbing influences which operate on the gun-projectile combination during the in-bore phase of the firing cycle. Ultimately it is the vibration imparted to the sabot/rod combination during launch that is of concern. The sabot (which is discarded soon after muzzle exit) is composed of three or four lengthwise petals and serves to: (1) Provide axial and lateral support to the rod during launch; (2) Seal off the high pressure gun gases behind the projectile; and (3) Grip the grooved surface of the rod and transfer axial forces to it across the sabot/rod interface during in-bore acceleration. To maximize rod velocity, the parasitic weight of the sabot should be minimized, but not at the expense of sabot strength.

MODELING CONSIDERATIONS AND METHODS. No real gun barrel is ever perfectly straight or perfectly rigid, thus even if rigidly clamped against all lateral motion, the KE projectile travels a slightly curved path, but at very high velocity, and is set in vibratory motion. In our 120-mm simulations the barrel is smooth-bore, is cantilevered at the breech end, and is fixed against axial recoil. Obviously, this is a rather simplified model of a real tank situation where complex recoil motions, breech block CG offsets (which cause overturning moments to be imposed on the structure), rigid body rotations of the barrel/breech assembly about the trunnions, and rigid body motions of the entire tank body are possible. The focus of the present work is limited to understanding cause-effect relationships between muzzle exit motions of the projectile and barrel/projectile interactions during in-bore travel.

The main sources of deviations in barrel straightness are gravity droop and inherent variations. Inherent variations include machining irregularities, erosion and wear seen during service, and thermoelastic deformations due to nonuniform patterns of heating and cooling in the structure at the time of firing. It should be kept in mind that inherent variations can vary considerably in a population of barrels. Gravity droop profiles are essentially constant over a population of similar-geometry barrels and can be computed quasistatically with great accuracy with FE codes such as ABAQUS (Hibbett, et al. 1985). The inherent centerline straightness is unique for each

barrel and must be determined by field measurements.

Unlike artillery shells which are rigid and therefore essentially immune to lateral loads, large length/diameter ration KE rounds are light and flexible and thus susceptible to lateral loads. A rough estimate of the peak lateral acceleration for a typical KE round is in the range 500-5000g's. Our experience has been that to obtain more refined estimates than this requires accurate modeling of the specific barrel, bore straightness profile, and projectile. The RASCAL and SHOGUN beam codes are useful for this purpose. The lateral loading can thus not be ignored in launch survival of KE rounds. Moreover, we believe the dynamic motion imparted to the projectile in-bore by these lateral loads influences dispersion.

BRIEF REVIEW OF 2D TRANSIENT IN-BORE RESULTS. The DYNA2D FE code, running on a Cray XMP/48 at BRL, was used to model the dynamic 2D response of the 120-mm configuration of Figure 1. All-elastic material properties were assumed for the components. No barrel was included and no sliding interfaces between parts were included (i.e., the sabot was "welded" to the rod) and no relative motion of the sabot petals was allowed. Also the plastic obturator was not modeled, although an estimated pressure history due to obturator/barrel engagement was applied. Figure 3 shows the base pressure-time history. Axial stress responses at the points A, B, C are shown in Figure 4 to underscore two important points: (1) The axial stress is the dominant stress component; and (2) A sign change in the axial stress (tension to compression) occurs within the rod. Once again, 2D transient analyses are useful for investigating structural integrity of the projectile during early in-bore motion, i.e., up until the time of peak pressure. During this time projectile travel and velocity remain relatively small so that 3D lateral motions due to interactions with the barrel are also small.

3D PROBLEMS (GUN ACCURACY-RELATED WORK). The second and third authors have carried out extensive 3D numerical modeling efforts on both 105-mm and 120-mm tank gun problems. The second author (D. Rabern) was the first to demonstrate the feasibility of accurately modeling with DYNA3D the projectile/barrel interactions for a 120-mm M829 KE projectile. This work also included extensive experimental firings to get data on the in-bore response of the M829 for model validation studies. Unfortunately, only the briefest coverage can be given here of the enormous amount of 3D modeling work that has been done in the 120-mm arena, hence reference is made to the PhD thesis and LANL reports (Rabern 1988, 1989, 1991) for more complete discussions. Comparisons of predicted and X-ray photographed rod profiles at the same barrel location were carried out to validate the 3D modeling procedures. In early stages of this work, the DYNA3D modeling exploited vertical symmetry as shown in the DYNA3D mesh of Figure 2; more recently this restriction was relaxed and full 360-degree grids have been used, both with the DYNA3D and PRONTO3D calculations. Mesh sensitivity studies with

the half-symmetry model indicated that at least 5000 elements (7000 nodes) were needed for accurate displacement data; run times on the LANL Cray XMP/416 were on the order of 6 CPU hours. The 360-degree grids required approximately 11000 elements and 14000 nodes; and here the runs were made on a Cray YMP, requiring approximately 9-15 CPU hours. Nonlinear material properties were used and sliding surfaces were defined between the barrel and projectile,and between sabot petals. No sliding surfaces were defined between the sabot and rod.

As may be evident after careful inspection of the grid in Figure 2, mesh fineness compromises had to made as compared to the 2D mesh of Figure 1. Namely, a relatively coarse 3D barrel mesh was used and a medium-coarse mesh for the projectile. Table 1 summarizes comparisons of DYNA3D-computed and experimental rod tip and tail lateral displacements at two locations along the barrel for the M829 fired out of a barrel model with a realistic bore straightness profile. In these calculations, and for two other M829 sabot design variations, the displacements were within 5-10% of experimental values; comparable agreement of computed/experimental values was obtained with full 360-degree models. The good agreement achieved here validates the 3D finite element modeling procedures we have developed for handling transient barrel/projectile interaction problems. This means that we have capabilities to model and investigate cause-and-effect relationships of in-bore KE round dynamics upon dispersion.

During 1990, in response to an internal BRL request, the third author (S. Wilkerson) initiated studies of a new sabot concept for the 105-mm XM900 projectile. 3D FE models of both the existing XM900 sabot design (building on previous modeling work done at LANL) and a new webbed sabot configuration were required. Aside from the complex modeling task involved in setting up the projectile models, the rifled 105-mm barrel had to be handled. This was accomplished by giving the barrel mesh the same twist angle as that of the rifling in the actual barrel; no attempt was made to model the details of the rifling lands and grooves. Figure 5 shows the straight and twisted barrel grids. Figure 6 shows the computational grids for the standard and webbed sabot design projectiles.

Of particular interest in the 105-mm studies was the extraction of data on rigid body motions of the projectile at muzzle exit. The purpose of the webbed sabot design was to increase the lateral stiffness of the projectile, but staying as close as possible to the XM900 weight (See Figure 7). By increasing lateral stiffness, perhaps the projectile's sensitivity to lateral forces in-bore could be mitigated. By reducing lateral vibration in-bore, then magnitudes of the affected jump components at muzzle exit could be reduced and thus dispersion on target also reduced. Plostins, et al. (1989) have identified (1) muzzle pointing angle, (2) muzzle crossing velocity, and (3) projectile CG jump at muzzle as the major contributions to dispersion attributable to in-bore causes.

Table 1

Comparison of DYNA3D-Predicted and Measured 120-mm KE Projectile Tip and Tail Displacements			
	Axial Location Back from Muzzel (Inches)		
	66	58	51
Predicted Tip Displacement	0.042	0.037	0.018
Measured Tip Displacement	0.048	0.043	0.025
Predicted Tail Displacement	-0.016	0.004	0.032
Measured Tail Displacement	-0.011	0.007	0.036

Two important results were also found from the 3D dynamic analysis of the XM900 projectile. The first was that the bending of the projectile can be significant depending mainly on the design of the sabot/long rod system. For example, using von Mises stress as a criteria of how close to yielding the projectile is during its acceleration down the gun tube, areas of concern can be easily identified. As it were, an axisymmetric analysis of the projectile will identify these areas of high stress with a fair degree of accuracy. However, an axisymmetric analysis does not take into account the projectiles traverse loading during its in bore travel. Such loading as introduced by an unbalanced breech, barrel droop due to gravity, tube heating, or bent gun tubes (Gun tubes are never perfectly straight) are not considered in an axisymmetric analysis. Therefore, when an equivalent 3D analysis is compared with an axisymmetric computation, the results from the three dimensional calculation reveal details which can not otherwise be obtained. These type of results are summarized in Figure 8. where it can be seen that bending in the projectile can lead to higher stresses than would have been predicted by an equivalent axisymmetric analysis.

A second important result that can be found from the 3D analysis is the state of the projectiles traverse motion. Questions like, how fast is the projectile moving downward or outward and what is the projectiles rigid body motion, can now be

addressed. Moreover, by changing the initial conditions slightly, let say cocking one projectile slightly up and one down, variations in the state of the projectile's rigid body motion at shot exit have been observed. In reality a projectile has some clearances between the front bell and gun tube to ease needed loading and unloading operations. This small clearance then allows the projectile to seat slightly off line with the centerline of the gun tube and the projectiles alignment is more or less a random function of loading. By using 3D transient analysis techniques the differences between these initial seating conditions can be measured in terms of projectile velocity variations at shot exit. Therefore, these variations in velocity can be equated directly to a loss of accuracy. By better understanding the mechanisms leading to a loss of accuracy the analyst now has a unique opportunity to improve a projectiles initial design and test his theory prior to its manufacturing.

CONCLUSIONS. For the first time, 3D transient finite element modeling techniques are being successfully applied to solving extremely difficult problems in tank gun saboted-rod kinetic energy ammunition design. The transient KE projectile/barrel interactions of actual rifled 105-mm and smoothbore 120-mm tank gun systems have been modeled. 2D models provide useful information about the axial performance of saboted rod systems. A problem posed by transient 3D FE analyses is the huge amount of data generated that must be interpreted to glean useful performance information. This problem can be mitigated by judicious post-processing using, for example, computer animation techniques to present the data. Even with the present generation of supercomputers, compromises must still be made in 3D analyses due to CPU time and storage limitations. In the present context, these limitations are not particularly serious but do place practical restrictions on how much of tank system can be modeled.

A final comment on barrel/projectile interaction modeling is in order. Just how detailed the 2D or 3D modeling of the barrel/projectile interaction really needs to be remains an open matter. Whether solid continuum finite elements or even beam element models will be adequate, depends on the application. BRL and LANL have in fact assembled a hierarchy of barrel/projectile interaction models of differing levels of sophistication. These range from the RASCAL and SHOGUN beam-element codes running on PC's, to full transient continuum codes such as DYNA2D or DYNA3D and PRONTO3D running on Cray supercomputers.

REFERENCES

Erline, T. F., Kregel, M. D., and Pantano,M, "Gun and Projectile Flexural Dynamics Modeled by the Little Rascal - A User's Manual," BRL-TR-3122, U.S. Army Ballistic Research Laboratory, MD, July 1990.

Hallquist, J., "User's Manual for DYNA2D -- An Explicit Two-Dimensional Hydrodynamic Finite Element Code with Interactive Rezoning." LLNL Report UCID-18756 Rev. 2, Lawrence Livermore Laboratory, Livermore, CA, January 1984.

Hallquist, J. and Benson, D. J.,"DYNA3D User's Manual." LLNL Report UCID-19592 Rev. 2, Lawrence Livermore Laboratory, Livermore, CA, March 1986.

Hibbett, Karlsson, & Sorenson, Inc., ABAQUS User's Manual Version 4.5a, Providence, RI, 1985.

Hopkins, D. A., "SHOGUN - 3-D Gun Dynamics User's Manual," BRL-TR-3128, U.S. Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, August 1990.

Plostins, P., Celmins, I., and Bornstein, J., "The Effect of Sabot Front Borerider Stiffness on the Launch Dynamics of Fin-Stabilized Kinetic Energy Ammunition," in Proceedings of the 11th International Symposium on Ballistics, Vol. I, pp. 535-549, Royal Military Academy, Brussels, Belgium, May 1989.

Rabern, D., "Axially Accelerated Saboted Rods Subjected to Lateral Forces." PhD Dissertation, University of Arizona, 1988.

Rabern, D., "Axially Accelerated Saboted Rods Subjected to Lateral Forces." LANL Report LA-11494-MS, Los Alamos National Laboratory, Los Alamos, NM, March 1989.

Rabern, D., "Numerical Simulations of Gun-Launched Kinetic Energy Projectiles Subjected to Axisymmetric Base Pressure." LANL Report MEE13-91-445, Los Alamos National Laboratory, Los Alamos, NM. July 1991.

Taylor, L. M., and Flanagan, D. P., "PRONTO3D - A Three-Dimensional Transient Solid Dynamics Program," Sandia Report SAND87-1912, Sandia National Laboratories, Albuquerque, NM, March 1989.

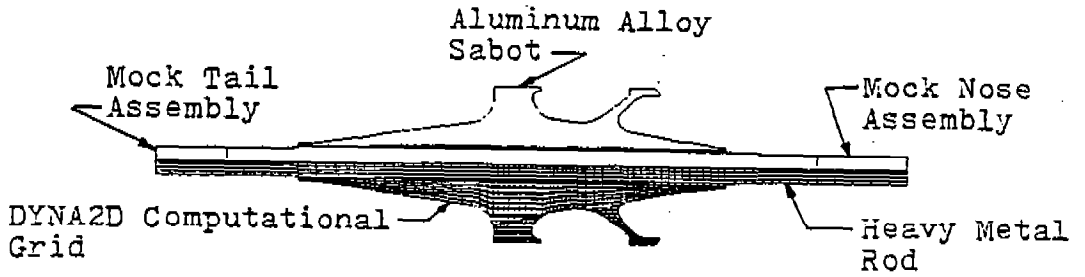


Figure 1. Representative 120mm Saboted Tank Gun Round

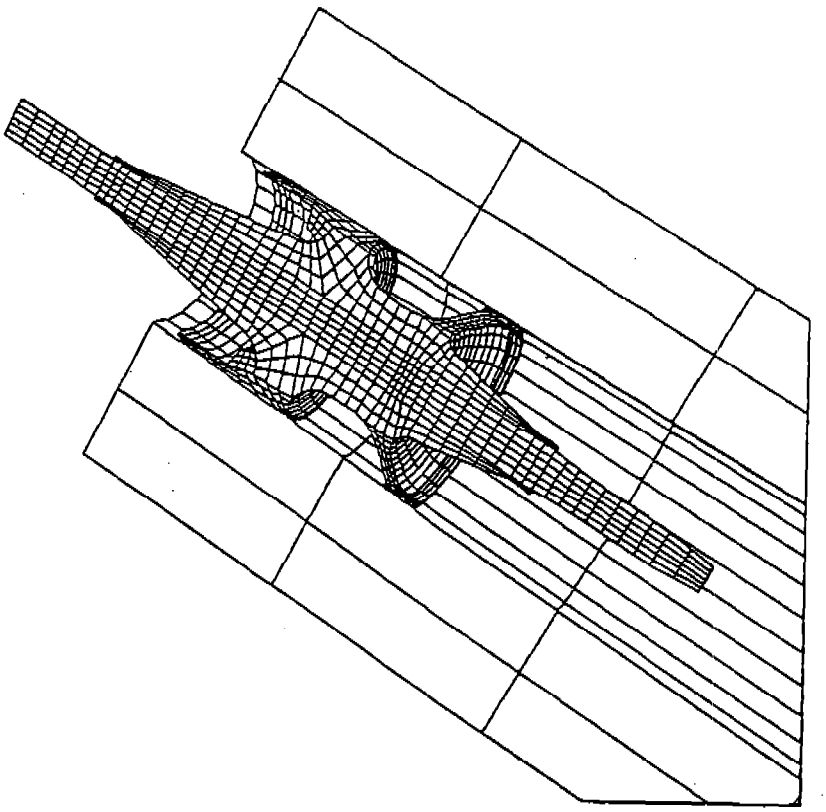


Figure 2. M829 Finite Element Mesh and SN81 Launch Tube

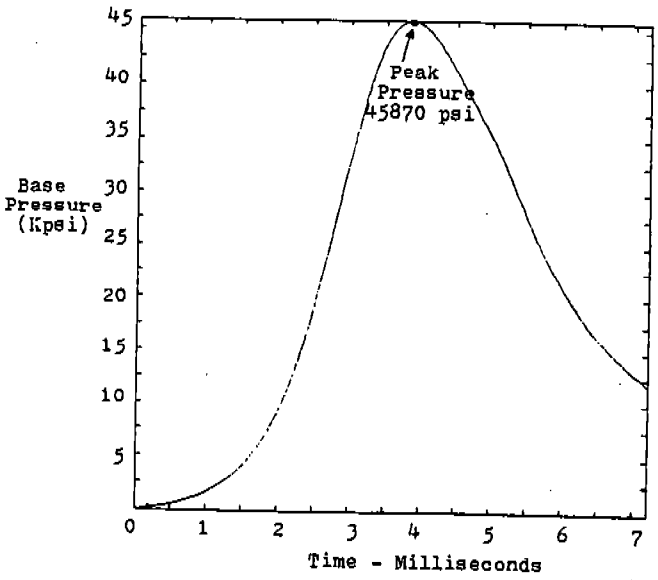


Figure 3. Base Pressure History

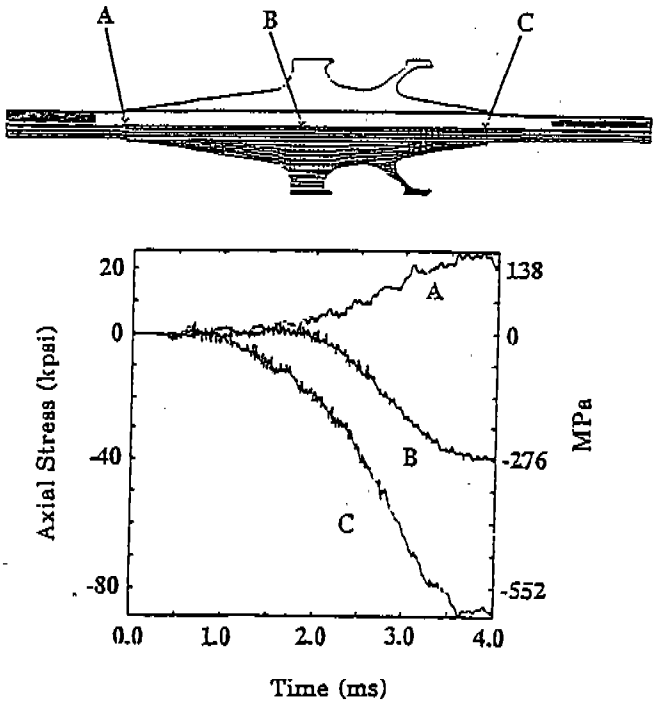
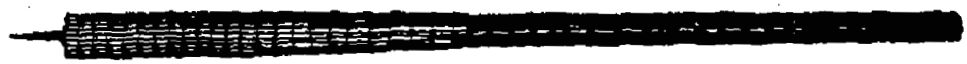
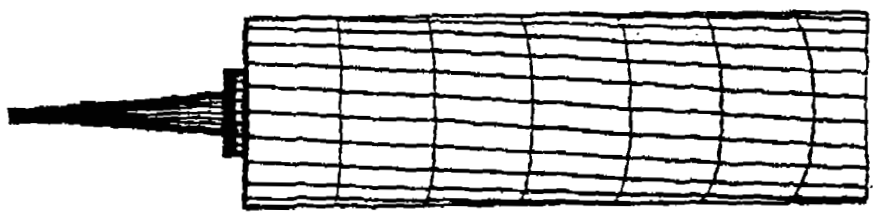


Figure 4. Two-Dimensional Finite Element Analysis



Finite element mesh without spin.

Finite element mesh with spin.



Enlarged view of finite element mesh with spin.

Figure 5. Straight and Twisted Barrel Grids7

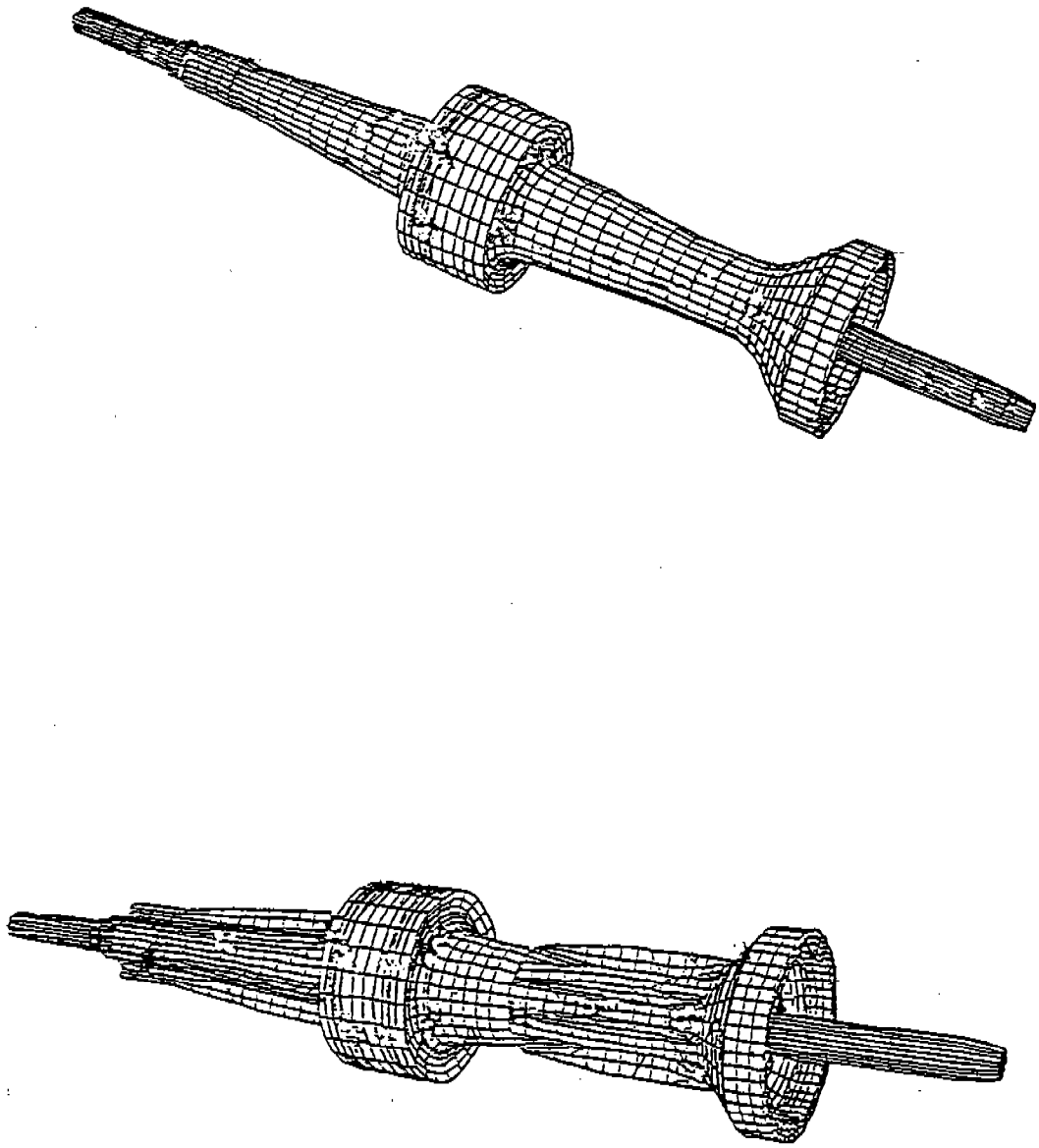


Figure 6. Computational Grids for Standard and Webbed Sabot Design

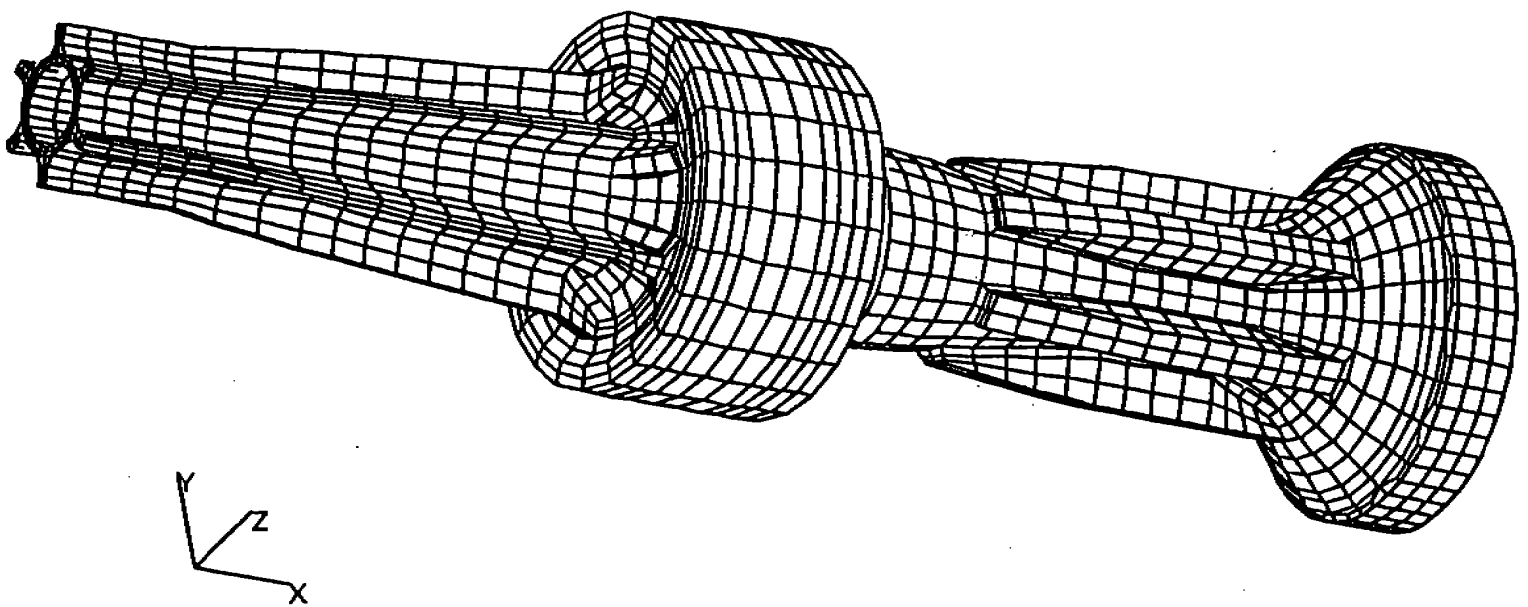


Figure 7. Enlargement of Webbed Sabot Design

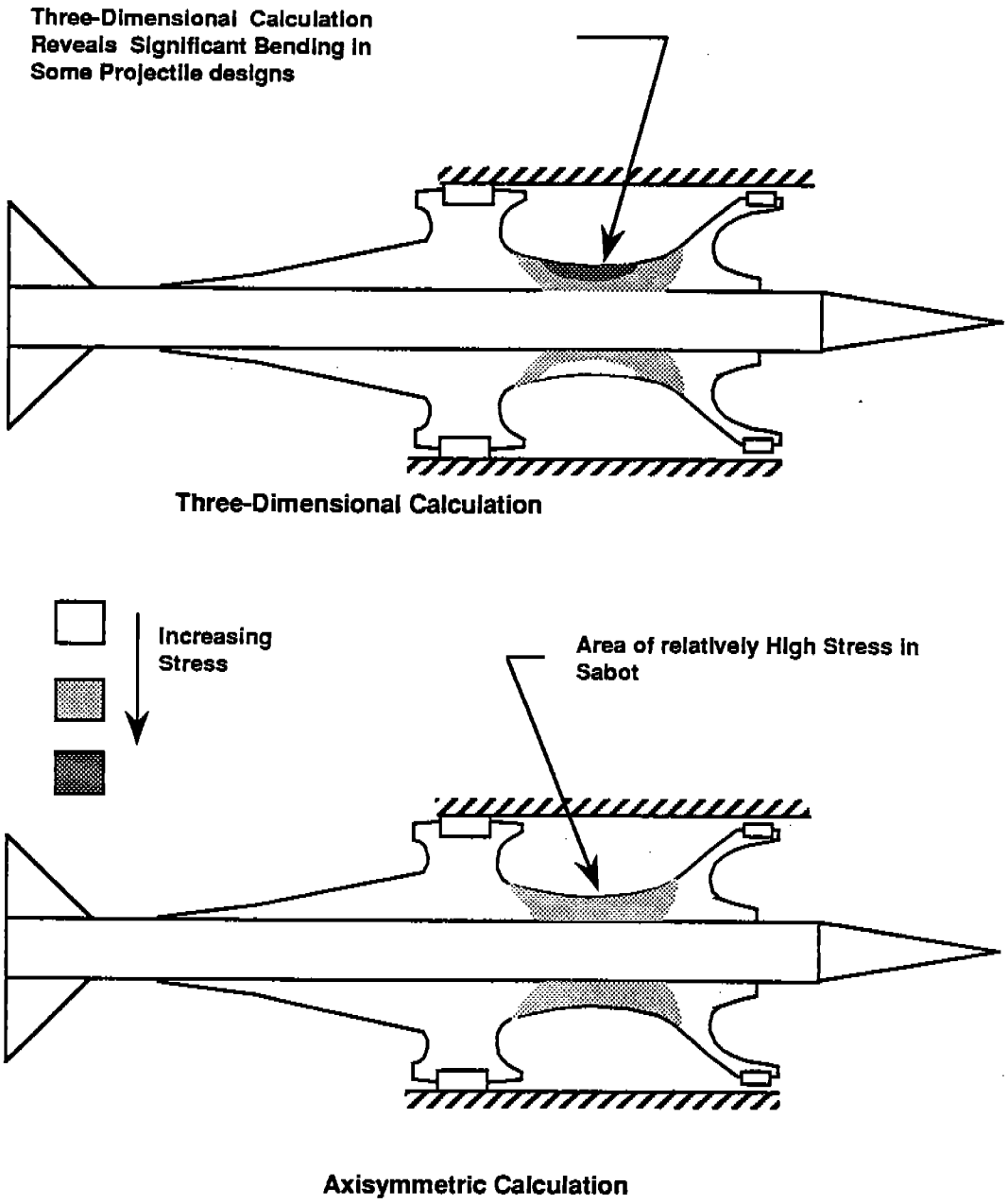


Figure 8. Von-Mises Stress Comparison between Axisymmetric and Three-Dimensional Calculations at Peak Pressure

Computing the PSVD of Two-by-Two Triangular Matrices

Gary E. Adams, Adam W. Bojanczyk and Franklin T. Luk
School of Electrical Engineering
Cornell University
Ithaca, NY 14853, USA

Abstract

In this paper, we propose a method for computing the SVD of a product of two 2×2 triangular matrices. We show that our method is numerically desirable in that all relevant residual elements will be numerically small.

1. Introduction

The problem of computing the singular value decomposition (SVD) of a product of two matrices has many applications; see, e.g., [4] and [5]. The problem is also closely related to finding a generalized SVD of two matrices (cf. [6]). A crucial step in either the product SVD (PSVD) or the generalized SVD (GSVD) problem is the accurate computation of the PSVD of two 2×2 triangular matrices.

We wish to achieve two objectives: first, to ensure that the transformations applied to the triangular matrices must leave the matrices triangular, and second, to ensure that the product of the transformed matrices must be diagonal. As discussed in a recent paper by Bai and Demmel [1], these two properties are essential to guarantee stability of the GSVD method [6]. Several strategies have been proposed to preserve these two properties. In [1] examples are presented where these strategies can fail, and a new method that overcomes the exposed drawbacks is then proposed.

In this paper we propose an alternative approach. Our new method, which we will call a *half-recursive* method, is a slight variation of the *fully-recursive* method proposed in [2] for computing the SVD of a product of several matrices. We show that while our algorithm enjoys the same nice numerical properties as the one in [1], it is simpler to implement.

Our paper is organized as follows. In Section 2 we describe the PSVD of two 2×2 upper triangular matrices. A criterion for numerical stability is given in Section 3. We present our new algorithm in Section 4, and an error analysis in Section 5. Finally, some detailed proofs can be found in Appendices A and B, and a numerical example in Appendix C.

2. Problem Definition

Given two upper triangular matrices:

$$A_1 = \begin{pmatrix} a_1 & b_1 \\ 0 & d_1 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} a_2 & b_2 \\ 0 & d_2 \end{pmatrix},$$

we call the product A:

$$A = A_1 A_2 ,$$

and let

$$A = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} .$$

Our objective is to find three orthogonal matrices Q_1, Q_2, Q_3 such that

$$A' = Q_1 A Q_3^T = \begin{pmatrix} a' & 0 \\ 0 & d' \end{pmatrix} \tag{2.1}$$

and

$$A_i' = Q_i A_i Q_{i+1}^T = \begin{pmatrix} a_i' & b_i' \\ 0 & d_i' \end{pmatrix} , \tag{2.2}$$

for $i = 1, 2$. The two equations (2.1) and (2.2) imply that

$$A' = A_1' A_2' .$$

In words, we would like to find *three* transformations Q_1, Q_2 and Q_3 to zero out *four* elements, namely, the off-diagonal elements of A and the sub-diagonal elements of A_1 and A_2. The extra requirement, although mathematically feasible, may cause numerical difficulty if not treated with care; see examples in [1] and [2]. Our goal is to develop an algorithm so that properties (2.1) and (2.2) will be satisfied except for very small numerical errors. In this paper, we use the vector and matrix 2-norms:

$$\| \cdot \| = \| \cdot \|_2 .$$

2.1. Relationship with GSVD

The basic step in a GSVD of two 2×2 triangular matrices A_1 and A_2 is to compute the SVD of the product $A_1 \cdot adjoint(A_2)$, where

$$adjoint(A_2) = \begin{pmatrix} d_2 & -b_2 \\ 0 & a_2 \end{pmatrix} .$$

It is therefore obvious that our two-by-two PSVD method can also be applied to the two-by-two GSVD problem.

3. Criterion for Numerical Stability

Recall that A_1', A_2' and A' denote the three matrices A_1, A_2 and A, respectively, after the equivalence transformations as defined in (2.1) and (2.2) have been performed. Let ϵ denote the relative precision of the floating-point arithmetic, and let $\bar{A}_1'$, $\bar{A}_2'$ and $\bar{A}'$ represent the computed A_1', A_2' and A', respectively. We want the product A' to be diagonal:

$$A' = A_1' A_2' = \begin{pmatrix} a' & 0 \\ 0 & d' \end{pmatrix} . \tag{3.1}$$

Assume that, given the exact upper triangular matrices A_i', for $i = 1, 2$, we compute using floating-point arithmetic the product:

$$\bar{A}' := \mathrm{fl}(\prod_{i=1}^{2} A_i') . \tag{3.2}$$

Due to rounding errors, we can hope for only

$$\bar{A}' = \begin{pmatrix} \bar{a}' & \bar{b}' \\ 0 & \bar{d}' \end{pmatrix} , \tag{3.3a}$$

where $\bar{b}'$ satisfies the relation:

$$| \bar{b}' | = O(\epsilon f') . \tag{3.3b}$$

The quantity f', defined by

$$f' = |a_1'| \, |b_2'| + |b_1'| \, |d_2'| , \tag{3.3c}$$

provides an upper bound on the rounding error for $\bar{b}'$. Thus, the best that we can aim for is to compute $\bar{A}_i'$ such that

$$\| \bar{A}_i' - A_i' \| = O(\epsilon) , \tag{3.4a}$$

with $\tilde{b}'$ satisfying

$$| \tilde{b}' | = O(\epsilon \tilde{f}') , \tag{3.4b}$$

and

$$\tilde{f}' = |\tilde{a}_1'| \, |\tilde{b}_2'| + |\tilde{b}_1'| \, |\tilde{d}_2'| . \tag{3.4c}$$

The relation (3.4a) implies that the (2,1) element $\tilde{e}_i'$ of $\tilde{A}_i'$ will satisfy

$$| \tilde{e}_i' | = O(\epsilon \ \| A_i \|) , \tag{3.5}$$

for $i = 1, 2$.

We prove in Section 5 that by using our new method, the computed matrices $\tilde{A}_1'$ and $\tilde{A}_2'$ will satisfy condition (3.5) and $\tilde{A}'$ will satisfy a condition somewhat weaker than (3.4b), namely that

$$|\tilde{b}'| \leq \epsilon ||\tilde{A}|| . \tag{3.6}$$

The conditions (3.5) and (3.6) are equivalent to the conditions proposed in [1] for computing the GSVD of the two matrices A_1 and $adjoint(A_2)$.

4. New Algorithm

In this section, we propose a new algorithm for the PSVD problem. Our algorithm is a modification of the algorithm presented in [2] for a product of several matrices. The tool we use is a transformation discussed in Charlier et al. [3]:

$$Q = \begin{pmatrix} s & c \\ -c & s \end{pmatrix} , \tag{4.1}$$

where $c^2 + s^2 = 1$. We may regard the transformation as a permuted reflection:

$$Q = \begin{pmatrix} c & s \\ s & -c \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} .$$

The reason behind using permuted reflections is that we actually deal with an $n \times n$ problem. The permutation that is incorporated into Q corresponds to the so called odd-even order of eliminations in one sweep of a Jacobi-SVD procedure.

While each transformation Q_i is defined by the cosine-sine pair:

$$c_i = \cos\theta_i \quad \text{and} \quad s_i = \sin\theta_i \ ,$$

we also associate Q_i with the tangent

$$t_i = \tan\theta_i \ .$$

Given t_i, we can easily recover c_i and s_i using the relations

$$c_i = \frac{1}{\sqrt{1+t_i^2}} \quad \text{and} \quad s_i = t_i c_i \ . \tag{4.2}$$

Following the exposition in [2] we consider the result of applying the left and right transformations Q_l and Q_r to a 2×2 upper triangular matrix A:

$$A' = Q_l A Q_r^T = \begin{pmatrix} a' & b' \\ e' & d' \end{pmatrix} = \begin{pmatrix} s_l & c_l \\ -c_l & s_l \end{pmatrix} \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \begin{pmatrix} s_r & c_r \\ -c_r & s_r \end{pmatrix}^T . \tag{4.3}$$

We can derive from (4.3) these four relations:

$$e' = c_l c_r(-at_r + dt_l - b) \ , \tag{4.4a}$$

$$b' = c_l c_r(-at_l + dt_r + bt_l t_r) \ , \tag{4.4b}$$

$$a' = c_l c_r(bt_l + d + at_l t_r) \ , \tag{4.4c}$$

$$d' = c_l c_r(a - bt_r + dt_l t_r) \ , \tag{4.4d}$$

where $t_l = \tan\theta_l$ and $t_r = \tan\theta_r$. The postulates that both e' and b' be zeros define two conditions on t_l and t_r, so that (4.3) represents an SVD of A. The postulate that e' be zero defines a condition relating θ_l to θ_r, so that if one is known the other can be computed in order to reduce A' to an upper triangular form. For ease of exposition, assume for now that $abd \neq 0$; this condition will be removed in Section 5.2. This assumption implies that $c_l c_r \neq 0$, and so the postulate that $e' = 0$ in (4.4a) becomes

$$-at_r + dt_l - b = 0 \ . \tag{4.4e}$$

The consequence of (4.4e) is that (4.4c) and (4.4d) simplify to

$$a' = c_l c_r(t_l^2 + 1)d \ , \tag{4.4f}$$

and

$$d' = c_l c_r(t_r^2 + 1)a \ , \tag{4.4g}$$

respectively. The relations (4.4f) and (4.4g) imply that

$$a'd' = ad \ .$$

For the SVD problem, both e' and b' are zeros, and we can use (4.4e) to reduce (4.4b) either to an equation in t_l:

$$b' = c_l c_r\left(\frac{bd}{a}\right)\left(t_l^2 + 2t_l\sigma_l - 1\right) \ , \tag{4.5a}$$

where

$$\sigma_l = \frac{1}{2d}\left(\frac{d^2 - a^2}{b} - b\right) \ ,$$

or to an equation in t_r:

$$b' = c_l c_r \left(\frac{ab}{d}\right)\left(t_r^2 + 2t_r\sigma_r - 1\right) , \tag{4.5b}$$

where

$$\sigma_r = \frac{1}{2a}\left(\frac{d^2 - a^2}{b} + b\right) .$$

From (4.5a) we get a quadratic equation by setting b' to zero:

$$t_l^2 + 2\sigma_l t_l - 1 = 0 , \tag{4.5c}$$

and from (4.5b) we get

$$t_r^2 + 2\sigma_r t_r - 1 = 0 . \tag{4.5d}$$

The two equations (4.5c) and (4.5d) are solved by the formulas given in [2]:

$$r = \frac{(d-a)(d+a)}{b} , \tag{4.6a}$$

$$\sigma_l = \frac{r-b}{2d} , \tag{4.6b}$$

$$\sigma_r = \frac{r+b}{2a} , \tag{4.6c}$$

$$t_l = \frac{1}{\sigma_l + \mathrm{sign}(\sigma_l)\sqrt{\sigma_l^2 + 1}} , \tag{4.6d}$$

$$t_r = \frac{1}{\sigma_r + \mathrm{sign}(\sigma_r)\sqrt{\sigma_r^2 + 1}} . \tag{4.6e}$$

In finite-precision arithmetic, either one of t_l and t_r can be computed with a higher relative precision. In particular, if

$$\mathrm{sign}(r) = -\mathrm{sign}(b) ,$$

then (4.6d) will produce a very accurate t_l, whereas if

$$\mathrm{sign}(r) = \mathrm{sign}(b) ,$$

then (4.6e) will produce a very precise t_r. If $r = 0$, then both t_l and t_r will be computed with the same relative accuracy.

Now, let $r \neq 0$. We first present a lemma relating the sizes of t_l and t_r to those of a and d.

Lemma 4.1. Let $abdr \neq 0$. If $| a | > | d |$, then $| \sigma_l | > | \sigma_r |$ and $| t_l | < | t_r |$. Conversely, if $| a | < | d |$, then $| \sigma_l | < | \sigma_r |$ and $| t_l | > | t_r |$.

Proof. See [2]. □

We are ready to present an algorithm for computing the three orthogonal matrices Q_1, Q_2 and Q_3, such that (2.1) and (2.2) are satisfied. The algorithm proceeds in two stages. In the first stage, we calculate the product A explicitly:

$$a = a_1 a_2 , \tag{4.7a}$$

$$b = a_1 b_2 + b_1 d_2 , \tag{4.7b}$$

$$d = d_1 d_2 \ . \tag{4.7c}$$

We use (4.6a) to calculate r, and then compute either σ_l or σ_r so that the corresponding tangent defines the smaller angular rotation. Hence we obtain either t_1 or t_3. In the second stage we use the relation (4.4e) with t_1 or t_3 as the reference tangent to compute the remaining transformations. Suppose that t_1 is known. Then t_2 and t_3 are generated by the forward substitutions:

$$t_2 = \frac{d_1 t_1 - b_1}{a_1} \ , \tag{4.8a}$$

$$t_3 = \frac{d t_1 - b}{a} \ . \tag{4.8b}$$

On the other hand, if t_3 is known, then t_2 and t_1 are generated by the backward substitutions:

$$t_2 = \frac{a_2 t_3 + b_2}{d_2} \ , \tag{4.8c}$$

$$t_1 = \frac{a t_3 + b}{d} \ . \tag{4.8d}$$

If t_1 is computed first as the reference tangent, then (4.8a) will guarantee that A_1' will be numerically upper triangular and (4.8b) will guarantee that A' will be numerically diagonal. As will be shown later these two properties will guarantee that A_2' will be numerically upper triangular and hence both (3.5) and (3.6) will be satisfied.

It appears that the half-recursive method is equivalent to the method proposed by Bai and Demmel in [1] in the sense that it also computes a very accurate PSVD of $A_1 A_2$, and that it uses essentially the same criterion in deciding whether the middle transformation Q_2 should be computed from Q_1 or Q_3. A proof that the two methods use the same condition for computing Q_2 is given in Appendix B.

We refer to the method defined by (4.8a)-(4.8b) or (4.8c)-(4.8d) as *half-recursive*, to differentiate it from the *fully-recursive* method proposed in [2] for computing the PSVD of several matrices. The fully-recursive method also picks the smaller outer angular rotation as the starting point for the recursion, from which all remaining rotations are computed. However, there the other outer rotation is computed from the previous rotation in the sequence. For example, in the case of a product of two matrices, the tangent t_3 in (4.8b) would be computed from t_2 using (4.4e):

$$t_3 = \frac{d_2 t_2 - b_2}{a_2} \ . \tag{4.9}$$

Note how (4.8b) uses the product A whereas (4.9) uses the matrix A_2. It was shown in [1] that the fully-recursive method may fail to satisfy (3.6) and thus is not recommended for the GSVD problem. On the other hand, the fully-recursive method easily extends to any number of factors in the product. It is not clear what is an appropriate extension of the half-recursive method for the case of a product of more than two matrices.

5. Backward Error Analysis

In this section, we present a backward error analysis of our computation. We assume that our initial parameters are perturbed, and use the "bar" symbol. For example, instead of initial values

a, b and d, we have the perturbed values $\bar{a}$, $\bar{b}$ and $\bar{d}$. We assume further that exact arithmetic will be performed by using these perturbed initial values. We use the "tilde" symbol for the exact values based on the perturbed data. For example, $\tilde{r}$ will denote the exact result using formula (3.8a) for the perturbed data $\bar{a}$, $\bar{b}$ and $\bar{d}$.

The symbol $\mathrm{fl}(a)$ will be used to denote the computed result of the parameter a. In our error analysis, we adopt a convention that involves a liberal use of Greek letters. For example, by α we mean a relative perturbation of an absolute magnitude not greater than ϵ, where ϵ denotes the machine precision. All terms of order ϵ^2 or higher will be ignored.

We start our procedure by computing elements of the product matrix A. For the elements of the computed product matrix A we have

$$\bar{a} := \mathrm{fl}(a_1 a_2) = a_1 a_2(1+\alpha)\,, \tag{5.1a}$$

$$\bar{d} := \mathrm{fl}(d_1 d_2) = d_1 d_2(1+\delta)\,, \tag{5.1b}$$

$$\bar{b} := \mathrm{fl}(a_1 b_2 + b_1 d_2) = a_1 b_2(1+2\beta_1) + b_1 d_2(1+2\beta_2)\,, \tag{5.1c}$$

where, according to our convention, the parameters α_1, δ_1, β_1, β_2, and β_3 are all quantities whose absolute values are bounded by ϵ. From (5.1) it follows that

$$\bar{A} = (A_1 + \delta A_1)(A_2 + \delta A_2)\,,$$

with $\| \delta A_i \| \leq \epsilon \ \| A_i \|$. This property, which in general does not hold for a product of more than two 2×2 upper triangular matrices, will allow us to prove backward error type assertions on the half-recursive method.

Our analysis is divided into two parts. In Section 5.1, we consider a regular case where all elements of the computed matrix product are numerically significant with respect to the maximal-in-magnitude element, i.e.,

$$\min(|\,\bar{a}\,|\,,\,|\,\bar{b}\,|\,,\,|\,\bar{d}\,|\,) > \epsilon\, \max(|\,\bar{a}\,|\,,|\,\bar{b}\,|\,,|\,\bar{d}\,|\,)\,. \tag{5.2}$$

In Section 5.2, we consider special cases where at least one element of the computed A is numerically insignificant.

5.1. Regular Case

Without loss of generality we assume that $rb < 0$, i.e., $\mathrm{sign}(r) = -\mathrm{sign}(b)$. Thus we compute t_1 first as the reference tangent from which t_2 and t_3 will be next determined via (4.8a) and (4.8b) respectively. We recall several lemmas from [2].

Lemma 5.1. Let $\tilde{t}_1$ and $\bar{t}_1$ be the exact and computed solutions, respectively, of equation (3.7c) with data $\bar{a}, \bar{b}, \bar{d}$. Moreover, let $\tilde{c}_1, \tilde{s}_1$ and $\bar{c}_1, \bar{s}_1$ be the exact and computed cosines and sines using (3.4) with the tangent value $\bar{t}_1$. Then

$$\bar{t}_1 = \tilde{t}_1(1+10\epsilon_1)\,, \tag{5.3a}$$

$$\bar{c}_1 = \tilde{c}_1(1+3\mu_1)\,, \tag{5.3b}$$

$$\bar{s}_1 = \tilde{s}_1(1+4\nu_1)\,, \tag{5.3c}$$

where $|\epsilon_5| < \epsilon$, $|\mu_1| < \epsilon$, and $|\nu_1| < \epsilon$.

Proof. See [2]. □

In words, Lemma 5.1 states that the procedure (4.6a)–(4.6e) for solving (4.5c) is numerically stable in the forward sense. Three lemmas follow, leading to our main result of Theorem 5.1.

Lemma 5.2. The recurrences (4.8a) and (4.8b) yield $\bar{t}_2$ and $\bar{t}_3$ such that

$$\tilde{a}_1\bar{t}_2 - \tilde{d}_1\bar{t}_1 + b_1 = 0\,, \tag{5.4a}$$

$$\tilde{a}\bar{t}_3 - \tilde{d}\bar{t}_1 + \bar{b} = 0\,, \tag{5.4b}$$

with

$$\tilde{a}_1 = a_1(1+2\psi_1)\,,\ \tilde{d}_1 = d_1(1+\phi_1)\,, \tag{5.4c}$$

$$\tilde{a} = \bar{a}(1+2\psi)\,,\ \tilde{d} = \bar{d}(1+\phi). \tag{5.4d}$$

Proof. The proof easily follows from (4.8a) and (4.8b). □

Lemma 5.3. The recurrence (4.8b) yields $\bar{t}_3$ such that $\bar{t}_3 = \tilde{t}_3(1+13\gamma)$.

Proof. From (4.8b)

$$\bar{t}_3 = \left(\frac{\tilde{d}\bar{t}_1(1+11\psi)-\bar{b}}{\bar{a}}\right)(1+2\gamma_1) = \left(\frac{\tilde{d}\bar{t}_1-\bar{b}}{\bar{a}} + \frac{11\psi\tilde{d}\bar{t}_1}{\bar{a}}\right)(1+2\gamma_1) = \left(\tilde{t}_3 + 11\psi\tilde{t}_3\frac{\tilde{d}}{\bar{a}}\frac{\bar{t}_1}{\tilde{t}_3}\right)(1+2\gamma_1)\,.$$

Since $|\tilde{d}/\bar{a}| \le 1$ and $|\bar{t}_1/\tilde{t}_3| \le 1$, we get

$$\bar{t}_3 = \tilde{t}_3(1+13\gamma)\,.$$

□

We now show that $\bar{a}'$ and $\bar{d}'$ are computed with high relative precision.

Theorem 5.1. Let $\bar{a}'$ and $\tilde{d}'$ be the exact singular values of the computed product $\bar{A}$. If $\tilde{a}'$ and $\tilde{d}'$ are computed via relations (4.4c) and (4.4d) then the computed singular values $\bar{a}'$ and $\bar{d}'$ satisfy the following relations

$$\bar{a}' = \tilde{a}'(1+\alpha_4)\ ,\ \bar{d}' = \tilde{d}'(1+\delta_4)\,. \tag{5.5}$$

Proof. From (4.4f) and (4.4g), we get

$$\tilde{a}' = \bar{d}(\tilde{t}_1^2+1)\hat{c}_1\hat{c}_3 \quad \text{and} \quad \tilde{d}' = \bar{a}(\tilde{t}_3^2+1)\hat{c}_1\hat{c}_3\,,$$

where $\tilde{t}_1$ and $\tilde{t}_3$ are the exact tangnts corresponding to the data $\bar{a}$, $\bar{b}$ and $\bar{d}$ and $\tilde{t}_i = \hat{s}_i/\hat{c}_i$. Thus, the lemma follows from Lemmas 5.1 and 5.3. □

Theorem 5.2. Suppose that the computed tangent values are $\bar{t}_1$ and $\bar{t}_3$. Let $\tilde{c}_1$, $\tilde{s}_1$, $\tilde{c}_3$ and $\tilde{s}_3$ be the corresponding exact cosine and sine values. Let

$$\tilde{e}' := \tilde{c}_1\tilde{c}_3[-\bar{a}\bar{t}_3 + \tilde{d}\bar{t}_1 - \bar{b}]\,, \tag{5.6}$$

$$\tilde{b}' := \bar{c}_1\bar{c}_3[-\bar{a}\bar{t}_1 + \bar{d}\bar{t}_3 + \bar{b}\bar{t}_1\bar{t}_3] \,. \tag{5.7}$$

That is, $\tilde{e}'$ and $\tilde{b}'$ are the exact values of e' and b', respectively, corresponding to the computed data $\bar{a}$, $\bar{b}$, $\bar{d}$, $\bar{t}_1$ and $\bar{t}_3$. Then

$$| \tilde{e}' | \leq K_1\epsilon \ \| \bar{A} \| \,, \tag{5.8}$$

$$| \tilde{b}' | \leq K_2\epsilon \ \| \bar{A} \| \,, \tag{5.9}$$

where K_1 and K_2 are some positive constants.

Proof. See Appendix A. ■

Theorems 5.1 and 5.2 together state that the SVD of the upper triangular matrix $\bar{A}$ is computed very accurately. We now justify why the (2,1) element in the computed matrix A_i' can be set to zero, by showing that $| \tilde{e}_i' |$ corresponds to a relative and elementwise perturbation of A_i' of the order of ϵ. Let the cosine and sine pairs $\bar{c}_i$ and $\bar{s}_i$ satisfy $\bar{t}_i = \bar{s}_i/\bar{c}_i$, for $i = 1, 2, 3$. From (4.2) we can derive that

$$\bar{c}_i := \mathrm{fl}(\bar{c}_i) = \tilde{c}_i(1 + 3\mu_i) \,, \tag{5.10a}$$

$$\bar{s}_i := \mathrm{fl}(\bar{s}_i) = \bar{s}_i(1 + 4\nu_i) \,. \tag{5.10b}$$

Let $\tilde{A}_i'$ denote the exact updated matrix derived from A_i, $\bar{c}_i$ and $\bar{s}_i$. Our next results provide a bound on the element $\tilde{e}_i'$, $i = 1, 2$, defined by the relation

$$\tilde{e}_i' := -\bar{c}_i\bar{s}_{i+1}a_i + \bar{s}_i\bar{c}_{i+1}d_i - \bar{c}_i\bar{c}_{i+1}b_i \,. \tag{5.11}$$

Theorem 5.3. The matrices $\tilde{A}_1'$ and $\tilde{A}_2'$ are almost upper triangular in that their (2,1) elements $\tilde{e}_1'$ and $\tilde{e}_1'$ satisfy the inequalities:

$$| \tilde{e}_1' | \ \leq 3\,\epsilon \ \| A_1 \| \,. \tag{5.12a}$$

and

$$| \tilde{e}_2' | \ \leq K_3\,\epsilon \ \| A_2 \| \,. \tag{5.12b}$$

Proof. Note that $\tilde{A}_1'$ is the same for both fully-recursive and half-recursive methods. The proof that $\tilde{A}_1'$ is almost upper triangular in the sense that (5.12a) holds can be found in [2].

In order to prove the second part of the theorem note that from (5.4a)-(5.4d) and (5.1a)-(5.1c) we get the following two relations to first order of the machine precision:

$$a_1(1 + 2\psi_1)\bar{t}_2 - d_1(1 + \phi_1)\bar{t}_1 + b_1 = 0 \,, \tag{5.13a}$$

$$a_1a_2(1 + \alpha + 2\psi)\bar{t}_3 - d_1d_2(1 + \delta + \phi)\bar{t}_1 + a_1b_2(1 + 2\beta_1) + b_1d_2(1 + 2\beta_2) = 0 \,, \tag{5.13b}$$

By multiplying both sides of (5.13a) by $d_2(1 + 2\beta_2)$ and subtracting from (5.13b) we obtain

$$a_1\{a_2(1 + \alpha + 2\psi)\bar{t}_3 - (\frac{d_1d_2}{a_1})(\delta + \phi - \phi_1 + 2\beta_2)\bar{t}_1 + b_2(1 + 2\beta_1) - d_2(1 + 2\beta_2 + 2\psi_1)\bar{t}_2\} = 0 \,,$$

or, since $a_1 \neq 0$,

$$a_2(1 + \alpha + 2\psi)\bar{t}_3 - (\frac{d_1d_2}{a_1})(\delta + \phi - \phi_1 + 2\beta_2)\bar{t}_1 + b_2(1 + 2\beta_1) - d_2(1 + 2\beta_2 + 2\psi_1)\bar{t}_2$$

$$= a_2\bar{t}_3 - d_2\bar{t}_2 + b_2 + \Delta = 0\,,$$

where

$$\Delta = a_2(\alpha + 2\psi)\bar{t}_3 - \left(\frac{d}{a}\right) a_2(\delta + \phi - \phi_1 + 2\beta_2)\bar{t}_1 + b_2\beta_1 - d_2(2\beta_2 + 2\psi_1)\bar{t}_2\,.$$

Thus, we can rewrite (5.11) for $i = 2$ as

$$e_2' = -\bar{c}_2\bar{s}_3 a_2 + \bar{s}_2\bar{c}_3 d_2 - \bar{c}_2\bar{c}_3 b_2 + \bar{c}_3\bar{c}_2(a_2\bar{t}_3 - d_2\bar{t}_2 + b_2 + \Delta)\,. \tag{5.13c}$$

Now, as we start the half-recursive method from t_1, it means that $|\bar{t}_1| \leq 1$ and $|\bar{d}| \leq |\bar{a}|$. Hence from (5.10a), (5.10b) and (5.13c), we derive the inequality:

$$|e_2'| \leq |\bar{s}_3\bar{c}_2 a_2(\alpha + 2\psi)| + |\bar{c}_3\bar{c}_2 a_2(\delta + \phi - \phi_1 + 2\beta_2)| + |\bar{c}_3\bar{c}_2 b_2\beta_2| + |\bar{c}_3\bar{s}_2 d_2(2\beta_2 + 2\psi_1)|$$

$$\leq K_3\epsilon||A_2||\,,$$

completing the proof. □

In summary, we have proved two results using backward error analysis. First, the computed matrix product $\bar{A}'$ is almost diagonal in that inequalities (5.8) and (5.9) both hold. Second, we can safely set each computed matrix $\bar{A}_i'$, $i = 1, 2$, to a triangular form because (5.12a) and (5.12b) are valid. As a final note, even though we have assumed that $rb < 0$, we can easily prove similar results for the case where $rb \geq 0$.

5.2. Special Cases

In this subsection, we assume that inequality (5.2) is violated. To be specific, define

$$\gamma := \min(|\,\bar{a}\,|\,,\,|\,\bar{b}\,|\,,\,|\,\bar{d}\,|\,) \tag{5.14}$$

and

$$\Gamma := \max(|\,\bar{a}\,|\,,\,|\,\bar{b}\,|\,,\,|\,\bar{d}\,|\,)\,. \tag{5.15}$$

Now,

$$\gamma \leq \epsilon\,\Gamma\,, \tag{5.16}$$

i.e., one of the elements of $\bar{A}$ is numerically insignificant. This situation requires modifications to our algorithm, since the proposed formulas may break down. In particular, we do not solve a quadratic equation to determine either $\bar{t}_1$ or $\bar{t}_3$. Instead, we set one of the two tangents to zero and attempt to compute all the other tangents from the recurrences. We divide the special cases into three groups: one,

$$|\,\bar{a}\,| + |\,\bar{d}\,| \neq 0 \quad \text{and} \quad |\,\bar{b}\,| \neq 0\,, \tag{5.17}$$

two,

$$|\,\bar{a}\,| + |\,\bar{d}\,| = 0 \quad \text{and} \quad |\,\bar{b}\,| \neq 0\,, \tag{5.18}$$

and three,

$$|\,\bar{b}\,| = 0\,. \tag{5.19}$$

First, assume that (5.17) holds. Hence at least one, but not all, of the following three conditions hold:

$$\gamma = \bar{b}, \quad \gamma = \bar{a} \quad \text{or} \quad \gamma = \bar{d}\,.$$

We set $\bar{t}_1$ to zero if

$$|\bar{a}| > |\bar{d}| , \tag{5.20}$$

and set $\bar{t}_3$ to zero if

$$|\bar{a}| \leq |\bar{d}| . \tag{5.21}$$

Thus, the sizes of the diagonal elements of $\bar{A}$ will be compared to decide which one of $\bar{t}_1$ or $\bar{t}_3$ should be zeroed. Without loss of generality, assume that (5.20) holds; hence, $\bar{t}_1$ becomes the reference angle. So, $\bar{t}_2$ and $\bar{t}_3$ are computed from recurrence (4.8a) and (4.8b). Further, since $\bar{t}_1 = 0$ it follows that $\bar{t}_3 = -\bar{b}/\bar{a}$. Substituting these values into (5.6) and (5.7), we can verify that Theorem 5.2 holds. Similarly, Theorem 5.3 follows from (5.11). We note that it is very important to decide which reference angle to choose, even for the case when $\bar{b}$ is numerically zero. At first, the choice of the reference angle may seem arbitrary for a "small" $\bar{b}$, since either $\bar{t}_1$ or $\bar{t}_3$ can be set to zero. However, an unnecessarily large error may occur unless we pay special care.

Second, assume that (5.18) holds. Then, at least one of the a_i's equals zero and at least one of the d_j's also equals zero, for $i, j = 1, 2$. A solution is to permute either the rows or the columns, in order to ensure that the transformed product is diagonal and that the data are reordered. Hence for this case, we may set the two extreme tangents $\{\bar{t}_1, \bar{t}_3\}$ to $\{0, \infty\}$, resulting in the transformations being rotations of negative ninety and zero degrees, respectively. To be specific, consider the case where one or more a_i's equal zero. If $a_1 = 0$, set $\bar{t}_1 = 0$ and $\bar{t}_2 = \bar{t}_3 = \infty$. If $a_1 \neq 0$ and $a_2 = 0$, set $\bar{t}_1 = 0$, compute $\bar{t}_2$ from the forward recurrence, and set $\bar{t}_3 = \infty$. Note that we may also choose to determine the tangents using the values of the d_j's.

Third, assume that (5.19) holds. We need to account for the fact that we are really solving an $n \times n$ problem. Although the 2×2 subproblem is already numerically diagonal, it is not sufficient to set $\bar{t}_1 = \bar{t}_3 = \infty$, which will leave the 2×2 product unchanged. The $n \times n$ data need to be reordered, calling for $\bar{t}_1 = \bar{t}_3 = 0$, i.e., the affected rows and columns will be permuted. Unfortunately, while applying the symmetric permutation, the triangular structures of both $\bar{A}_1$ and $\bar{A}_2$ are destroyed. Therefore, $\bar{t}_2$ is determined from the recurrence.

6. Concluding Remark

In this paper we have presented a simple and accurate way to calculate the PSVD or GSVD of two 2×2 upper triangulat matrices. In Appendix C we present an example which shows that our half-recursive method produces identical numerical results as the method in [1].

7. Acknowledgements

G. E. Adams and F. T. Luk were supported in part by the Army Research Office under grant DAAL03-90-G-0104, and A. W. Bojanczyk also by the Army Research Office under grant DAAL03-90-G-0092.

8. References

[1] Z. Bai and J.W. Demmel, "Computing the Generalized Singular Value Decomposition", Report No UCB/CSD 91/645, Computer Science Division, University of California, Berkeley, August 1991.

[2] A.W. Bojanczyk, L.M. Ewerbring, F.T. Luk and P. Van Dooren, "An Accurate Product SVD Algorithm", *Signal Processing*, 25 (1991), to appear.

[3] J. P. Charlier, M. Vanbegin and P. Van Dooren, "On efficient implementations of Kogbetliantz's algorithm for computing the singular value decomposition," *Numer. Math.*, 52 (1988), pp. 279–300.

[4] K. V. Fernando and S. J. Hammarling, "A product induced singular value decomposition for two matrices and balanced realisation," in *Linear Algebra in Signals, Systems and Control*, B. N. Datta et al., Eds., SIAM, Philadelphia, Penn., 1988, pp. 128–140.

[5] M. T. Heath, A. J. Laub, C. C. Paige, and R. C. Ward, "Computing the SVD of a product of two matrices," *SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 1147–1159.

[6] C. C. Paige, "Computing the generalized singular value decomposition," *SIAM J. Sci. Statist. Comput.*, 7 (1986), pp. 1126–1146.

A Proof of Theorem 5.2

We first present a lemma.

Lemma A.1. Let $\bar{\sigma}_1$ and $\bar{t}_1$ be the exact values corresponding to the given data $\bar{a}$, $\bar{b}$ and $\bar{d}$, and let $\bar{t}_1$ be the computed value of $\bar{t}_1$. Define a residual r_1 by

$$r_1 := \frac{\bar{b}\bar{d}}{\bar{a}}(\bar{t}_1^2 + 2\bar{\sigma}_1\bar{t}_1 - 1)\,. \tag{A.1}$$

Then

$$|\, r_1 \,| \;\leq\; K_4\epsilon|\bar{b}|\,, \tag{A.2}$$

where K_4 is a positive constant.

Proof. See the proof of Lemma 5.2 in [2]. □

We now have the necessary tools for proving the theorem.

Proof (of Theorem 5.2). First, from Lemma 5.2 and relation (5.4b) we get

$$\tilde{e}' = \tilde{c}_1\tilde{c}_3[(-\bar{a}\bar{t}_3 + \bar{d}\bar{t}_1 - \bar{b}) + (\bar{a}\bar{t}_3 - \tilde{d}\bar{t}_1 + \bar{b})] = (\bar{a} - \bar{a})\tilde{c}_1\tilde{s}_3 - (\tilde{d} - \bar{d})\bar{s}_1\tilde{c}_3\,.$$

Using (5.1a)-(5.1b) and (5.4d) we prove the inequality:

$$|\, \tilde{e}' \,| \leq K\epsilon\,(|\, a \,| \;+ |\, d \,|\,) \leq K_1\epsilon \;\|\, \bar{A} \,\| \quad . \tag{A.3}$$

Second, rewrite (A.1) as

$$r_1 = \frac{1}{\bar{a}}[\bar{d}\bar{b}\bar{t}_1^2 + \bar{t}_1(\bar{d}^2 - \bar{a}^2 - \bar{b}^2) - \bar{d}\bar{b}] = \frac{1}{\bar{a}}[(\bar{d}\bar{t}_1 - \bar{b})(\bar{b}\bar{t}_1 + \bar{d}) - \bar{t}_1\bar{a}^2]\,. \tag{A.4}$$

From (5.6) we obtain

$$\frac{1}{\bar{a}}(\bar{d}\bar{t}_1 - \bar{b}) = \bar{t}_3 + \frac{\tilde{e}'}{\tilde{c}_1\tilde{c}_3\bar{a}}\,. \tag{A.5}$$

Substituting (A.5) into (A.4) and rearranging terms, we get

$$-\bar{a}\bar{t}_1 + \bar{d}\bar{t}_3 + \bar{b}\bar{t}_1\bar{t}_3 = r_1 - \frac{\tilde{e}'(\bar{b}\bar{t}_1 + \bar{d})}{\tilde{c}_1\tilde{c}_3\bar{a}}\,,$$

and so

$$\tilde{b}' = \tilde{c}_1\tilde{c}_3 r_1 - \frac{\tilde{e}'(\bar{b}\bar{t}_1 + \bar{d})}{\bar{a}}\,. \tag{A.6}$$

From (4.6d) we derive

$$|\, \bar{t}_1\bar{\sigma}_1 \,| \;\leq\; \frac{1}{2}\,,$$

and from (4.6b) we get

$$|\, \bar{\sigma}_1 \,| \;= |\, \frac{\bar{r} - \bar{b}}{2\bar{d}} \,| \;\geq |\, \frac{\bar{b}}{2\bar{d}} \,|\quad .$$

It follows that

$$| \tilde{t}_1 | \leq | \frac{\bar{d}}{\bar{b}} | < | \frac{\bar{a}}{\bar{b}} | , \tag{A.7}$$

since we have assumed that $| \bar{d} | < | \bar{a} |$. Finally, recall from (5.3) that $\bar{t}_1 = \tilde{t}_1(1 + 10\epsilon_5)$, and use (A.6), Lemma A.1 and (A.5) to obtain

$$| \tilde{b}' | \leq \tilde{c}_1 \tilde{c}_2 | r_1 | + 2 | \tilde{e}' | \leq K_2 \epsilon \| \bar{A} \| , \tag{A.8}$$

thus completing the proof. □

B How to Compute the Middle Transformation

As pointed out by Bai and Demmel in [1], a critical issue concerns how the middle transformation should be computed. They proposed the following scheme for its computation after both end transformations have been determined. In order to relate the test for computing Q_2 in [1] to the test in the half recursive method, we first translate our setting to that in [1]. Let

$$U^T \equiv \begin{pmatrix} c_1 & -s_1 \\ s_1 & c_1 \end{pmatrix} , \quad Q^T \equiv \begin{pmatrix} c_2 & -s_2 \\ s_2 & c_2 \end{pmatrix} \quad \text{and} \quad V^T \equiv \begin{pmatrix} c_3 & -s_3 \\ s_3 & c_3 \end{pmatrix} .$$

Note that the relation, given by

$$Q_1 A_1 = \begin{pmatrix} s_1 & c_1 \\ -c_1 & s_1 \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ 0 & d_1 \end{pmatrix} = \begin{pmatrix} s_1 a_1 & s_1 b_1 + c_1 d_1 \\ -c_1 a_1 & -c_1 b_1 + s_1 d_1 \end{pmatrix} \tag{B.1a}$$

upon permuting rows and changing the signs of the top row, is equivalent to

$$U^T A_1 = \begin{pmatrix} c_1 & -s_1 \\ s_1 & c_1 \end{pmatrix} \begin{pmatrix} a_1 & b_1 \\ 0 & d_1 \end{pmatrix} = \begin{pmatrix} c_1 a_1 & c_1 b_1 - s_1 d_1 \\ s_1 a_1 & s_1 b_1 + c_1 d_1 \end{pmatrix} \equiv G . \tag{B.1b}$$

Similarly,

$$A_2 Q_3^T = \begin{pmatrix} a_2 & b_2 \\ 0 & d_2 \end{pmatrix} \begin{pmatrix} s_3 & -c_3 \\ c_3 & s_3 \end{pmatrix} = \begin{pmatrix} s_3 a_2 + c_3 b_2 & -c_3 a_2 + s_3 b_2 \\ c_3 d_2 & s_3 d_2 \end{pmatrix} . \tag{B.2a}$$

By changing the sign of the second columns, and permuting columns we obtain

$$V^T adjoint(A_2) = \begin{pmatrix} c_3 & -s_3 \\ s_3 & c_3 \end{pmatrix} \begin{pmatrix} d_2 & -b_2 \\ 0 & a_2 \end{pmatrix} = \begin{pmatrix} c_3 d_2 & -c_3 b_2 - s_3 a_2 \\ s_3 d_2 & -s_3 b_2 + c_3 a_2 \end{pmatrix} \equiv H . \tag{B.2b}$$

In [1] Bai and Demmel used (B.1b) and (B.2b) as a starting point for computing Q_2. Their argument is as follows. After postmultiplications of both (B.1b) and (B.2b) by Q_2, the (1,2) elements of G and H should become zeros. Now, one should compute Q_2 from the one product, either G or H, for which the computed element in the (1,2) position has a smaller error relative to the norm of the row in which it resides. The magnitude of that error can be only bounded and hence the test for the choice is based on the bounds of the errors. It is easy to see that the bound g for the relative error in the (1,2) element of the computed G is

$$g = \frac{|c_1 b_1| + |s_1 d_1|}{|c_1 a_1| + |c_1 b_1 - s_1 d_1|} , \tag{B.3a}$$

while the bound h for the relative error in the (1,2) element of the computed H is

$$h = \frac{|c_3 b_2| + |s_3 a_2|}{|c_3 d_2| + |c_3 b_2 + s_3 a_2|} . \tag{B.3b}$$

Now, if $g \leq h$ then Bai and Demmel compute Q_2 from $U^T A$ and otherwise from $V^T B$. The next lemma states that the conditions that specify how Q_2 is computed in [1] and by the half-recursive method are essentially equivalent.

Lemma B.1. In exact arithmetic the condition

$$g \leq h \tag{B.4a}$$

where g is defined by (B.3a) and h is defined by (B.3b), is equivalent to the condition

$$a \geq d \tag{B.4b}$$

Proof. First note that (B.3a) and (B.3b) can be simplified to

$$g = \frac{|b_1| + |t_1 d_1|}{|a_1| + |t_1 d_1 - b_1|} \tag{B.5a}$$

and

$$h = \frac{|b_2| + |t_3 a_2|}{|d_2| + |t_3 a_2 + b_2|} , \tag{B.5b}$$

respectively. By using (4.8a) and (4.8c) the relations (B.3a) and (B.3b) simplify further to

$$g = \frac{|b_1| + |t_1 d_1|}{|a_1|(1 + |t_2|)} \tag{B.6a}$$

and

$$h = \frac{|b_2| + |t_3 a_2|}{|d_2|(1 + |t_2|)} , \tag{B.6b}$$

respectively. Hence (B.4a) is equivalent to

$$|b_1 d_2| + |t_1 d| \leq |a_1 b_2| + |a t_3| . \tag{B.7}$$

We now prove that (B.4b) implies (B.4a). The proof that $d < a$ implies that $h < g$ is analogous and is omitted. The proof is elementary but tedious as it requires us to consider a large number of cases. So we assume that $a \geq b$. Then Lemma 3.1 implies that $t_3 \geq t_1$. From (4.8b) we see that

$$|a t_3 + b| = |d t_1| ,$$

and as $|a t_3| \geq |a t_1|$ we conclude that

$$sign(a t_3) = -sign(b) = -sign(a_1 b_2 + b_1 d_2) , \tag{B.8}$$

as from (4.7b) $b = a_1 b_2 + b_1 d_2$. Substituting (4.8b) into (B.7) and using (4.7b) again we get that (B.7) is equivalent to the following inequality:

$$|b_1 d_2| + |a t_3 + a_1 b_2 + b_1 d_2| \leq |a_1 b_2| + |a t_3| . \tag{B.9}$$

Case 1. $-|b| \geq |b_1 d_2| - |a_1 b_2|$.

Then

$$|at_3| \geq |dt_1| - |b| \geq |dt_1| + |b_1 d_2| - |a_1 b_2| ,$$

establishing (B.7).

Case 2a. $-|b| > |b_1 d_2| - |a_1 b_2|$ and $|at_3| > |b|$.

Then $|a_1 b_2| > |b_1 d_2|$ and using (B.8) we obtain that

$$|b_1 d_2| + |dt_1| = |b_1 d_2| + |at_3 + a_1 b_2 + b_1 d_2| = |at_3| + 2\,|b_1 d_2| - |a_1 b_2| ,$$

from which (B.7) follows.

Case 2b. $-|b| > |b_1 d_2| - |a_1 b_2|$ and $|at_3| \leq |b|$.

Then again $|a_1 b_2| > |b_1 d_2|$. Now from (B.8)

$$|b_1 d_2| + |dt_1| = |b_1 d_2| + |at_3 + a_1 b_2 + b_1 d_2|$$

$$= |b_1 d_2| - |at_3| + |a_1 b_2| - |b_1 d_2| = |a_1 b_2| - |at_3| ,$$

from which (B.7) again follows.

Remark. Note that there might be a slight difference in using (B.4a) or (B.4b) as the lemma holds only in exact arithmetic. In finite precision computation, the relations (B.4a) and (B.4b) may not always be equivalent. However, we have not been able to find any numerical example where these two conditions are not equivalent. Moreover, as shown in this paper the consequences of numerical non-equivalence are numericaly insignificant.

C Numerical Example

It has been proved in Appendix B that the half-recursive procedure computes essentially the same numerical results as the *direct* method of [1]. For both methods, the end transformations are computed explicitly from the product $A = A_1 A_2$, and the middle transformation is computed from the same direction. The greatest difference between the fully-recursive method and the other two occurs when there is cancellation in forming the product $A = A_1 A_2$. In the following PSVD example, A_1 and A_2 each has an $O(1)$ norm, but the product $A_1 A_2$ has an $O(10^{-5})$ norm. Hence errors which are small relative to the initial matrices may be large relative to the product.

$$A_1 = \begin{pmatrix} 2.316797292247488e+00 & -1.437687878748196e-01 \\ 0 & -5.208536329107726e-06 \end{pmatrix} ,$$

$$A_2 = \begin{pmatrix} 2.472499811756353e-05 & 2.624474233535929e-01 \\ 0 & 4.229273187671001e+00 \end{pmatrix} ,$$

$$A_1 A_2 = \begin{pmatrix} 5.728280868959543e-05 & -1.110223024625157e-16 \\ 0 & -2.202832304370565e-05 \end{pmatrix} .$$

The three methods all compute the left transformation from the explicit product, and calculate the middle transformation from A_1. We use the subscipts dir, hr, and fr to distinguish between results computed via the direct, half-recursive, and fully-recursive methods, respectively. The computed values of $A'_{1,dir}$, $A'_{1,hr}$, and $A'_{1,fr}$ are *numerically identical* in that the corresponding entries are numerically equal:

$$\bar{A}'_{1,dir} = \begin{pmatrix} 2.321253790030786e+00 & 2.775557561562891e-17 \\ 3.225930076892087e-07 & -5.198536633811768e-06 \end{pmatrix},$$

$$\bar{A}'_{1,hr} = \begin{pmatrix} -5.198536633811768e-06 & -3.225930076892087e-07 \\ -2.775557561562891e-17 & 2.321253790030786e+00 \end{pmatrix},$$

$$\bar{A}'_{1,fr} = \begin{pmatrix} -5.198536633811768e-06 & -3.225930076892087e-07 \\ -2.775557561562891e-17 & 2.321253790030786e+00 \end{pmatrix}.$$

The computed values of $A'_{1,dir}$, $A'_{1,hr}$, and $A'_{1,fr}$ are numerically triangular but now the (1,2) element in $\bar{A}'_{1,fr}$ is significantly different than the corresponding element in $\bar{A}'_{1,dir}$ or $\bar{A}'_{1,hr}$:

$$\bar{A}'_{2,dir} = \begin{pmatrix} 2.467752941777026e-05 & 5.551115123125783e-17 \\ 1.531353724707768e-06 & 4.237408446913959e+00 \end{pmatrix},$$

$$\bar{A}'_{2,hr} = \begin{pmatrix} 4.237408446913959e+00 & -1.531353724707768e-06 \\ -5.551115123125783e-17 & 2.467752941777026e-05 \end{pmatrix},$$

$$\bar{A}'_{2,fr} = \begin{pmatrix} 4.237408446913959e+00 & -1.531363362694676e-06 \\ 0 & 2.467752941777026e-05 \end{pmatrix}.$$

To maintain triangularity, $\bar{A}'_1$ and $\bar{A}'_2$ are truncated by setting the appropriate elements to zero. Let A''_1 and A''_2 denote the truncated matrices. The product $\bar{A}'' = A''_1 \cdot A''_2$ should be diagonal:

$$\bar{A}''_{dir} = \begin{pmatrix} 5.728280868959542e-05 & 0 \\ 1.615587133892632e-27 & -2.202832304370564e-05 \end{pmatrix},$$

$$\bar{A}''_{hr} = \begin{pmatrix} -2.202832304370564e-05 & -1.615587133892632e-27 \\ 0 & 5.728280868959542e-05 \end{pmatrix},$$

$$\bar{A}''_{fr} = \begin{pmatrix} -2.202832304370564e-05 & 5.010342801562901e-17 \\ 0 & 5.728280868959542e-05 \end{pmatrix}.$$

Clearly, $\bar{A}''_{hr}$ and $\bar{A}''_{dir}$ are numerically diagonal, but $\bar{A}''_{fr}$ fails the criterion of diagonality. Forcing $\bar{A}''_{fr}$ to be a diagonal matrix requires a truncation of $O(10^{-17})$, which is significant with respect to $\|\bar{A}''\|$. The matrices $\bar{A}''_{dir}$ and $\bar{A}''_{hr}$ require only insignificant truncations to obtain diagonality, but we have previously made $O(10^{-17})$ truncations during their computation to force $\bar{A}''_{2,dir}$ and $\bar{A}''_{2,hr}$ to triangular forms. Thus, equal amounts of absolute truncation errors have been committed by all three methods; the only difference is that the relative truncation error is largest for the fully-recursive method.

It is interesting to note that if triangularity is not enforced and the factors $\bar{A}'_1$ and $\bar{A}'_2$ are multiplied, then none of the products can be considered diagonal. One may say that the numerical diagonality of $\bar{A}''_{hr}$ and $\bar{A}''_{dir}$ is a consequence of the truncation to triangular forms.

$$\bar{A}'_{1,dir} \cdot \bar{A}'_{2,dir} = \begin{pmatrix} 5.728280868959542e-05 & 2.464671807471544e-16 \\ 1.615587133892632e-27 & -2.202832304370564e-05 \end{pmatrix},$$

$$\bar{A}'_{1,hr} \cdot \bar{A}'_{2,hr} = \begin{pmatrix} -2.202832304370564e-05 & -1.615587133892632e-27 \\ -2.464671807471544e-16 & 5.728280868959542e-05 \end{pmatrix},$$

$$\bar{A}'_{1,fr} \cdot \bar{A}'_{2,fr} = \begin{pmatrix} -2.202832304370564e-05 & 5.010342801562901e-17 \\ -1.176117105626251e-16 & 5.728280868959542e-05 \end{pmatrix}.$$

In conclusion, our example shows that the half-recursive and direct methods produce numerically identical results, while the fully-recursive method fails to meet the diagonality criterion.

An Asynchronous Array Design for MVDR Beamformers

Moon S. Jun
Physical Science Laboratory
New Mexico State University
Las Cruces, New Mexico 88003

Shietung Peng
Computer Science Department
University of Maryland, MD 21228

Abstract

In this paper[1], we present an asynchronous array design for the minimum variance distortionless response (MVDR) beamformers. The array transforms the constrained problem into unconstrained form, enabling an unconstrained processor to compute the beamformer output. The key component of the array is a communication protocol which controls input data flow properly and efficiently. In the design, instead of using global control, self-timed processing elements (PEs) and communication protocols are provided. The asynchronous array for MVDR beamformers can significantly speed up the total computation time. Finally, we present an algorithm in Occam[2] languages for the asynchronization scheme of the processes. It is felt that the array has promise for real-time beamforming with planar array antennas.

1 Introduction

Due to advances in VLSI technology, there is much interest in using array processors to improve the throughput rate of various signal processing algorithms. The use of systolic arrays for adaptive beamforming technology has been proposed and developed by several authors [1, 2, 3, 6–9]. In these works, the adaptive beamforming has

[1]This work is partially supported by the Army of Research Office(ARO) under contracts DAAL03-90-G-0211.

[2]Occam are trademarks of the INMOS Group of Companies

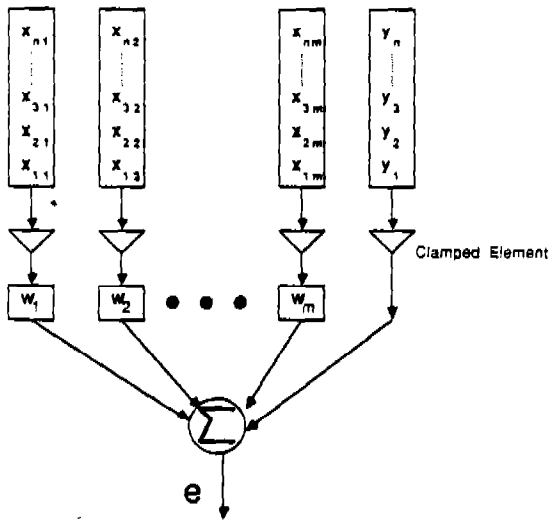


Figure 1: The functional diagram of a MVDR beamformer.

been formulated as a least-squares problem and implemented using triangular systolic array by means of the QR algorithm. In this paper, we propose an asynchronous array which can perform the QR decomposition needed in the solution of the MVDR beamformers.

In general, a minimum variance distortionless response (MVDR) beamformer has $(m+1)$ sensor elements and a beam-pattern forming network comprising (m) weights that have to be determined in order to maximize the array response to the desired signals. The objective of an optimal beamforming system is to minimize the total error power via manipulation of the weight values subject to the clamped weight constraint. The $(m+1)$-th (reference) sensor element is constrainted to a constant value $\mu(j)$ $(0 < \mu(j) \leq 1)$. The functional diagram of a MVDR beamformer is shown in Figure 1. At each sample time t_i, evaulate the *a posteriori* residuals

$$
\begin{aligned}
e(t_i) &= \begin{bmatrix} x(t_1,1) & x(t_1,2) & \cdots & x(t_1,m) \\ x(t_2,1) & x(t_2,2) & \cdots & x(t_2,m) \\ \vdots & \vdots & \vdots & \vdots \\ x(t_n,1) & x(t_n,2) & \cdots & x(t_n,m) \end{bmatrix} \cdot \begin{bmatrix} w(1) \\ w(2) \\ \vdots \\ w(m) \end{bmatrix} + \begin{bmatrix} y(t_1) \\ y(t_2) \\ \vdots \\ y(t_n) \end{bmatrix} \\
&= \begin{bmatrix} e(t_1) \\ e(t_2) \\ \vdots \\ e(t_n) \end{bmatrix}
\end{aligned} \tag{1}
$$

where $x(t_i, j)$ is the j-th element vector of (complex) signal samples received by the array at time t_i, $y(t_i)$ is the value at time t_i of an additional reference signal, and $w(j)$

is the j-th element vector of (unconstrainted) weights which minimizes the equantity for $1 \leq j \leq n$

$$minimize\left\{\ e(t_i)\ \right\} = \| \sum_{j=1}^{m} \{x(t_i, j) \cdot w(j)\} + y(t_i) \|, \tag{2}$$

subject to a linear equatity constraint of the form

$$\mu(0 < \mu(j) \leq 1) = \begin{bmatrix} \mu(1) \\ \mu(2) \\ \vdots \\ \mu(m) \end{bmatrix} = \begin{bmatrix} c(t_i, 1) \cdot w(1) \\ c(t_i, 2) \cdot w(2) \\ \vdots \\ c(t_i, m) \cdot w(m) \end{bmatrix}, \quad 0 \leq i \leq n. \tag{3}$$

The symbol $\| \cdot \| = \| \cdot \|_2$ denotes the euclidean norm.

The key components of an adaptive antenna system are illustrated in Figure 1 and Equation 1. The amplitude and phase weights are selected by a beampattern controller that continuously updates them in response to the element outputs. In this paper, we describe an asynchronous algorithm and architectures for high performance, digital, adaptive beamforming.

Section 2 describe an efficient linear equation using the Givens rotation [3, 5] and the QR decomposition algorithm [1, 2, 7–9]. Most previous array [1, 2, 6–9] may be designed more delay time and more complicated implementations. To solve these problems, the top boundary PEs receive both $x(t_i, m)$ and $c(t_i, m)$ and compute $a(t_i, m)$ from two data. The *a posteri* residual equation is solved by using the Givens rotations [3, 5]. The dervied equation is more understandable and can get higher performances. To achieve maximal parallelism of constrained MVDR beamformers, section 3 shows data dependencies in computations and systolic recurrence equations.

Section 4 discusses an asynchronous design and its time analysis for MVDR. In an asynchronous design, self-timed PEs and communication protocols are provided. With the Occam programs, it will be shown that the triangular systolic beamformer can simultaneously and concurrently process the input data from the rows and columns of the array antennas with a speed comparable to McWhirter's systolic array for MVDR beamformer.

2 A Constrainted MVDR Beamformer

The MVDR beamformer problem can be summarized from Equation 1. Given a data matrix X and a desired vector Y, find the tap weight vector W which minimizes the least-squares error

$$\|e(t_i)\| = \|\sum_{j=1}^{m}\{x(t_i,j)\cdot w(j)\} + y(t_i)\|, \tag{4}$$

where $x(t_i,j) \in X$, $w(j) \in W$, and $y(t_i) \in Y$. McWhirter and Shepherd [6–8] have developed an algorithm that directly extracts the residuals element $e(t_i)$, without using the weight vector $w(j)$, by QR decomposition which consists of a sequence of unitary transformations applied to the measured signal matrix $x(t_i,j)$ to transform it to a triangular matrix. Assuming that a QR decomposition [6–8] has been carried out on the data matrix $x(t_i,j)$ so that

$$\|Q(j,t_i)\cdot e(t_i)\| = \|Q(j,t_i)\cdot\sum_{j=1}^{m}\{x(t_i,j)\cdot w(j)\} + Q(j,t_i)\cdot y(t_i)\|, \tag{5}$$

where $Q(j,t_i)$ represents a sequence of elementary Givens rotations used to annihilate each element of a new data vector $x(t_i,j)$. Then the equation 5 can be expressed in the form

$$\|E(j)\| = \|Q(j,t_i)\cdot\sum_{j=1}^{m}\{x(t_i,j)\cdot w(j)\} + Y(j)\|, \tag{6}$$

where both $E(j)$ and $Y(j)$ are a $(m\times 1)$ matrix, respectively. The weight vector $w(j)$ determines the characteristics of the beamformer. For a MVDR beamformer, $w(j)$ can be chosen to minimize the output power from equation 3. The solution to this constrainted least squares minimization problem can be given by the the following formulation

$$\left[\; w(j) \;\right]_{m\times 1} = \{w(j)|w(j) = \frac{\mu(j)}{c(t_i,j)}, 1\le i\le n, 1\le j\le m\} \tag{7}$$

It follows that the *a posteriori* residual at *j-th* (reference) sensor element is given by

$$\left[\; E(j) \;\right]_{m\times 1} = \left[\; Q(j,t_i) \;\right]_{m\times n}\left[\; \frac{x(t_i,i)}{c(t_i,j)} \;\right]_{n\times m}\left[\; \mu(j) \;\right]_{m\times 1} + \left[\; Y(j) \;\right]_{m\times 1.} \tag{8}$$

Similarly [McWhirter 89], the Givens rotations $Q(j, t_i)$ can be used to annihilate each element of a new data vector

$$a(t_i, j) = \frac{x(t_i, j)}{c(t_i, j).} \tag{9}$$

The inner product of $Q(j, t_i)$ and $a(t_i, j)$ is $Z(j, k)$ with a $(m \times 1)$ upper triangular matrix. It is simply given by

$$E(j) = \gamma \sum_{k=j}^{m} Z(j, k) \cdot \mu(j) + Y(j), \tag{10}$$

where $1 \leq j \leq m$ and γ is an coefficient rate of residuals.

To efficiently implement Equation 10, we can use a triangular systolic array which can be carried out using the Givens rotations. The Givens rotation method has been found to be particularly suitable for adaptive application since the triangularization process is recursively updated as each new row of data enters the computation. In the next section, we describe the Givens rotation and consider the systolic array design.

3 A Systolic Design for Constrainted MVDR

To achieve maximal parallelism of the MVDR algorithm, we must try to find data dependencies in a computations. The following recurrences are defined over the index space:

$$1 \leq k, j \leq m$$

where m is the number of processing elements in row and column. Assume that $A_{k,j}(t_i)$, $R_{k,j}(t_i)$, $sin_{k,j}(t_i)$, and $cos_{k,j}(t_i)$ are computed in the processing element with below index (k, j) at time t_i. In the equation of MVDR, the dependency of $R_{k,j}(t_i)$ is local, while the dependencies of $A_{k,j}(t_i)$, $cos_{k,j}(t_i)$, and $sin_{k,j}(t_i)$ are global. There are two kinds of processes as shown in Figure 2.

1. Givens Rotation for $PE_{j,j}$: Given ($1 \leq j \leq m$)

$$\begin{bmatrix} R_{j,j}(t_i) \\ A_{j,j}(t_i) \end{bmatrix} \in \Re^2,$$

where $\Re$ is a real number, compute parameters $sin_{j,j}$ and $cos_{j,j}$ such that

$$sin^2_{j,j} + cos^2_{j,j} = 1$$

and

$$\begin{bmatrix} R_{j,j}(t_{i+1}) \\ A_{j,j}(t_{i+1}) \end{bmatrix} = \begin{bmatrix} cos_{j,j} & sin_{j,j} \\ -sin_{j,j} & cos_{j,j} \end{bmatrix} \cdot \begin{bmatrix} R_{j,j}(t_i) \\ A_{j,j}(t_i) \end{bmatrix} = \begin{bmatrix} \sqrt{\|R_{j,j}(t_i)\|^2 + \|A_{j,j}(t_i)\|^2} \\ 0 \end{bmatrix},$$

where two funchtions, f_{sin} and f_{cos}, are

$$sin_{k,j} = f_{sin}\left(\begin{bmatrix} R_{j,j}(t_i) \\ A_{j,j}(t_i) \end{bmatrix}\right) := \frac{A_{j,j}(t_i)}{R_{j,j}(t_{i+1})} \tag{11}$$

$$cos_{j,j} = f_{cos}\left(\begin{bmatrix} R_{j,j}(t_i) \\ A_{j,j}(t_i) \end{bmatrix}\right) := \frac{R_{j,j}(t_i)}{R_{j,j}(t_{i+1})} \tag{12}$$

$$Q_{j,j} = \begin{bmatrix} cos_{j,j} & sin_{j,j} \\ -sin_{j,j} & cos_{j,j} \end{bmatrix}. \tag{13}$$

That is, the plane rotation $Q_{j,j}$ is determined in terms of the elements $R_{j,j}(t_i)$ and $A_{j,j}(t_i)$ to annihilate $A_{j,j}(t_{i+1})$. The parameters of this process may be described functions as follows.

$$R_{j,j}(t_{i+1}) = \sqrt{\|R_{k,j}(t_i)\|^2 + \|A_{j,j}(t_i)\|^2}$$

and

$$A_{j,j}(t_{i+1}) = 0.$$

2. Apply Rotation for $PE_{k,j}$, k < j :

$$\begin{bmatrix} R_{k,j}(t_{i+1}) \\ A_{k,j}(t_{i+1}) \end{bmatrix} = \begin{bmatrix} cos_{j,j} & sin_{j,j} \\ -sin_{j,j} & cos_{j,j} \end{bmatrix} \cdot \begin{bmatrix} R_{k,j}(t_i) \\ A_{k,j}(t_i) \end{bmatrix}. \tag{14}$$

Then $R_{k,j}(t_{i+1})$ and $A_{k,j}(t_{i+1})$ are:

$$R_{k,j}(t_{i+1}) = cos_{j,j} \cdot R_{k,j}(t_i) + sin_{j,j} \cdot A_{k,j}(t_i)$$

and

$$A_{k,j}(t_{i+1}) = -sin_{j,j} \cdot R_{k,j}(t_i) + cos_{j,j} \cdot A_{k,j}(t_i)$$

Once computed, a rotation is applied successively to each column of the affected pair of rows. Since each process applied to a pair of elements in adjacent rows, it can be identified by the indexes of the top element involved. This identifier is referred to as the *process index. A process dependence* graph is a graph whose set of nodes is a set

of processes to be executed, and whose arcs represent an ordering relation between these processes. A processes dependence graph for computing

$$\begin{bmatrix} R_{j,j}(t_{i+1}) & \cdots & R_{j,m}(t_{i+1}) \\ \\ 0 & \cdots & A_{j,m}(t_{i+1}) \end{bmatrix} = \begin{bmatrix} \\ Q_{j,j} \\ \\ \end{bmatrix} \cdot \begin{bmatrix} R_{j,j}(t_i) & \cdots & R_{j,m}(t_i) \\ \\ A_{j,j}(t_i) & \cdots & A_{j,m}(t_i) \end{bmatrix} \tag{15}$$

is given in Figure 2. In the figure, processes are identified by their process index. Each rotation application is indicated in Figure 2 by a rectangular vertex. The parameters followed by their application. This is not the only process dependence graph that is compatible with a QR factorization based on Givens rotations. Its cellular structure, however, makes it amenable to realization as a systolic array.

Algorithm 1: A systolic version for MVDR

$$initial = \begin{cases} X_{1,j}(t_i) \longleftarrow x(t_i, j) \\ C_{1,j}(t_i) \longleftarrow c(t_i, j) \end{cases} \tag{16}$$

$$A_{k+1,j}(t_{j+1}) \longleftarrow \begin{cases} k \leq j: \quad cos_{k,j}(t_j) \cdot R_{k,j}(t_j) + sin_{k,j}(t_i) \cdot A_{k,j}(t_i) \\ where \quad A_{1,j}(t_i) = \begin{pmatrix} X_{1,j}(t_i) \\ - - - \\ C_{1,j}(t_i) \end{pmatrix} \\ i > j: \quad 0 \end{cases} \tag{17}$$

$$R_{k,j}(t_{i+1}) \longleftarrow \begin{cases} k < j: \quad cos_{k,j}(t_i) \cdot R_{k,j}(t_{i-1}) + sin_{k,j}(t_i) \cdot A_{k,j}(t_i) \\ k = j: \quad \sqrt{\|A_{k,j}(t_i)\|^2 + \|R_{k,j}(t_i)\|^2} \end{cases} \tag{18}$$

where

$$sin_{k,j+1}(t_{i+1}) \longleftarrow \begin{cases} k < j: \quad sin_{k,j}(t_{i-1}) \\ k = j: \quad \begin{pmatrix} A_{k,j}(t_i) \\ - - - \\ R_{k,j}(t_{i+1}) \end{pmatrix} \end{cases} \tag{19}$$

$$cos_{k,j+1}(t_{i+1}) \longleftarrow \begin{cases} k < j: \quad cos_{k,j}(t_{i-1}) \\ k = j: \quad \begin{pmatrix} R_{k,j}(t_i) \\ - - - - \\ R_{k,j}(t_{i+1}) \end{pmatrix} \end{cases} \tag{20}$$

$$\gamma_{k+1,k+1}(t_{j+1}) \longleftarrow \begin{cases} k \neq 1: \quad \gamma_{k,k}(t_i) \cdot cos_{k,k}(t_i) \\ k = 1: \quad 1 \end{cases} \tag{21}$$

Two ordering constraints must be respected as shown in Figure 2.

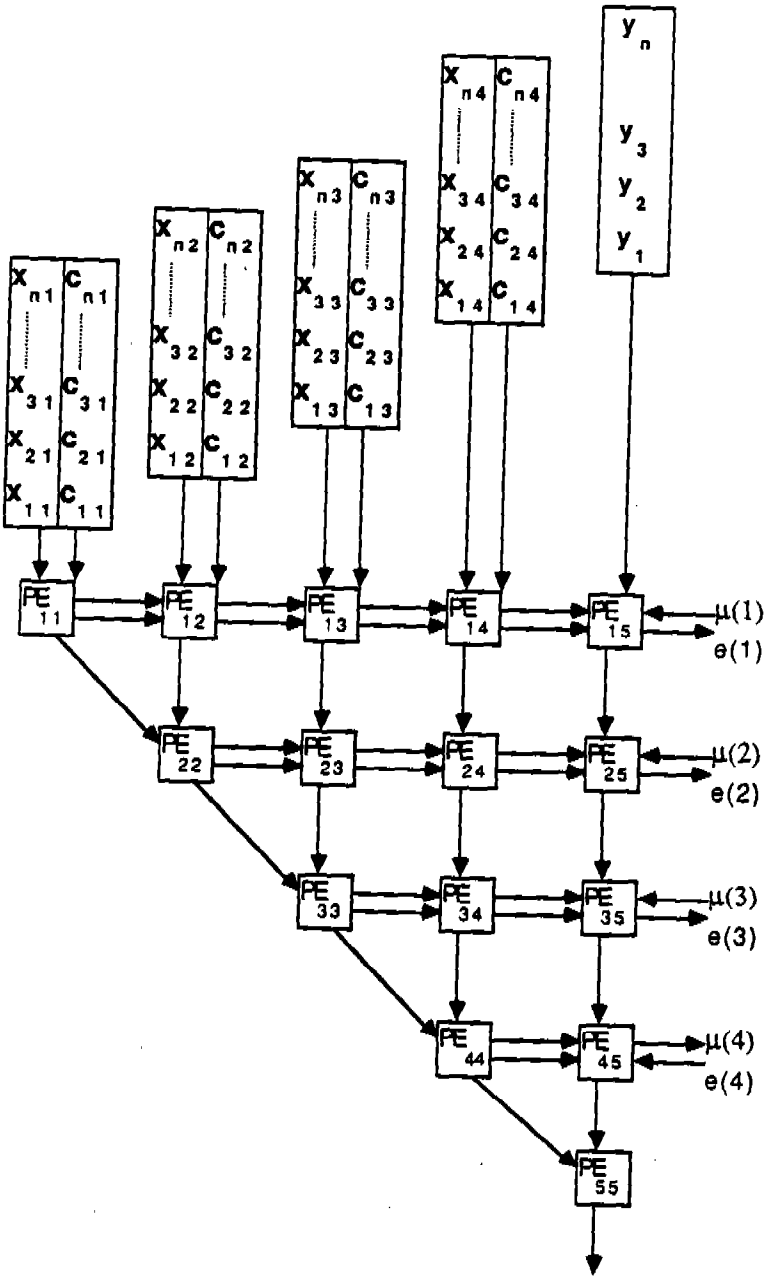


Figure 2: Systolic implementation of the MVDR beamformer.

1. The plane rotation application performed by PE(1,2) must be complete before PE(2,2) can compute its rotation parameters. This is true in general: The plane rotation application performed by PE $(k-1, j)$ must be compute before PE (k, j) can compute its rotation parameters.

2. Rotation application for PE (k, j) must be complete before PE $(k+1, j)$ can be apply its rotation parameters, since both affect PE $(k+1, j)$.

Both of these ordering constraints are represented in the process dependence graph by the horizontal arcs between processes: No data flow across the horizontal arcs, these arcs impose only a partial order on the processes.

4 An Asynchronous Design

A majority of the signal processing algorithms require a lot of the computations. In a systolic array, each PE receives the data, carries out the computations, and pumps the results rhythmically to the neighboring PEs. One problem with previous systolic arrays is the global control of data movement in different PEs. To assure proper timing and synchronization in systolic arrays, extra delays are needed. This slows down the computation, therefore decreasing throughtput rate. Moreover, for large scale arrays this synchronization could become very tedious.

To overcome these difficulties and to speed up the computation time, design of asynchronous arrays was explored. In an asynchronous design, instead of using global clock, self-timed PEs and communication protocols are provided. The advantage is that the whole period of a clock unit for multiplication, addition, and routing can be separated into several small steps and some of these steps can be executed simultanously. The concept of asynchronous computations can be specified as below steps:

1. send an acknowledge signal to previous processors while getting data from them and send a request signal to next processors while forwarding data to them.

2. transfer data to next processors.

3. execute input data and accumulate the results.

Note that step 2 and 3 can be executed simultaneously. In this section, we will develop a protocol to implement the above processes. The idea is to use self-timed PEs in which the inner product operations is triggered by the availability of the data. The major different between the two architectures is the fact that the new array transfers the data to the next cell asynchronously by its local control unit, while systolic arrays require global timing for the control of data flows. Therefore, a PE does not have to wait for data until the previous PE complete its computation. It has the basic features of the previous systolic array with the exception that the data routing and computing in each PE can be operated simultaneously.

To make the data flow independent of the operations in each PE, we need a protocol to control the flow of data such that the values of input variables will not be overwritten during their computting periods. As shown in the proposed protocol of Figure 3, three kinds of signals (R, A, ans E) are introduced: two external signals and one internal signal. The function of **a R signal** is to report to the next PE that the data in its output port is ready for transmission. The function of **an A signal** is to report the previous PE that its input port is ready to receive new data. The function of **an E signal** is to report the emptiness of the input port. The protocol can be described formally as below.

1. Each PE receives a request from previous PE when the data (*sin*, *cos*, γ, and A) in the output port of previous PE are ready to be transmitted.

2. The PE sends an acknowledge to previous PE when completely receives new data.

3. Each PE has a internal signal, E, which report the emptiness of the input port.

In Figure 3, we depict a detailed configuration of this protocol. Communication and processing in the array are usually executed in asynchronous parallel to reduce loss time in the processing elements. The loss time may yield some divergence between synchronized concurrent processes, and it decreases a efficiency of the system. An example of the algorithm can be described with Occam programs. Occam programs are built from three primitive processes:

variable := expression assign value of expression to variable
channel ? variable input a value from channel to variable
channel ! expression output the value of expression to channel

R request flag

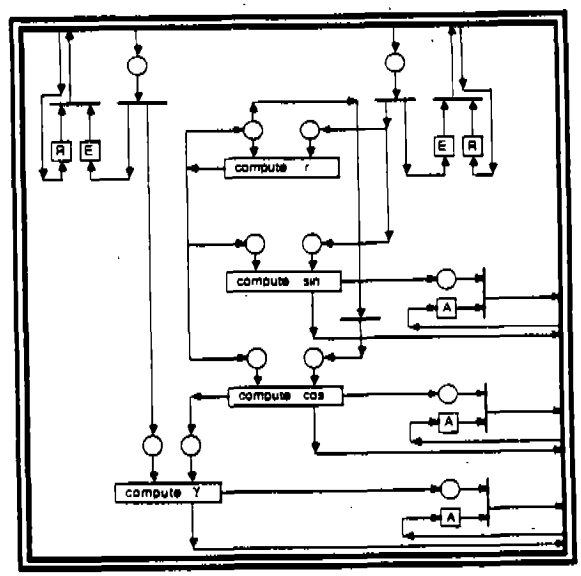
E emty-input buffer flag

A acknowledge flag

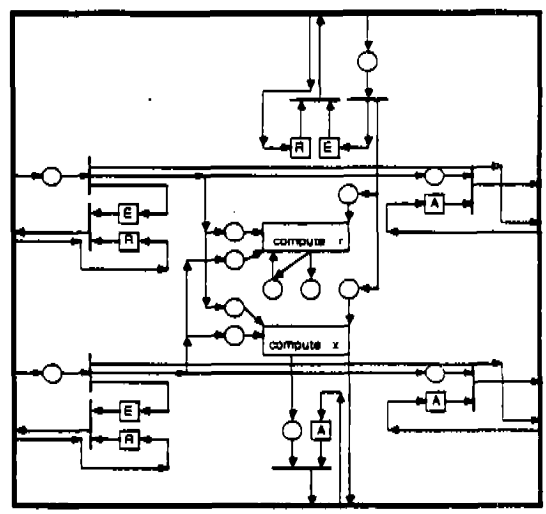
data line

data buffer

compute compute transition



(a) $PE_{j,j}$ $(1 \leq j \leq m)$



(b) $PE_{k,j}$ $(1 \leq k < j \leq m)$

Figure 3: The proposed protocols in the asynchronous array.

A sequential-construct is represented by

```
SEQ P1
    ⋮
    Pn.
```

The component processes P1, ···, and Pn are executed one after another. A parallel-construct is represented by

```
PAR P1
    ⋮
    Pn.
```

The component processes P1, ···, and Pn are executed together. The following algorithms of PE1 and PE2 show a fragment of Occam program for this asynchronization scheme of the processes in Figure 3 and Appendix.

It is easy to see that the above algorithm described correctly implements the asynchronous version for MVDR beamformers. Since the new input data are received only whenever input ports are ready to receive, indicating the completeness of internal computations, it guarantees that overwritting of input data will never occur.

5 Conclusions

In this paper, we have shown an asynchronous array processing algorithm and new PE designs for the array of MVDR beamformers. The design procedure should be applicable to other adaptive signal processings. It will be of great interest to design efficient systolic arrays and asynchronous arrays for those radar signal processings. The asynchronous array improves the performance of the systolic array further as indicated in our simulation. Some additional hardwares may be needed for implementing protocols, but a reduction of computing time is significant for large scale computations. It might be possible to improve the proposed protocol for data communication. Issues about implementation and evaulation of the asynchronous array deserve more research attention. More research can be conducted in this direction.

References

[1] Bohme, J. F., and Yang, B., "Systolic implementation of a general adaptive processing algorithm". *IEEE Int. Conf. ASSSP*, 1988, pp.2785-2788.

[2] Bojanczyk, Adam W., and Luk, F., "Reasearch Note: A unified systolic array for adaptive beamforming". *J. of Parallel and Distributed Computing 8*, 1990, pp. 388-392.

[3] Gentleman, W. M., "Least squares computations by Givens transformations without square roots," *J. Inst. Math & Appl.*, 1973, 12, p. 329-336.

[4] Griffiths, L. J., and Jim, C. W., "An alternative approach to linearly constrainted adaptive beamforming". *IEEE Trans. on Antennas and propagation*, AP-30:27-34, Jan. 1982.

[5] Golub, G. H., and Van Loan, C. F. *Matrix Computations*. Johns Hopkins University Press, Baltimore, MD. 1983.

[6] McCanny, J. V., and McWhirter, J. G.. "Some systolic array developments in the United Kingdom." *IEEE Computer* (July 1987), pp.51-63.

[7] McWhirter, J. G., "Recursive least squares minimization using a systolic array," In Bromley, K.(Ed.). *Real-Time Signal Processing VI*, pp. 105-112 (proceedings SPIE, Vol. 432, 1983).

[8] McWhirter, J. G., and Shepherd, T. J., "A systolic array for constraint least squares problems". In Speiser, J. M. (Ed.). *Adavanced Algorithms and Architectures for Signal processing I*, pp. 80-87 (Processings SPIE, Vol. 696, 1986)

[9] Veen, B. V., "Systolic preprocessors for linearly constrainted beamforming". *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol-37, no. 4, April, 1989, pp. 600-604.

(PROC proc.$PE_{j,j}$)

```
PAR
    SEQ ch1 ? req5
        :
        ch1 ? γ[ time ]
        ch1 ! ack5
    SEQ ch2 ? req9
        :
        ch2 ? A[ time ]
        ch2 ! ack9
        R[time+1] := SQRT( ABS(R[time]) ** 2 + ABS(A[time]) ** 2 )
        PAR
            SEQ sin[time + 1] = A[time] / R[time]
                ch3 ! req3
                :
                ch3 ! sin[time+1]
                ch3 ? ack3
            SEQ cos[time + 1] = R[time] / R[time]
                PAR
                    SEQ ch4 ! req4
                        :
                        ch4 ! cos[time + 1]
                        ch4 ? ack4
                    SEQ γ[time+1] := γ[time] * cos[time+1]
                        ch5 ! req1
                        :
                        ch5 ! γ[time+1]
                        ch5 ? ack1
    time := time + 1
:
```

(PROC proc.$PE_{k,j}$)

```
R[1] := 1.0(REAL32)
SEQ
    PAR
        SEQ ch3 ? req3
            :
            ch3 ? sin[ time ]
            ch3 ! ack3
            ch7 ! req3
            :
            ch7 ! sin[time + 1]
            ch7 ? ack3
        SEQ ch4 ? req4
            :
            ch4 ? cos[ time ]
            ch4 ! ack4
            ch8 ! req4
            :
            ch8 ! cos[time + 1]
            ch8 ? req4
        SEQ ch6 ? req9
            :
            ch6 ? A[ time ]
            ch6 ! ack9
    PAR R[time + 1] := cos[time] * R[time] + sin[time] * A[time]
        A[time + 1] := cos[time] * A[time] - sin[time] * R[time]
                       SEQ ch9 ! req6
                           :
                           ch9 ! A[time + 1]
                           ch9 ? ack6
    time := time + 1
:
```

General Algorithm Based Error Correction and Orthogonal Polynomials

Daniel Boley [1]

Computer Science Dept.
University of Minnesota
Minneapolis, Minnesota

Abstract

We explore the relationships between sequences of orthogonal polynomials and the process of error correction based on the use of weighted checksums, often called Algorithm Based Fault Tolerance. We show that the latter process can be reduced to a problem in orthogonal polynomials. We then use methods for generating sequences of orthogonal polynnomials to solve the error correction problem, where the checksums are computed using rather general sets of weights. The methods are equivalent to the matrix Lanczos tridiagonalization process. We give a simple numerical example.

1. Introduction

The Lanczos Algorithm was originally proposed by Lanczos [19] as a method for the computation of eigenvalues of symmetric and nonsymmetric matrices. The idea was to reduce a general matrix to tridiagonal form, from which the eigenvalues could be easily determined. For symmetric matrices, the Lanczos Algorithm has been studied extensively [7, 22]. In that case, the convergence of the algorithm, when used to compute eigenvalues, has been extensively analyzed in [18, 21, 25, 26] [27, p270ff]. This algorithm is particularly suited for large sparse matrix problems. A block Lanczos analog has been studied and analyzed by Underwood (cf. Golub and Underwood [13], Cullum and Willoughby [7] and Parlett [22]). However, until recently, the nonsymmetric Lanczos Algorithm has received much less attention. Some recent computational experience with this algorithm can be found in [6]. Besides some numerical stability problems, the method suffered from the possibility of an incurable breakdown from which the only way to "recover" was to restart the whole process from the beginning with different starting vectors [27, p388ff]. More recently, several modifications allowing the Lanczos process to continue after such breakdowns have been

[0] [1]This research was partially supported by the National Science Foundation under grant CCR-8813493 and by the Minnesota Supercomputer Institute.

proposed by Parlett et al [24] and by Gutknecht [15], and a numerical implementation has been developed in [9, 10]. The close connection between the modified Non-symmetric Lanczos Algorithm and orthogonal polynomials with respect to indefinite inner products is discussed by Golub and Gutknecht [12] and Boley et. al. [2]. Recently, Parlett [23] noticed the close relation between the Lanczos Algorithm and the controllability- observability structure of dynamical systems. In this paper, we show how the matrix Lanczos Algorithm may be used to transform a sequence of polynomials into another "orthogonal" sequence, how this relationship can be exploited to handle general sorts of error correction process in Algorithm Based Fault Tolerance (ABFT) based on checksums.

The Lanczos Algorithm [19] is an example of a method that generates bases for Krylov subspaces starting with a given vector. The Arnoldi Algorithm [3] can be thought of as a "one-sided" method, which generates one sequence of vectors that span the reachable space. In this paper, we extend this idea to the use of a two-sided method, the non-symmetric Lanczos Algorithm, which generates two sequences of vectors spanning the left and right Krylov spaces.

This paper is organized as follows. First we give a short description of the Lanczos process in a rather general setting, then we show how this process can be used to generate a sequence of polynomials orthgogonal to an indefinite bilinear function ("inner product") given only the "moments," then we show how this polynomial construction applies to the error correction problem in signal processing.

2. Description of the Lanczos Process

We give a brief description of the non-symmetric Lanczos process we have implemented. For clarity, we describe the algorithms at a level of detail appropriate for a MATLAB environment, omitting the specific methods used for the basic linear algebra computations.

We consider a real vector space $\mathbf{V}$ with an associated inner product $(\mathbf{x},\mathbf{y})$ of vectors $\mathbf{x}$ such that $0 < (\mathbf{x},\mathbf{x}) < \infty$ with $(\mathbf{x},\mathbf{x}) = 0$ only if $\mathbf{x} = 0$. We suppose that there exists an orthonormal basis $\mathbf{e}_1, \mathbf{e}_2, ...$, and we express all the vectors in $\mathbf{V}$ in terms of this basis:

$$\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2 + ... = \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \end{bmatrix}$$

In this basis, a linear operator on $\mathbf{V}$ will be expressed as a matrix A, and the transpose (adjoint) A^{T} will satisfy $(A^{\mathrm{T}}\mathbf{x},\mathbf{y}) = (\mathbf{x}, A\mathbf{y})$. We will write $\mathbf{x}^{\mathrm{T}}\mathbf{y} = (\mathbf{x},\mathbf{y})$. We now describe the Lanczos algorithm in the general setting so that we may apply it to possibly infinite vectors in the sequel. We will also discuss the "nonsingularity" and "rank" of a matrix, but only for finite dimensional ones, so we have the usual definitions of these concepts.

We use the following notation, to keep the description concise. Vectors are represented by lower case bold letters ($\mathbf{b}$), matrices by upper case italic (B), and linear spaces by upper face bold ($\mathbf{B}$); all other typefaces are scalars or indices. The notation $span[\mathbf{v}_0, \mathbf{v}_1, ...]$ denotes the space spanned by the column vectors $\mathbf{v}_0, \mathbf{v}_1, ...$ If $\mathbf{v}_k = A\mathbf{v}_{k-1}$ for all k, so that $\mathbf{v}_k = A^{k-1}\mathbf{v}_1$, the sequence of vectors $\mathbf{v}_1, \mathbf{v}_2, ...$ is called a *Krylov sequence*, and the space $span[\mathbf{v}_1, \mathbf{v}_2, ...]$ is called the *right Krylov space* $\mathbf{K}$ generated by the vector $\mathbf{v}_1$. We let $\mathbf{K}_k$ denote the truncated space generated by the first k vectors: $K_k \equiv [\mathbf{b}_1, A\mathbf{b}_1, ..., A^{k-1}\mathbf{b}_1]$. Likewise, we let $\mathbf{L}$ denote the *left Krylov space* $span[\mathbf{c}_1, A^{\mathrm{T}}\mathbf{c}_1, ...]$, and $\mathbf{L}_k$ the truncated space generated by $L_k \equiv [\mathbf{c}_1, A^{\mathrm{T}}\mathbf{c}_1, ..., (A^{\mathrm{T}})^{k-1}\mathbf{c}_1]$.

Given an operator A on $\mathbf{V}$ and two non-null, vectors $\mathbf{b}_1$, $\mathbf{c}_1$ in $\mathbf{V}$, all represented as a matrix or vectors, respectively, in a particular orthogonal basis, the algorithm generates two sequences of vectors $B \equiv [\mathbf{b}_1, \mathbf{b}_2, ...]$ and $C \equiv [\mathbf{c}_1, \mathbf{c}_2, ...]$ such that

$$span[\mathbf{b}_1, ..., \mathbf{b}_k] = \mathbf{K}_k \text{ and } span[\mathbf{c}_1, ..., \mathbf{c}_k] = \mathbf{L}_k \text{ for all } k. \tag{1}$$

Given vectors $\mathbf{b}_1, ..., \mathbf{b}_k$ and $\mathbf{c}_1, ..., \mathbf{c}_k$, the vectors $\mathbf{b}_{k+1}$ and $\mathbf{c}_{k+1}$ are computed by the formulas

$$\mathbf{b}_{k+1} = A\mathbf{b}_k - [\mathbf{b}_1, ..., \mathbf{b}_k]\mathbf{h}_k$$

and

$$\mathbf{c}_{k+1} = A^{\mathrm{T}}\mathbf{c}_k - [\mathbf{c}_1, ..., \mathbf{c}_k]\mathbf{g}_k,$$

for some $(k-1)$-vectors of coefficients $\mathbf{h}_k$ and $\mathbf{g}_k$ so that (1) is satisfied automatically. The $\mathbf{h}_k$ and $\mathbf{g}_k$ are chosen to enforce certain other conditions, principally the "bi-orthogonality" condition to be described below.

The bi-orthogonality condition that we would like the vectors to satisfy is

$$\mathbf{b}_{k+1}^{\mathrm{T}}[\mathbf{c}_1, ..., \mathbf{c}_k] = 0 \text{ and } \mathbf{c}_{k+1}^{\mathrm{T}}[\mathbf{b}_1, ..., \mathbf{b}_k] = 0. \tag{2}$$

But this may not always be possible. We consider two cases. If the $k \times k$ matrix

$$[\mathbf{c}_1, ..., \mathbf{c}_k]^{\mathrm{T}}[\mathbf{b}_1, ..., \mathbf{b}_k] \text{ (or equivalently } L_k^{\mathrm{T}} K_k) \tag{3}$$

is nonsingular, then we can find the $\mathbf{h}_k$ and $\mathbf{g}_k$ to enforce (2) by the formulas

$$\mathbf{h}_k = ([\mathbf{c}_1, ..., \mathbf{c}_k]^{\mathrm{T}}[\mathbf{b}_1, ..., \mathbf{b}_k])^{-1}[\mathbf{c}_1, ..., \mathbf{c}_k]^{\mathrm{T}} A\mathbf{b}_k$$

$$\mathbf{g}_k = ([\mathbf{b}_1, ..., \mathbf{b}_k]^{\mathrm{T}}[\mathbf{c}_1, ..., \mathbf{c}_k])^{-1}[\mathbf{b}_1, ..., \mathbf{b}_k]^{\mathrm{T}} A^{\mathrm{T}}\mathbf{c}_k. \tag{4}$$

We will see below that all but two entries of $\mathbf{h}_k$ and $\mathbf{g}_k$ turn out to be zero, so that the resulting algorithm is exactly the nonsymmetric Lanczos algorithm as described in [19] [27, p388ff].

If (3) is singular (or optionally the condition number is above a certain tolerance), then we let j denote the largest index less than k such that

$$[\mathbf{c}_1, ..., \mathbf{c}_j]^{\mathrm{T}}[\mathbf{b}_1, ..., \mathbf{b}_j] \text{ (or equivalently } L_j^{\mathrm{T}} K_j) \text{ is nonsingular} \tag{5}$$

(or sufficiently well conditioned). Then we may enforce the partial bi-orthogonality condition

$$\mathbf{b}_{k+1}^{\mathrm{T}}[\mathbf{c}_1, ..., \mathbf{c}_j] = 0 \text{ and } \mathbf{c}_{k+1}^{\mathrm{T}}[\mathbf{b}_1, ..., \mathbf{b}_j] = 0, \tag{6}$$

by the formulas

$$\mathbf{h}_k = \begin{bmatrix} ([\mathbf{c}_1, ..., \mathbf{c}_j]^{\mathrm{T}}[\mathbf{b}_1, ..., \mathbf{b}_j])^{-1}[\mathbf{c}_1, ..., \mathbf{c}_j]^{\mathrm{T}} A\mathbf{b}_k \\ \\ \mathbf{h'}_k \end{bmatrix} \tag{7}$$

and

$$\mathbf{g}_k = \begin{bmatrix} ([\mathbf{b}_1, ..., \mathbf{b}_j]^{\mathrm{T}}[\mathbf{c}_1, ..., \mathbf{c}_j])^{-1}[\mathbf{b}_1, ..., \mathbf{b}_j]^{\mathrm{T}} A^{\mathrm{T}}\mathbf{c}_k \\ \\ \mathbf{g'}_k \end{bmatrix}, \tag{8}$$

where the $\mathbf{h'}_k$, $\mathbf{g'}_k$ are two $(k-j)$-vectors. If the intervening vectors $\mathbf{b}_{j+1}, ..., \mathbf{b}_k$ and $\mathbf{c}_{j+1}, ..., \mathbf{c}_k$ are all formed by this prescription, then condition (6) will be satisfied by any choice for $\mathbf{h'}_k$, $\mathbf{g'}_k$. So we will make the choice to orthogonalize (in the usual sense) the vectors $\mathbf{b}_{j+1}, ..., \mathbf{b}_k$ among themselves and the $\mathbf{c}_{j+1}, ..., \mathbf{c}_k$ among themselves.

Algorithm 1.

1. For $k = 1, 2, \ldots$ until stopped
2. Expand Krylov spaces: Set $\mathbf{b}_{k+1}^{(0)} = A\mathbf{b}_k$ and $\mathbf{c}_{k+1}^{(0)} = A^{\mathrm{T}}\mathbf{c}_k$.
3. Let j be the largest index s.t. (5) holds.
4. Enforce bi-orthogonality condition (6) by setting

$$\mathbf{b}_{k+1}^{(1)} = \mathbf{b}_{k+1}^{(0)} - [\mathbf{b}_0, \ldots, \mathbf{b}_j]\mathbf{h}_k$$
$$\mathbf{c}_{k+1}^{(1)} = \mathbf{c}_{k+1}^{(0)} - [\mathbf{c}_0, \ldots, \mathbf{c}_j]\mathbf{g}_k$$

5. Orthogonalize within last un-bi-orthogonalized cluster by setting

$$\mathbf{b}_{k+1} = \mathbf{b}_{k+1}^{(1)} - [\mathbf{b}_{j+1}, \ldots, \mathbf{b}_k]\mathbf{h}'_k \text{ s.t. } \mathbf{b}^{\mathrm{T}}[\mathbf{b}_{j+1}, \ldots, \mathbf{b}_k] = 0,$$
$$\mathbf{c}_{k+1} = \mathbf{c}_{k+1}^{(1)} - [\mathbf{c}_{j+1}, \ldots, \mathbf{c}_k]\mathbf{g}'_k \text{ s.t. } \mathbf{c}^{\mathrm{T}}[\mathbf{c}_{j+1}, \ldots, \mathbf{c}_k] = 0,$$

We note that there are several choices for the stopping condition in step 1. We choose the following. The process continues until $\mathbf{b}_{r+1} = 0$ for some r, or $\mathbf{c}_{s+1} = 0$ for some s. Suppose without loss of generality that $\mathbf{c}_{s+1} = 0$ occurs first. We may then continue expanding the right Krylov sequence $\mathbf{b}_{s+1}, \ldots, \mathbf{b}_r$ by appending a sequence of zero vectors to the corresponding left Krylov sequence $\mathbf{c}_{s+1} = \ldots = \mathbf{c}_r = 0$, but for our purposes in the next section, it will suffice to stop at step s.

The resulting vectors generated from this algorithm will satisfy certain important properties that we mention. Let $B = [\mathbf{b}_1, \ldots, \mathbf{b}_r]$ and $C = [\mathbf{c}_1, \ldots, \mathbf{c}_r]$ be the matrices of all the vectors generated. The vector $\mathbf{b}_{k+1}$ is a linear combination of $A\mathbf{b}_k$ and previous vectors $\mathbf{b}_i$, $i \leq k$. Thus the matrix B of generated vectors satisfies

$$AB = BH,$$

where H is a unit upper Hessenberg matrix consisting of all the coefficients $\mathbf{h}_k$, $k = 1, \ldots, r$. Likewise, the matrix C satisfies

$$A^{\mathrm{T}}C = CG,$$

where G is a unit upper Hessenberg matrix, consisting of all the coefficients $\mathbf{g}_k$, $k = 1, \ldots, r$. That is, the k-th columns of H and G are, respectively:

$$H_{\cdot k} = \begin{bmatrix} \mathbf{h}_k \\ 1 \\ \mathbf{0} \end{bmatrix} \text{ and } G_{\cdot k} = \begin{bmatrix} \mathbf{g}_k \\ 1 \\ \mathbf{0} \end{bmatrix},$$

where each "1" entry above occupies the $k+1$-th position, lying on the sub-diagonal of H and G, respectively, for $k = 1, \ldots, r$. The bi-orthogonality conditions (6) (4) become

$$C^{\mathrm{T}}B = D,$$

where D is a block diagonal matrix in which the lower right corners of the diagonal blocks fall exactly on those elements d_{jj} for all indices j satisfying (5). Since $C^{\mathrm{T}}AB = C^{\mathrm{T}}BH = DH$, and $B^{\mathrm{T}}A^{\mathrm{T}}C = B^{\mathrm{T}}CG = D^{\mathrm{T}}G$, we have the relation

$$G^{\mathrm{T}}D = DH. \tag{9}$$

Since a block diagonal matrix times a upper Hessenberg matrix is block upper Hessenberg, it follows that G and H are block tridiagonal, with the partitioning defined by the cluster dimensions. This implies that in computing the coefficients $\mathbf{h}_k$, $\mathbf{g}_k$ at each stage, the only nonzero entries of $\mathbf{h}_k$ and $\mathbf{g}_k$ are those corresponding to the last two diagonal blocks of the part of D generated to date: that is, $h_{\hat{j}+1,k}, ..., h_{jk}, ..., h_{kk}$ are the only nonzero entries in vector $\mathbf{h}_k$, where where $\hat{j}$, j are the last two consecutive indices less than k satisfying (5), and likewise for $\mathbf{g}_k$.

In particular, if (5) were satisfied for *every* index j, then H and G would be *scalar* tridiagonal. So step 5 of Algorithm 1 would be empty, and step 4 would reduce to

4. Enforce bi-orthogonality condition (2) by setting

$$\mathbf{b}_{k+1} = \mathbf{b}_{k+1}^{(0)} - [\mathbf{b}_{k-1}, \mathbf{b}_k] \begin{bmatrix} h_{k-1,k} \\ h_{kk} \end{bmatrix}$$
$$\mathbf{c}_{k+1} = \mathbf{c}_{k+1}^{(0)} - [\mathbf{c}_{k-1}, \mathbf{c}_k] \begin{bmatrix} g_{k-1,k} \\ g_{kk} \end{bmatrix}.$$

In this case with the current scalings, both H and G have subdiagonals all equal to 1. By equating matrix elements in (9) it follows that $G = H$. If we instead scaled the vectors $\mathbf{b}_{k+1}$, $\mathbf{c}_{k+1}$ to have unit norm, then G and H would be related by $G^{\mathrm{T}} = H$, as noted in [27, p388ff].

3. Application to Sequences of Polynomials

We explore the following problem. Suppose we have two sequences of polynomials $q_0, q_1, ...$ and $p_0, p_1, ...$ of exact degree. And suppose there exists a real-valued bilinear functional $b(f, g)$ which satisfies some of the usual properties for an inner product

$$b(f, \alpha g + h) = b(\alpha g + h, f) = \alpha b(g, f) + b(h, f)$$

and

$$b(xf, g) = b(f, xg)$$

for any real-valued functions f, g, h of x. The problem we would like to address is the problem of generating the q's to be "orthogonal" with respect to $b(\cdot, \cdot)$ knowing only the "moments"

$$\mu_k = b(p_k, p_0), \ k = 0, 1, 2, ... \tag{10}$$

where p_0 is a constant polynomial. In the case that $b(\cdot, \cdot)$ is an ordinary inner product (i.e. that $b(f, f) > 0$ for all nonzero f), this problem has been extensively studied in the literature (see e.g. section 5 as well as [11] and references therein). However, only recently has this problem been addressed for more general $b(\cdot, \cdot)$. In this section, we will show how the matrix Lanczos algorithm solves this very problem. This problem was addressed in [2] for the case that $b(\cdot, \cdot)$ was a discrete sum over a finite number of knots. The resulting algorithm is equivalent to the "non-generic modified Chebyshev algorithm" in [12].

Since the polynomials p_i, q_i, are of exact degree they obey a recurrence formula

$$x\mathbf{p}^{\mathrm{T}} = \mathbf{p}^{\mathrm{T}} Z_p \tag{11}$$

and

$$x\mathbf{q}^{\mathrm{T}} = \mathbf{q}^{\mathrm{T}} Z_q$$

where

$$\mathbf{p} = \begin{bmatrix} p_0(x) \\ p_1(x) \\ \cdot \\ \cdot \end{bmatrix} \text{ and } \mathbf{q} = \begin{bmatrix} q_0(x) \\ q_1(x) \\ \cdot \\ \cdot \end{bmatrix}$$

and Z_p, Z_q are unreduced infinite upper Hessenberg matrices. The p's and q's are also related by an infinite upper triangular matrix of coefficients U:

$$\mathbf{q}^{\mathrm{T}} = \mathbf{p}^{\mathrm{T}} U.$$

From the above definitions, we have that

$$\mathbf{p}^{\mathrm{T}} U Z_q = x\mathbf{p}^{\mathrm{T}} U = \mathbf{p}^{\mathrm{T}} Z_p U = \mathbf{p}^{\mathrm{T}} Z_p U \tag{12}$$

We are interested in exploring the relations between the the polynomials $p_0, p_1, \ldots$ with $q_0, q_1, \ldots$. We will make the simplifying assumption that the zero degree polynomials are scaled so that $p_0 = q_0$. Then (12) reduces to

$$Z_p U = U Z_q \tag{13}$$

The upper Hessenberg structure of Z_p implies, among other things, that

$$span[\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_k] = span[\mathbf{u}_1, Z_p\mathbf{u}_1, ..., Z_p^{k-1}\mathbf{u}_1] \tag{14}$$

for every k, where $\mathbf{u}_i$ denotes the i-th column of U.

We have already defined the moments (10). We define the matrix S of "mixed moments"

$$s_{ij} = b(p_{i-1}, q_{j-1}),\ i, j = 1, 2, 3, \ldots \tag{15}$$

The first column $\mathbf{s}_1$ of S is just the vector of moments $[\mu_0, \mu_1, ...]^{\mathrm{T}}$. We use the extended notation $S = b(\mathbf{p}, \mathbf{q}^{\mathrm{T}})$ following [12], where b applied to a vector means that b is applied individually to each element. By linearity we have that

$$Z_p^{\mathrm{T}} b(\mathbf{p}, \mathbf{q}^{\mathrm{T}}) = b(x\mathbf{p}, \mathbf{q}^{\mathrm{T}}) = b(\mathbf{p}, x\mathbf{q}^{\mathrm{T}}) = b(\mathbf{p}, \mathbf{q}^{\mathrm{T}}) Z_q \tag{16}$$

Equation (16) reduces to

$$Z_p^{\mathrm{T}} S = S Z_q \tag{17}$$

As with the U matrix, this implies that for all k

$$span[\mathbf{s}_1, ..., \mathbf{s}_k] = span[\mathbf{s}_1, Z_p^{\mathrm{T}}\mathbf{s}_1, ..., (Z_p^{\mathrm{T}})^{k-1}\mathbf{s}_1], \tag{18}$$

where $\mathbf{s}_i$ denotes the i-th column of S.

We now discuss some specific choices for the polynomials p and q. First of all, if the polynomials $p_i = x^i$ are the "monomials," then the recurrence matrix Z_p reduces to the "shift-down" matrix

$$\begin{bmatrix} 0 & & \\ 1 & 0 & \\ & 1 & \cdot \\ & & \cdot \cdot \\ & & \cdot \end{bmatrix}.$$

In this case, the column $\mathbf{u}_i$ of the matrix U will hold exactly the coefficients of the polynomial q_i. And the moments μ_i will be the usual classical moments with respect to the unknown bilinear functional b.

If instead we choose the p's to be a sequence of orthogonal polynomials with respect with a "usual" positive definite inner product, then the matrix Z_p will be a tridiagonal matrix, and with certain scalings, symmetric. In this case, the matrix Z_p can be generated by the symmetric Lanczos algorithm ([8, 4, 14]).

In any case, the computations that we will describe below depend on having in hand the recurrence matrix Z_p.

Independently of the choice for the p's, we can make arbitrary choices for the q's. If in particular we choose the q's to be "orthogonal" with respect to $b(\cdot,\cdot)$, then thethe corresponding matrix condition is that the matrix

$$D = b(\mathbf{q}, \mathbf{q}^{\mathrm{T}}) = b(\mathbf{q}, \mathbf{p}^{\mathrm{T}})U = S^{\mathrm{T}}U \tag{19}$$

be diagonal. We then observe that the conditions (13), (17), (19) and the Krylov sequence conditions (14) and (18) exactly match the properties of the vectors generated by the Lanczos process when started with the matrix Z_p and right and left vectors $\mathbf{u}_1 = \mathbf{e}_1$ and $\mathbf{s}_1$, respectively. It follows that if such a sequence of orthogonal q's exist, then the vectors generated by the Lanczos process will satisfy (19), and viceversa. However, if the matrix D resulting from the Lanczos process is only block diagonal, then no such complete sequence of orthogonal q's exists.

We now discuss the computation of the leading finite-dimensional part of the above infinite vectors. Suppose we are given only the first $2m-1$ moments $\mu_0, \mu_1, ...\mu_{2m-2}$ as well as the leading $2m-1 \times 2m-1$ part of Z, which we refer henceforth as Z for simplicity. Because of the lower Hessenberg form of Z^{T}, we know the first $2m-2$ entries in $Z^{\mathrm{T}}\mathbf{s}_1$, the first $2m-3$ components of $(Z^{\mathrm{T}})^2\mathbf{s}_1$, and so on. Hence, we will know the *leading anti triangle* of the Krylov sequence

$$span[\mathbf{s}_1, Z_p^{\mathrm{T}}\mathbf{s}_1, (Z_p^{\mathrm{T}})^2\mathbf{s}_1, ...,] \tag{20}$$

containing the leading $m \times m$ principal submatrix of (20). The vectors $\mathbf{u}_1, \mathbf{u}_2, ...$ and $\mathbf{s}_1, \mathbf{s}_2, ...$ satisfying (19) can be generated by applying an oblique Gram-Schmidt process to the Krylov sequences (20) and

$$span[\mathbf{u}_1, Z_p\mathbf{u}_1, Z_p^2\mathbf{u}_1, ...,]. \tag{21}$$

Due to the upper triangular nature of the vectors $\mathbf{u}_i$, the conditions (19) for the first m vectors involve only the first m entries of both the $\mathbf{u}$ and $\mathbf{s}$ vectors.

The Lanczos process will generate a sequence of vectors $\mathbf{u}_1, \mathbf{u}_2, ...$ and $\mathbf{s}_1, \mathbf{s}_2, ...$ With the first $2m-1$ entries of $\mathbf{s}_1$ known and $\mathbf{u}_1 = \mathbf{e}_1$, the Lanczos algorithm will generate at least the vectors $\mathbf{u}_1, ..., \mathbf{u}_m$ and leading m entries of $\mathbf{s}_1, ..., \mathbf{s}_m$. Each polynomial q_k will be defined in terms of the originally given set of p polynomials by the relation $q_k(x) = \mathbf{p}^{\mathrm{T}}(x)\mathbf{u}_{k+1}$i, for $K = 0, 1, 2, ...$ The moments involving q_k are the entries of $\mathbf{s}_{k+1}$:

$$\mathbf{s}_{k+1} = \begin{bmatrix} b(p_0, q_k) \\ b(p_1, q_k) \\ \cdot \\ \cdot \end{bmatrix} \tag{22}$$

If k is an index such that

$$[\mathbf{s}_1, ..., \mathbf{s}_k]^{\mathrm{T}}[\mathbf{u}_1, ..., \mathbf{u}_k] \text{ is nonsingular,} \tag{23}$$

then $\mathbf{s}_{k+1}$ will orthogonal to $[\mathbf{u}_1, ..., \mathbf{u}_k]$. Because of the upper triangular structure of U, this means simply that the first k entries of (22) will be zero. Note that this is a condition involving only finitely many leading entries of (22). So for such indices k, the polynomial q_k will be orthogonal to all polynomials of lower degree, with respect to $b(\cdot, \cdot)$. The condition that $D = S^{\mathrm{T}}U$ be diagonal is equivalent to condition that (23) hold for every k, which implies that S will be lower triangular. If D is only block diagonal, then (23) holds for only certain values of k, corresponding to the ends of the diagonal blocks. In this case S will be block lower triangular.

4. Algorithm Based Error Correction

A standard problem in ABFT [16, 17] is the correction of errors in a data sequence given a collection of checksums. This problem can be expressed in terms of sequences of polynomials as described in the previous section. In the error correction problem, we have a data sequence

$$a_1, a_2, ..., a_n, \tag{24}$$

and a collection of checksums

$$c_1, c_2, ..., c_m, \tag{25}$$

where each checksum c_j is a weighted sum of the data values

$$c_j = \sum_{i=1}^{n} a_i p_{j-1}(x_i) \tag{26}$$

for some set of functions p_j defined over a set of distinct knots x_i. In the literature, the typical set of functions are the monomials $p_j = x^j$, and the knots proposed have been $x_i = i$ [20] and $x_i = 2^{i-1}$ [17]. However, it will be seen that when the techniques of the previous section are used, the p's may be any sequence of polynomials of exact degree, and the knots may be any set of distinct points. In [1, 5] a simple modification to this technique was presented that allows correction also among the checksums. In brief, this is accomplished by appending a set of parity values to the original data values, and then carrying out the entire computation, including the computation of the checksums, on the combined set of values. The parity values are chosen just to make the true checksums identically zero, so that in fact the checksums themselves may be completely omitted from the entire computation. In this paper, we do not discuss parity values any further, though all the methods of this paper still apply if we consider m of the data entries (24) to be parity values chosen to make the checksums (25) identically zero.

Suppose that during some process involving computation or transmission, the data items become corrupted to the erroneous sequence

$$\tilde{a}_1, \tilde{a}_2, ...$$

The error correction problem is then to compute the *errors* $\omega_i = (\tilde{a}_i - a_i)$, from which we may recover the true values a_i. For this purpose, we compute the *syndromes*

$$\mu_j = \sum_{i=1}^{n} \omega_i p_{j-1}(x_i) = \left(\sum_{i=1}^{n} \tilde{a}_i p_{j-1}(x_i) \right) - c_j. \tag{27}$$

To express this problem in terms of sequences of polynomials, we define the bilinear functional

$$b(f, g) = \sum_{i=1}^{n} f(x_i) g(x_i) \omega_i$$

Then the syndromes are given by (10).

Lemma 1. Let k be the number of errors (ω values) and denote the nonzero errors by

$$\omega_{i_1}, ..., \omega_{i_k}.$$

Then there is a unique (up to scaling) polynomial $r(x)$ of lowest degree such that

$$b(f, r) = 0 \tag{28}$$

for all polynomials f of degree up to $k-1$, and the degree of r is k. On the other hand, if $q(x)$ is any nonzero polynomial of degree l satisfying $0 = b(f, q)$ for all polynomials f of degree at most $m-1$, then either $k \leq l$ or $k \geq m+1$.

Proof: The polynomial

$$r(x) = (x - x_{i_1})...(x - x_{i_k}) \tag{29}$$

is a polynomial satisfying (28). Define the Lagrange interpolating polynomials $\{t_j\}$ of degree $k-1$ over the points $x_{i_1}, ..., x_{i_k}$:

$$t_j(x_{i_j}) = \begin{cases} 1 \text{ if } j = \hat{j} \\ 0 \text{ if } j \neq \hat{j}. \end{cases}$$

If $s(x)$ were a polynomial of smaller degree satisfying (28), then $b(t_j, s) = 0$ for each j. But that means $s(x_{i_j}) = 0$ for each j, contradicting the assumption that s has degree less than k.

If $\hat{r}(x)$ were a second such polynomial of degree k, also scaled to be monic, then $s = \hat{r} - r$ would be a polynomial of smaller degree, also satisfying (28), so again we have a contradiction.

For the second part, if $k \leq m$ then $0 = b(t_j, q)$ for each t_j (since they have degree $< m$) so that $q(x_{i_j}) = 0$ for each $j = 1, ..., k$. So q must have degree $l \geq k$. □

The polynomial $r(x)$ (29) is called the "error locator polynomial," and it satisfies the following proposition, easily demonstrated from the above Lemma:

Proposition 1. For any $m \geq k-1$, the error locator polynomial (29) is the unique polynomial (up to scaling) of lowest degree satsifying (28) for all polynomials f of degree up to m. The degree of $r(x)$ is k and the zeroes of r are the knots corresponding to the nonzero ω values. □

In the previous section we considered a starting sequence of polynomials $\{p_0, p_1, ...\}$ and a second set to be generated $\{q_0, q_1, ...\}$. In this section we have already defined a sequence of polynomials $p_0, p_1, ...$ used to fix the checksum coefficients (26). We now propose to consider a second sequence of polynomials of exact degree $q_0, q_1, ...$, and we consider the problem of determining $r(x)$ in terms of the q's.

Express the error locator polynomial in terms of the q's and a coefficient vector $\mathbf{r}$:

$$r(x) = \mathbf{q}^{\mathrm{T}}\mathbf{r} = \mathbf{p}^{\mathrm{T}}U\mathbf{r}. \tag{30}$$

Then condition (28) is equivalent to

$$0 = b\left(\begin{bmatrix} p_0 \\ \cdot \\ \cdot \\ p_{k-1} \end{bmatrix}, r\right) = b\left(\begin{bmatrix} p_0 \\ \cdot \\ \cdot \\ p_{k-1} \end{bmatrix}, [q_0, ..., q_k]\right)\mathbf{r}$$

This can be written in terms of the mixed moments s_{ij} (15):

$$0 = \begin{bmatrix} s_{11} & \cdot & \cdot & s_{1k} & s_{1,k+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ s_{k1} & \cdot & \cdot & s_{kk} & s_{k,k+1} \end{bmatrix} \begin{bmatrix} r_1 \\ \cdot \\ \cdot \\ r_{k+1} \end{bmatrix} \tag{31}$$

Proposition 1 guarantees that (31) has a solution with $r_{k+1} \neq 0$, and the solution is unique once r_{k+1} is fixed.

If the number k of errors is unknown in advance, but it is known that $k \leq m$ for some given number m, then Proposition 1 guarantees that the error locator polynomial is determined by condition (28) for $k = 0, 1, ..., m-1$. So we need to solve the following overdetermined set of equations

$$0 = b\left(\begin{bmatrix} p_0 \\ \cdot \\ \cdot \\ p_{m-1} \end{bmatrix}, r\right) = b\left(\begin{bmatrix} p_0 \\ \cdot \\ \cdot \\ p_{m-1} \end{bmatrix}, [q_0, ..., q_k]\right) \mathbf{r}$$

or expressed in terms of the mixed moments:

$$0 = \begin{bmatrix} s_{11} & \cdot & \cdot & s_{1k} & s_{1,k+1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ s_{m,1} & \cdot & \cdot & s_{m,k} & s_{m,k+1} \end{bmatrix} \begin{bmatrix} r_1 \\ \cdot \\ \cdot \\ r_k \\ r_{k+1} \end{bmatrix} \tag{32}$$

where k is the smallest value of the index for which these equations have a solution. The system (32), when r_{k+1} has been fixed, is an overdetermined system unless $k = m$.

If the q's are generated by a Lanczos procedure, then the resulting vectors $\mathbf{s}_1, \mathbf{s}_2, ...$ will be lower triangular, and linearly independent. At some stage k, the vector $\mathbf{s}_{k+1}$ will be zero, so that (32) will have the trivial solution

$$\begin{bmatrix} r_1 \\ \cdot \\ \cdot \\ r_k \\ r_{k+1} \end{bmatrix} = \begin{bmatrix} 0 \\ \cdot \\ \cdot \\ 0. \\ 1 \end{bmatrix}$$

From (30), that means $r(x) \equiv q_k(x)$.

Suppose that only the first l syndrome values $s_{11}, ..., s_{l1}$ are known. Since the $\mathbf{s}$ vectors are being generated by the lower Hessenberg matrix Z_p^{T}, only the first $l-1$ entries of $\mathbf{s}_2$ can be computed, only the first $l-2$ entries of $\mathbf{s}_3$ can be computed, and in general only the first $l-j+1$ entries of $\mathbf{s}_j$ can be computed for $j = 1, 2, ...$. Hence, to compute the first m entries of the vector $\mathbf{s}_{m+1}$ as in (32) requires knowing only the first $l = 2m$ entries of $\mathbf{s}_1$.

If the Lanczos algorithm is used, $\mathbf{s}_{m+1}$ is guarenteed to be zero by Proposition 1, but we need the leading $(m+1) \times (m+1)$ part of U and the leading $(m+1) \times m$ part of S to enforce the bi-orthogonality conditions and to recover the m-th degree polynomial $q_m = r$, assuming m errors have occured. To generate the leading part of S mentioned requires the first $2m$ elements of $\mathbf{s}_1$, as before. Hence in either case, we can conclude that only $l = 2m$ syndrome values are required to determine up to m errors among the data.

7. Numerical Example

We illustrate our method with a numerical example using the Chebyshev polynomials to generate the coefficients and the knots. In printing the numbers, we have rounded them to the digits shown, even though the computations were carried out in a precision of about 16 decimal digits on a Sun using Lisp.

Example 1. The first three Chebyshev polynomials are

$$p_0(x) = 1, \quad p_1(x) = x, \quad p_2(x) = 2x^2 - 1,$$

and it is well known that the subsequent polynomials are generated by the three term recurrence

$$p_{i+1}(x) = 2x\, p_i(x) - p_{i-1}(x), \qquad \text{for } i = 1, 2, \cdots.$$

The first 9 polynomials $p_0(x), p_1(x), \cdots, p_8(x)$ are related via the recurrence (11) with the tridiagonal recurrence matrix

$$Z_p = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}.$$

The knots are chosen as the zeros of p_8, which are the same as the eigenvalues of Z_p:

$$\begin{aligned} x_1 &= \cos 15\pi/16 = -0.980785 & \qquad x_5 &= \cos\ 7\pi/16 = +0.195090 \\ x_2 &= \cos 13\pi/16 = -0.831470 & \qquad x_6 &= \cos\ 5\pi/16 = +0.555570 \\ x_3 &= \cos 11\pi/16 = -0.555570 & \qquad x_7 &= \cos\ 3\pi/16 = +0.831470 \\ x_4 &= \cos\ 9\pi/16 = -0.195090 & \qquad x_8 &= \cos\ \pi/16 = +0.980785 \end{aligned}$$

We also allow for up to 3 errors, requiring 6 syndrome values. Thus the matrix $\{p_{j-1}(x_i)\}$ in (26) is given by

$$G = \begin{pmatrix} 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 & 1.0000 \\ -0.9808 & -0.8315 & -0.5556 & -0.1951 & 0.1951 & 0.5556 & 0.8315 & 0.9808 \\ 0.9238 & 0.3827 & -0.3827 & -0.9238 & -0.9238 & -0.3827 & 0.3827 & 0.9238 \\ -0.8315 & 0.1951 & 0.9808 & 0.5556 & -0.5556 & -0.9808 & -0.1951 & 0.8315 \\ 0.7071 & -0.7071 & -0.7071 & 0.7071 & 0.7071 & -0.7071 & -0.7071 & 0.7071 \\ 0.5556 & 0.9808 & -0.1951 & -0.8315 & 0.8315 & 0.1951 & -0.9808 & 0.5556 \end{pmatrix}.$$

The first three vectors in the Krylov sequences (20) and (21) are respectively given by

$$\begin{pmatrix} -2.0000 & -2.4433 & 1.2304 \\ -2.4433 & 1.2304 & -1.1794 \\ 4.4609 & 0.0845 & 0.6698 \\ 2.6125 & 0.1091 & 0.3543 \\ -4.2426 & 0.6241 & \times \\ -1.3643 & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1.0 & 0 & 0.5 \\ 0 & 1.0 & 0 \\ 0 & 0 & 0.5 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \end{pmatrix},$$

where the first column in the first sequence above is the vector of given syndrome values defined by (27), and the symbol "×" stands for entries depending on the further syndrome values that we do not have available. The Lanczos process will generate the respective sequences

$$[\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3] = \begin{pmatrix} -2.0000 & 0 & 0 \\ -2.4433 & 4.2153 & 0 \\ 4.4609 & -5.3651 & 0 \\ 2.6125 & -3.0824 & 0 \\ -4.2426 & 5.8071 & \times \\ -1.3643 & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{pmatrix}, \quad [\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3] = \begin{pmatrix} 1.0 & -1.2217 & 0.3378 \\ 0 & 1.0000 & 0.6364 \\ 0 & 0 & 0.5000 \\ 0 & 0 & 0 \\ \vdots & \vdots & \vdots \end{pmatrix},$$

where $j \geq 2$. Note that $\mathbf{s}_3$ is all zero, so number k of errors equals 2. The error locator polynomial (29) is determined by the entries of $\mathbf{u}_3$:

$$r(x) = q_2(x) = \mathbf{u}_3^{\mathrm{T}}\mathbf{p} = 0.3378p_0(x) + 0.6364p_1(x) + 0.5000p_2(x).$$

The zeroes of this polynomial are $x_2 = -0.8315$ and $x_5 = 0.0195$, indicating that the nonzero error (ω-values) are ω_2 and ω_5. We can then extract the corresponding columns from equation (27) to obtain a 2×2 system which we then solve for those ω-values:

$$\begin{pmatrix} -2.0000 \\ -2.4433 \end{pmatrix} = \begin{pmatrix} 1.0000 & 1.0000 \\ -0.8315 & 0.1951 \end{pmatrix} \begin{pmatrix} \omega_2 \\ \omega_5 \end{pmatrix} \text{ yielding the solution } \begin{pmatrix} \omega_1 \\ \omega_4 \end{pmatrix} = \begin{pmatrix} 2.0000 \\ -4.0000 \end{pmatrix}.$$

□

8. Conclusions

We have illustrated the close connection between a variety of methods in different problem domains: the Lanczos Algorithm in linear algebra, sequences of polynomials in the theory of moments, the error correction problem in Algorithm Based Fault Tolerance, The close relations among these methods yield very simple descriptions of the methods in the various domains.

REFERENCES

[1] **D. L. Boley, R. P. Brent, G. H. Golub and F. T. Luk**, "Algorithmic Fault Tolerance Using the Lanczos Method," To appear in *SIAM J. Matrix Anal.*, 1992.

[2] **D. L. Boley, S. Elhay, G. H. Golub and M. H. Gutknecht**, "Nonsymmetric Lanczos and finding orthogonal polynomials associated with indefinite weights," *Numerical Algorithms* 1, pp 21-44, 1991.

[3] **D. L. Boley, G. H. Golub**, The Lanczos Algorithm and Controllability; *Systems and Control Letters*, vol. 4 no. 6 (1984), pp 317-324.

[4] **D. L. Boley and G. H. Golub**, "A survey of matrix inverse eigenvalue problems," in *Inverse Problems* 3, pp. 595-622, Physics Trust Publications, Bristol, England, 1987.

[5] **D. L. Boley and F. T. Luk**, "A Well Conditioned Checksum Scheme for Algorithmic Fault Tolerance," Report TR 91-27, Computer Science Dept., Univ. of Minnesota, Twin Cities, Minnesota, July 1991.

[6] **J. Cullum, W. Kerner, R. Willoughby**, A generalized nonsymmetric Lanczos procedure; *Computer Physics Communications*, vol 53 (1989), pp 19-48.

[7] **J. Cullum, R. Willoughby**, *Lanczos Algorithms for Large Symmetric Eigenvalue Computations, vol I Theory*, Birkhäuser Boston, 1985.

[8] **C. de Boor and G. H. Golub**, "The numerically stable reconstruction of a Jacobi matrix from spectral data," *Lin. Alg Applics.* 21, pp. 245-260, 1978.

[9] **R. W. Freund, M. H. Gutknecht, N. M. Nachtigal**, *An Implementation of the Look-ahead Lanczos Algorithm for Non-Hermitian Matrices, Part I*, M.I.T. Math. Numerical Analysis report 90-10, 1990.

[10] **R. W. Freund, N. M. Nachtigal**, *An Implementation of the Look-ahead Lanczos Algorithm for Non-Hermitian Matrices, Part II*, M.I.T. Math. Numerical Analysis report 90-11, 1990.

[11] **W. Gautschi**, "On generating orthogonal polynomials"; *SIAM J. Sci. and Stat. Comput.* 3, pp 289-317,1982.

[12] **G. H. Golub and M. H. Gutknecht**, "Modified moments for indefinite weight functions," *Numer. Math.* 57, pp. 607-624, 1990.

[13] **G. H. Golub and R. Underwood**, The block Lanczos method for computing eigenvalues, in J. Rice ed: *Mathematical Software III*, pp 364-377, Acad. Press, New York, 1977.

[14] **G. H. Golub and J. Welsch**, "Calculation of Gauss quadrature rules," *Math. Comp.* 23, pp. 221-230, 1969.

[15] **M. H. Gutknecht**, A Completed Theory for the Lanczos Algorithm; preprint submitted to *SIAM J. Matrix Anal.*, 1989.

[16] **K. H. Huang and J. A. Abraham**, "Algorithm-based fault tolerance for matrix operations," *IEEE Trans. Comput.* C-33 #6, pp. 518-528, June 1984.

[17] **J. Y. Jou and J. A. Abraham**, "Fault-tolerant matrix arithmetic and signal processing on highly concurrent computing structures," *Proc. IEEE* 74 #5, *Special Issue on Fault Tolerance*, pp. 732-741, May 1986.

[18] **S. Kaniel**, "Estimates for some computational techniques in linear algebra"; *Math. Comp.* 20 (1966), pp 369-378.

[19] **C. Lanczos**, "An iteration method for the solution of the eigenvalue problem linear differential and integral operators"; *J. Res. Natl. Bur. Stand.* 45 (1950), pp 255-282.

[20] **F. T. Luk**, "Algorithm-based fault tolerance for parallel matrix equation solvers," *Proceedings of SPIE Vol. 564, Real Time Signal Processing VIII*, pp. 49-53, 1985.

[21] **C. C. Paige**, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*; Ph.D. Thesis, London Univ., 1971.

[22] **B. Parlett**, *The Symmetric Eigenvalue Problem*; Prentice Hall, Englewood Cliffs, NJ, 1980.

[23] **B. N. Parlett**, Reduction to Tridiagonal Form and Minimal Realizations; preprint submitted to *SIAM J. Matrix Anal.*, 1990.

[24] **B. N. Parlett, D. R. Taylor and Z. A. Liu**, "A look-ahead Lanczos algorithm for unsymmetric matrices," *Math. Comp.*, 44, pp. 105-124, 1985.

[25] **Y. Saad**, "On the rates of Convergence of the Lanczos and the block Lanczos methods"; *SIAM J. Num. Anal.* 17 (1980), pp 687-706.

[26] **D. Scott**, "Analysis of the symmetric Lanczos process"; Univ. of Calif., Berkeley, Electronic Res. Lab. report UCB/ERL M78/40, 1978.

[27] **J. H. Wilkinson**, *The Algebraic Eigenvalue Problem*; Clarendon Press, Oxford, 1965.

ACCURATE FREQUENCY ANALYSIS OF MEASURED TIME-DEPENDENT SIGNALS OVER SHORT TIME INTERVALS*

Reo Olson and Daniel H. Cress
USAE Waterways Experiment Station
Environmental Laboratory
Vicksburg, Mississippi 39180-6199

ABSTRACT. The resolution of the frequency utilizing the Fourier Transform on a time-dependent signal is generally considered to be restricted to the inverse of the length of the time interval. This suggests that a one second long signal will permit a frequency resolution of 1 Hz.

A much more precise determination of the frequency is possible by a careful review of the phase when the source has an unknown narrow-band frequency. For purposes of the analysis presented herein, narrow-band frequency implies a bandwidth less than the normally interpreted frequency resolution. The phase of the cross-spectral density of successive time intervals indicates the difference between the Fast Fourier Transform (FFT) frequency resolution and the frequency of the input signal. Since this calculation is a trivial step after the calculation of the FFT, this method can be easily implemented on real-time systems using existing hardware for the FFT.

This method has been applied to the acoustic data obtained from a helicopter. The improved analysis of the Doppler shift of the frequency for the moving aircraft permitted a good estimate of the velocity of the approaching helicopter and its range at the closest point of approach using one microphone on the ground.

INTRODUCTION. The standard FFT calculates the amplitude and phase at equal increments in the frequency domain as determined by the formula

$$\Delta f = 1/\Delta t \tag{1}$$

where

Δf = frequency resolution of the FFT
Δt = time interval length from which the FFT was derived

This inverse relationship between the frequency resolution and the length of the time interval being analyzed has led to the belief that it is not possible to obtain accurate frequency resolution over short time intervals. However, the resolution constraint in Equation 1 is not applicable if the phase relationship among successive time intervals of length Δt is predictable (not random). An example of such a situation occurs when the frequency is unknown but has a bandwidth less than Δf.

---

* Supported by Headquarters, US Army Corps of Engineers.

In the case of a narrow-band source signal, the phase term in the frequency domain permits the accurate calculation of the frequency of the input. For the purposes of this paper, "narrow-band" refers to a signal bandwidth that is much smaller than the normal frequency resolution Δf of the FFT as defined in Equation 1. The concept behind the increased resolution can be easily understood by considering how one cosine wave would best fit another cosine wave of a slightly different frequency.

CURVE FITTING OF TWO COSINE WAVES. Suppose that we are given a time signal that is a 5.2-Hz cosine wave. What would be the best least-squares fit of a 5-Hz cosine wave of arbitrary amplitude and phase over the time interval [0,1]? The least-squares error is achieved by perfectly aligning the two curves at the midpoint of the time interval and permitting equal magnitude (but different sign) errors at the two endpoints. This means that the best fit of a 5-Hz cosine wave over [0,1] is achieved with phase 0.1 cycle and amplitude close to one. This situation is displayed in Figure 1.

Similarly, the best fit of the 5.2-Hz cosine wave by a 5-Hz cosine wave over the time interval [1,2] is with phase 0.3 cycle and amplitude close to one (see Figure 1). The difference between the phases of the 5-Hz fit on succeeding time intervals is 0.2 cycle. The simple geometry of the situation guarantees that the phase difference over any two adjacent one second long time intervals will always be 0.2 cycle for the 5.2-Hz cosine wave being approximated by a 5-Hz wave.

APPLICATION TO FFT. The FFT calculation over any 1-sec interval will attempt to fit (in a least-squares sense) integer frequency cosine waves to the input signal. The previous heuristic argument shows that, if the input signal was a 5.2-Hz cosine wave, the difference (0.2 cycle) in the 5-Hz phases of consecutive 1-sec intervals is 0.2 Hz more than the 5-Hz FFT value.

DEFINITION OF THE CROSS-SPECTRAL DENSITY. The cross-spectral density (CSD) at time t of frequency f is defined to be the product of two complex numbers:

$$CSD_t(f) = X^*_{t-1}(f) \cdot X_t(f) \tag{2}$$

where

$*$ = denotes the complex conjugate operator
$X_t(f)$ = complex FFT for frequency f where the FFT was applied to amplitudes recorded over the time period $[t,t+1]$

This definition provides obvious relationships between the FFT and CSD:

Amplitude

(3) $$|CSD_t(f)| = |X_{t-1}(f)| \cdot |X_t(f)|$$

Phase

(4) $$Ph(CSD_t(f)) = Ph(X_t(f)) - Ph(X_{t-1}(f))$$

THE KEY FORMULA. The curve-fitting analogy presented previously suggests that the phase of the cross-spectral density (when measured in cycles) defines the difference between a narrow-band (less than Δf bandwidth) input frequency and the FFT analysis frequency f (both measured in hertz). In the general case where the FFT intervals could have length different from unity, the formula becomes:

(5) $$Frequency(Hz) = f\ Hz + \frac{CSD\ Phase\ (cycles)\ of\ f\ Hz}{Length\ of\ FFT\ intervals(sec)}$$

where f is the frequency with the largest CSD amplitude and the phase of the CSD is between -0.5 and +0.5 cycle.

Proof:

Assume the signal $x(t)$ has amplitude A, frequency β Hz, and phase ϕ cycles. Then

$$x(t) = Ae^{2\pi i(\beta t+\phi)}$$

The FFT over the time interval $[(n-1)\Delta t, n\Delta t]$ of length Δt at the frequency $f=k/\Delta t$, where n and k are positive integers, is

$$X_n\left(\frac{k}{\Delta t}\right) = \int_{(n-1)\Delta t}^{n\Delta t} x(t)\, e^{-2\pi i \frac{k}{\Delta t} t}\, dt$$

$$= A \int_{(n-1)\Delta t}^{n\Delta t} e^{2\pi i\left[\left(\beta-\frac{k}{\Delta t}\right)t+\phi\right]}\, dt$$

$$= Ae^{2\pi i\phi} \int_{(n-1)\Delta t}^{n\Delta t} e^{2\pi i\left(\beta-\frac{k}{\Delta t}\right)t}\, dt$$

$$= Ae^{2\pi i\phi} \left. \frac{e^{2\pi i\left(\beta-\frac{k}{\Delta t}\right)t}}{2\pi i\left(\beta - \frac{k}{\Delta t}\right)} \right|_{t=(n-1)\Delta t}^{n\Delta t}$$

$$= \frac{Ae^{2\pi i\phi}}{2\pi i\left(\beta - \frac{k}{\Delta t}\right)} \left[e^{2\pi i\left(\beta-\frac{k}{\Delta t}\right)n\Delta t} - e^{2\pi i\left(\beta-\frac{k}{\Delta t}\right)(n-1)\Delta t}\right]$$

$$= \frac{Aie^{2\pi i\phi}}{2\pi\left(\frac{k}{\Delta t} - \beta\right)} e^{2\pi i(\beta n\Delta t-k)} \left[1 - e^{2\pi i\left(\beta-\frac{k}{\Delta t}\right)(-\Delta t)}\right]$$

$$= \frac{Aie^{2\pi i\phi}}{2\pi\left(\frac{k}{\Delta t} - \beta\right)} e^{2\pi i\beta n\Delta t} \left[1 - e^{-2\pi i\beta\Delta t}\right]$$

The CSD at time $n\Delta t$ at frequency $f=k/\Delta t$ is

$$CSD_n\left(\frac{k}{\Delta t}\right) = X^*_{n-1}\left(\frac{k}{\Delta t}\right) \cdot X_n\left(\frac{k}{\Delta t}\right)$$

$$= \frac{-Aie^{2\pi(-i)\phi}}{2\pi\left(\frac{k}{\Delta t} - \beta\right)} e^{-2\pi i\beta n\Delta t} \left[1 - e^{+2\pi i\beta\Delta t}\right]$$

$$\cdot \frac{Aie^{2\pi i\phi}}{2\pi\left(\frac{k}{\Delta t} - \beta\right)} e^{2\pi i\beta(n+1)\Delta t} \left[1 - e^{-2\pi i\beta\Delta t}\right]$$

$$= \frac{A^2}{4\pi^2\left(\frac{k}{\Delta t} - \beta\right)^2} e^{2\pi i\beta\Delta t} \left[1 - e^{2\pi i\beta\Delta t}\right]\left[1 - e^{-2\pi i\beta\Delta t}\right]$$

$$= \frac{A^2}{4\pi^2\left(\frac{k}{\Delta t} - \beta\right)^2} e^{2\pi i\beta\Delta t} \left[1 - e^{2\pi i\beta\Delta t} - e^{-2\pi i\beta\Delta t} + 1\right]$$

$$= \frac{A^2}{4\pi^2\left(\frac{k}{\Delta t} - \beta\right)^2} e^{2\pi i\beta\Delta t} \left[2 - \left(e^{2\pi i\beta\Delta t} + e^{-2\pi i\beta\Delta t}\right)\right]$$

$$= \frac{A^2\,[1 - \cos\,(2\pi\beta\Delta t)]}{2\pi^2\left(\frac{k}{\Delta t} - \beta\right)^2}\; e^{2\pi i\beta\Delta t} \cdot e^{-2\pi ik}$$

$$= \frac{A^2}{2\pi^2}\;\frac{[1 - \cos\,(2\pi\beta\Delta t)]}{\left(\frac{k}{\Delta t} - \beta\right)^2}\; e^{2\pi i\left(\beta - \frac{k}{\Delta t}\right)\Delta t}$$

$$CSD_n(f) = \frac{A^2}{2\pi^2}\;\frac{[1 - \cos\,(2\pi\beta\Delta t)]}{(f - \beta)^2}\; e^{2\pi i(\beta - f)\Delta t}$$

This means that the CSD at frequency f of an input signal of amplitude A and frequency β measured over FFT time intervals of length Δt is independent of both the phase of the input signal and also the time when the FFT analysis is performed. The CSD amplitude and phase terms are constant.

Amplitude

(6) $$|CSD(f)| = \frac{A^2[1 - \cos\,(2\pi\beta\Delta t)]}{2\pi^2(f - \beta)^2}$$

Phase (measured in cycles)

(7) $$Ph(CSD(f)) = (\beta - f)\,\Delta t$$

Equation 6 clearly shows that the frequency f that yields the largest amplitude will be the frequency that is closest to the frequency β of the input. That frequency f is the value that yields a CSD phase of absolute value less than or equal to 0.5 cycle. For this frequency f, phase Equation 7 can be rewritten as

(8) $$\beta = f + [Ph(CSD(f))/\Delta t]$$

to define the frequency of the input signal in terms of the FFT analysis frequency f, the phase of the CSD, and the length of FFT time window.

Q.E.D.

APPLICATION TO DOPPLER SHIFT. The high resolution of frequencies over short time intervals has an important application to the acoustic signature analysis of the doppler shift of moving objects (see Weidner and Sells, 1965). Figure 2 shows the expected Doppler shift of a moving acoustic source (e.g. aircraft). This graph displays the frequency shift that would be recorded by one microphone. The principal assumptions on the acoustic source that apply to these calculations are that it emits a constant basic frequency (i.e. frequency in the source reference frame) and it is moving in a straight line at a constant velocity.

The formulas (see Olson and Cress, in preparation) that determine the basic frequency (i.e. frequency before Doppler shift), velocity, and range at the closest point of approach (CPA) all require an accurate knowledge of the frequency:

(9) $$f_o = \frac{2 f_a f_r}{f_a + f_r}$$

(10) $$V = C \cdot \frac{f_a - f_o}{f_a}$$

(11) $$R = \frac{-f_o \cdot V^2}{C \cdot \frac{df}{dt}(CPA)}$$

where

f_a = the far-field approaching frequency
f_r = the far-field retreating frequency
f_o = the basic acoustic frequency of the source
C = the velocity of sound
R = the range of the source to the microphone at CPA
V = the velocity of that source

In this example, the accurate calculation of the slope near CPA requires the accurate knowledge of the frequency calculated over short time intervals.

MOVING HELICOPTER DATA. Acoustic measurements were made of a moving helicopter travelling in a straight line at a near-constant velocity. The standard FFT analysis was applied to 1-sec time intervals. Figure 3 displays the FFT amplitudes of the first 59 Hz for all 135 sec of the recording. The high-resolution frequency analysis and Doppler shift formulas were applied to the acoustic signal generated by the main rotor blades that, with the FFT analysis, showed an approaching frequency of near 19 Hz and retreating frequency near 16 Hz. (The description of the acquisition and analysis of the field data is covered more fully in Olson and Cress, in preparation).

The high-resolution frequency analysis was performed on the approaching signal in the time interval [30,45] from 30 to 45 sec. The CSD predictions of frequency at 1-sec increments are displayed in Figure 4. The individual frequencies remained consistently near 19.1 Hz during the time from 37 to 44 second. The complex vector sum of all the individual CSD measurements (see Olson and Cress, in preparation) yielded an estimate of 19.10 Hz for the far-field approaching frequency. Similarly, the far-field retreating frequency was calculated to be 15.65 Hz during the time interval [90,105]. Applying these numbers to Equation 9 yielded an estimate of 17.21 Hz for the basic frequency. This agrees with the known basic frequency of 17.2 Hz for that helicopter. The application of the approaching and predicted basic frequency into the velocity (Equation 10) yielded and estimate of 35.0 m/sec -- slightly above the 33.5 m/sec velocity reported by the pilots (but less than the 36.0 m/sec

calculated by some global positioning system (GPS) data recorded in the helicopter). Hence the velocity measurement obtained by analyzing the frequency of the Doppler shift appears to be very accurate.

The analysis of the frequency was made as the helicopter passed the CPA during the time interval [60,75]. At the time when the helicopter is passing CPA there is no Doppler shift (i.e. it is sending out the basic frequency, 17.2 Hz for this data). The high-resolution CSD frequency analysis of the 1-sec FFT intervals is displayed in Figure 5. The constant line of the basic frequency at 17.2 was added to aid in the determination of when the sound of CPA reached the microphone. The intersection of this constant line with the frequency curve just after 71 sec into the run defines this event. Interestingly, the high-resolution frequency technique is sufficiently accurate to display a near-constant slope for a few seconds before CPA. This means that it would be possible to use the slope of the frequency a few seconds before CPA to guess the slope at CPA in Equation 11. This approach was applied to obtain Figure 6. Clearly the agreement between the CSD prediction and the range reported by the pilots is good. Also reasonable estimates of the anticipated range at CPA were obtained even before the helicopter reached CPA. These estimates could not be obtained by only using the standard FFT techniques.

COSTS FOR IMPLEMENTING THE NEW TECHNIQUE. The extreme simplicity of the calculation beyond the traditional FFT calculation has some surprising benefits! First no additional computer hardware should be required beyond that needed to perform the FFT calculation. Also, only a minuscule amount of additional CPU time should be required to perform the calculations.

CONCLUSIONS. On real data the CSD provides a significant improvement over the conventional use of the FFT in the accuracy of frequencies over short time intervals. The phase of the CSD resolves the frequencies between the traditional FFT frequency increments. In the case of a single constant frequency and no noise, the CSD is completely accurate over any length of interval. On the real acoustic data of a moving helicopter the CSD high-resolution frequency analysis appears to be accurate to within 0.1 Hz. This high accuracy when combined with the analysis of Doppler theory permitted accurate prediction of the velocity and range at CPA from one passive microphone with minimal extra computing cost. The accurate frequency analysis of measured time-dependent signals over short time intervals should have many other scientific applications.

ACKNOWLEDGEMENTS. The research reported on in this paper was conducted by the U.S. Army Corps of Engineers. Permission was granted by the Chief of Engineers to publish this information.

REFERENCES.

Olson, R. E., and Cress, D. H. "Passive Acoustic Range Estimation of Helicopters," Technical Report in preparation, U.S. Army Engineer Waterways Experiment Station, Vicksburg, MS.

Weidner, R. T., and Sells, R. L. 1965. Elementary Classical Physics, Vol 2, Allyn-Bacon, Boston, pp 1049-1056.

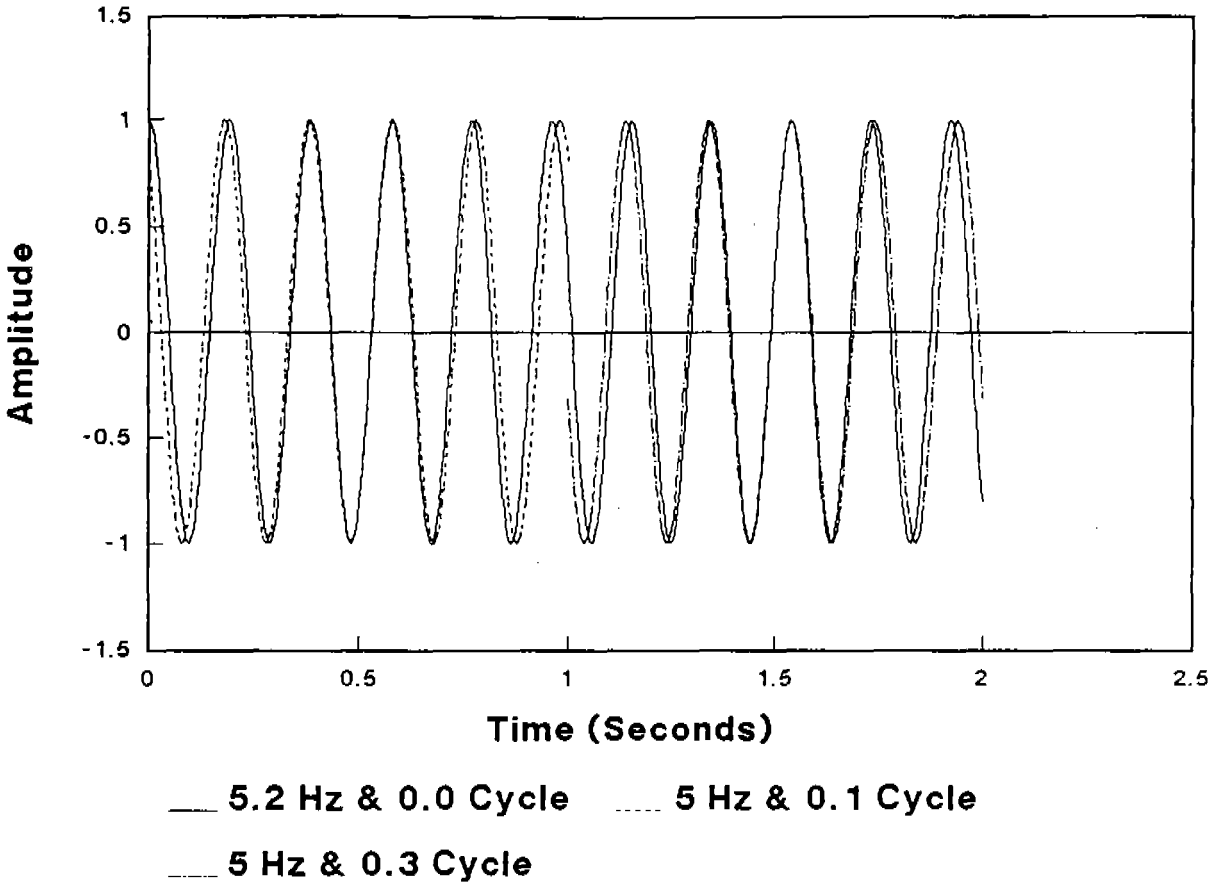


Figure 1. Curve fitting of a 5.2-Hz cosine wave by two 5-Hz cosine waves of phases 36 (0.1 cycle) and 108 degrees (0.3 cycle).

Note: The 5-Hz and 0.1-cycle curve is the best FFT fit of the 5.2-Hz and 0.0-cycle signal in the time interval [0,1]. Also, the 5-Hz and 0.3-cycle curve is the best FFT fit of the 5.2-Hz and 0.0-cycle signal in the time interval [1,2]. The difference, 0.2 cycle, between the two phases of the two 5-Hz FFT curves is caused by the fact that the frequency of the signal being fitted is 0.2-Hz different from those 5-Hz FFT curves

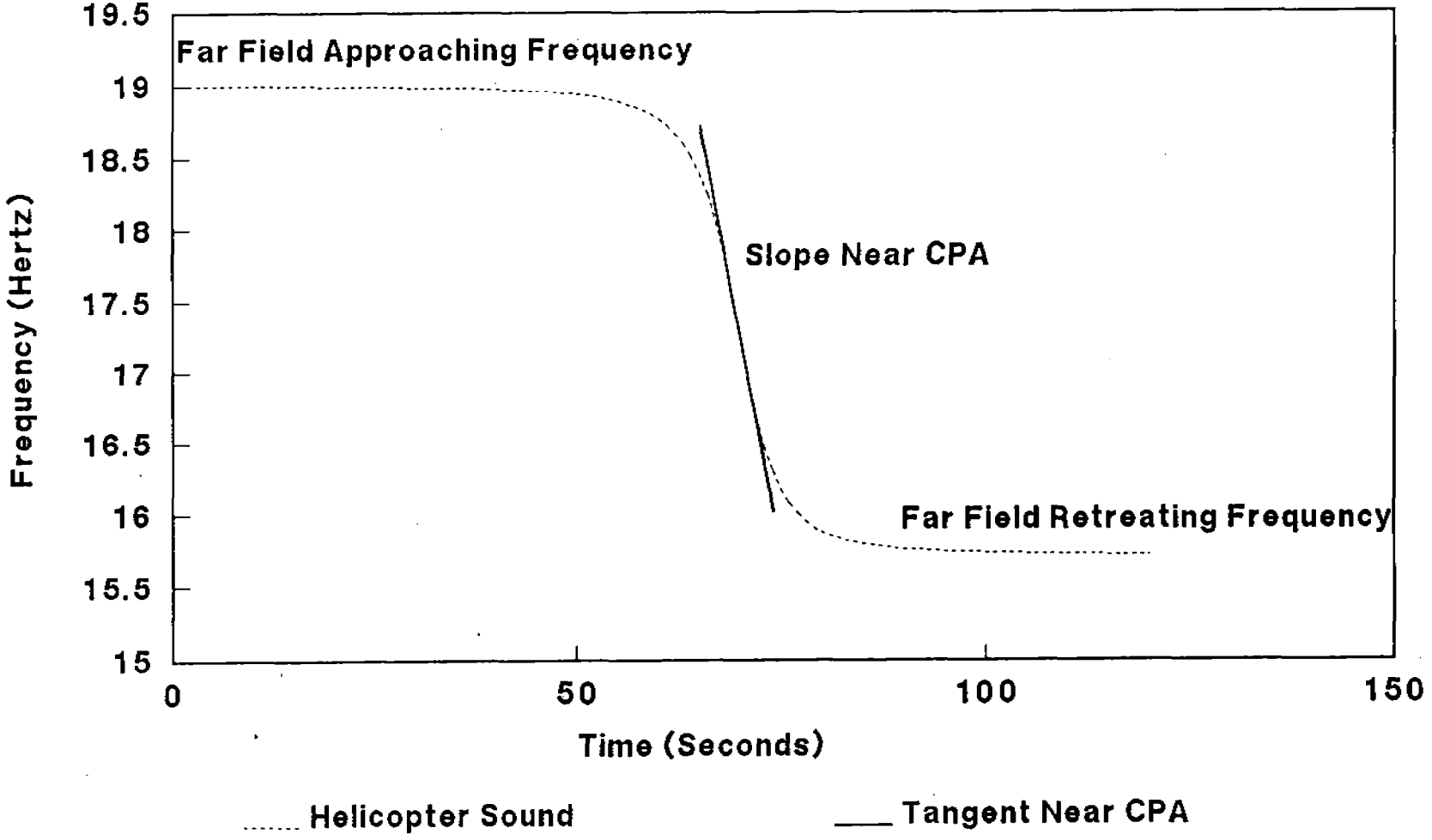


Figure 2. The key input data for passive estimation of velocity, range, basic frequency, and time of CPA. Note: if nothing is known about the helicopter, all of these data are required. If the basic frequency is known, the approaching frequency and slope are required

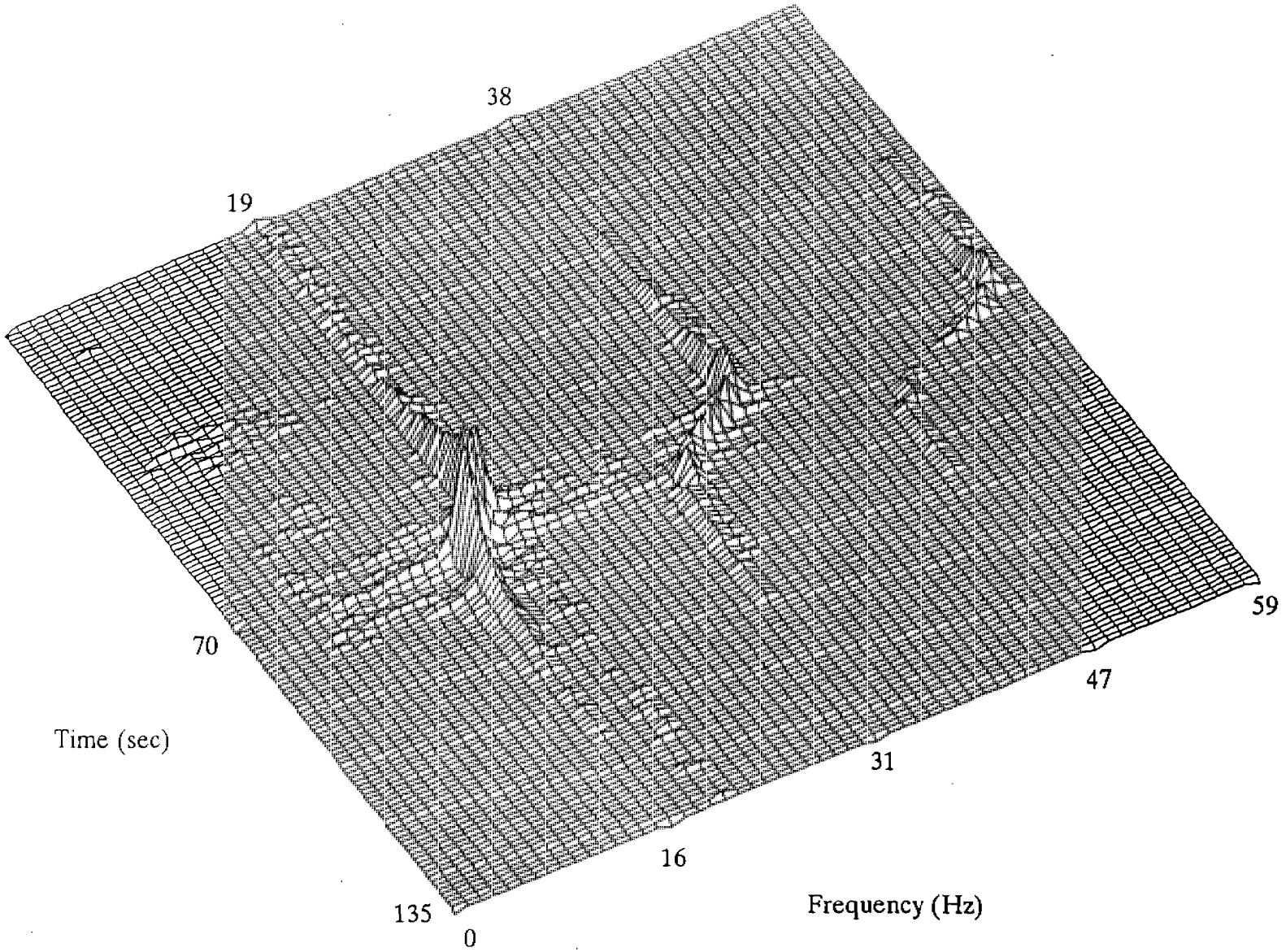


Figure 3. FFT analysis of a helicopter moving in a straight line at a constant velocity

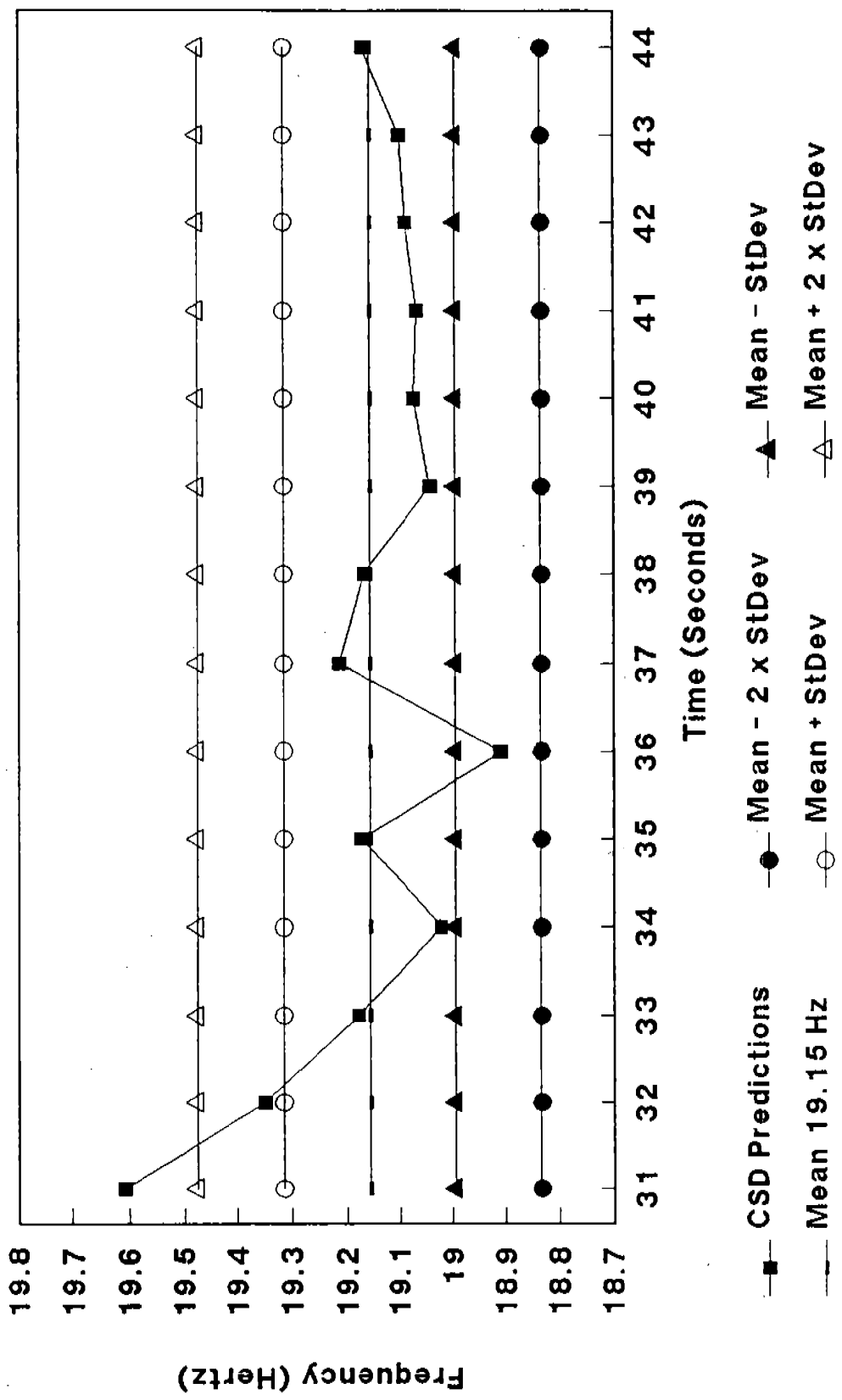


Figure 4. Frequency predicted by individual cross spectral densities of approaching main rotor energy. Note: the standard deviation (StDev) is 0.16 Hz

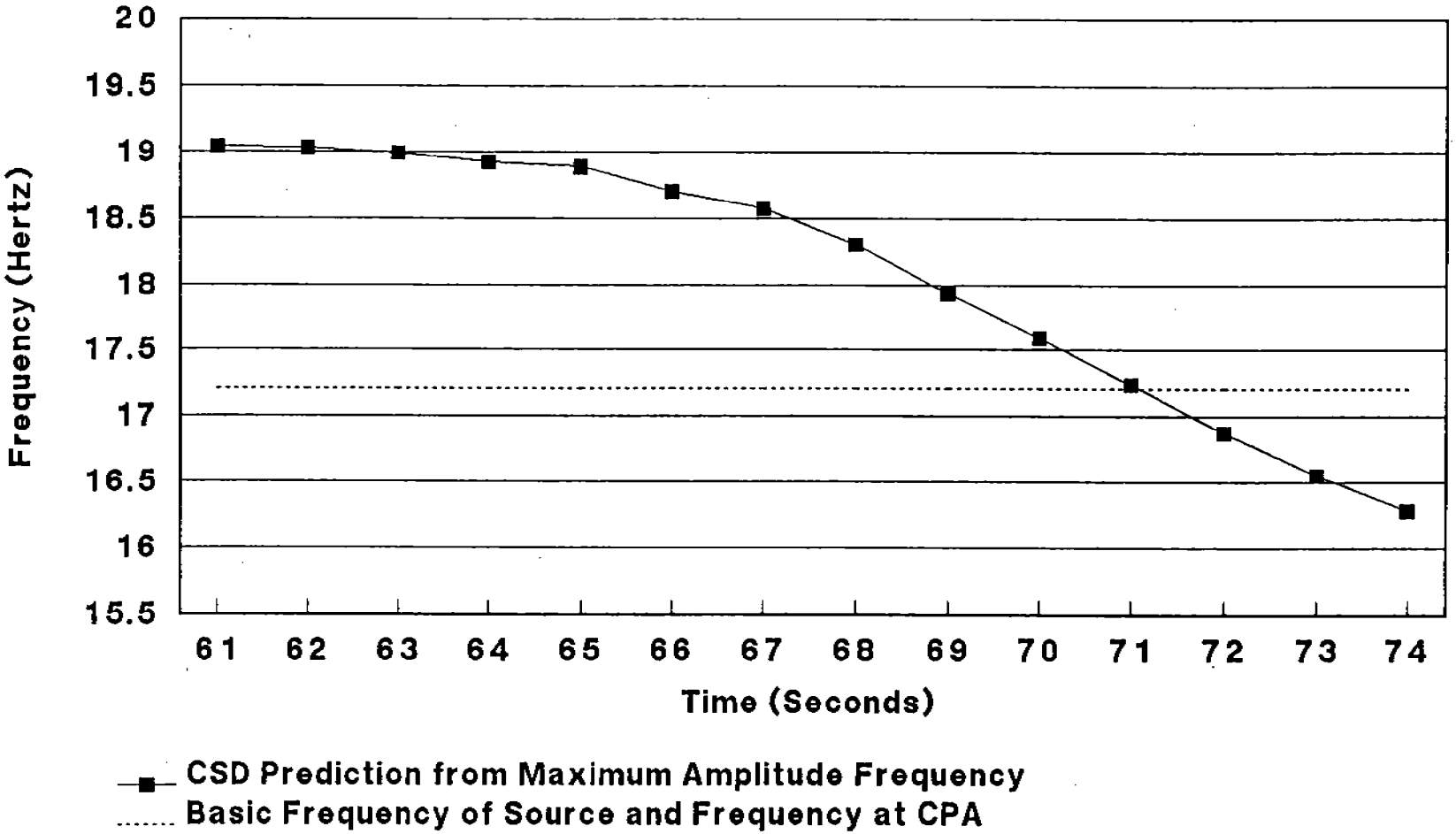


Figure 5. Frequency predicted by cross spectral density of the main rotor blades' energy crossing CPA

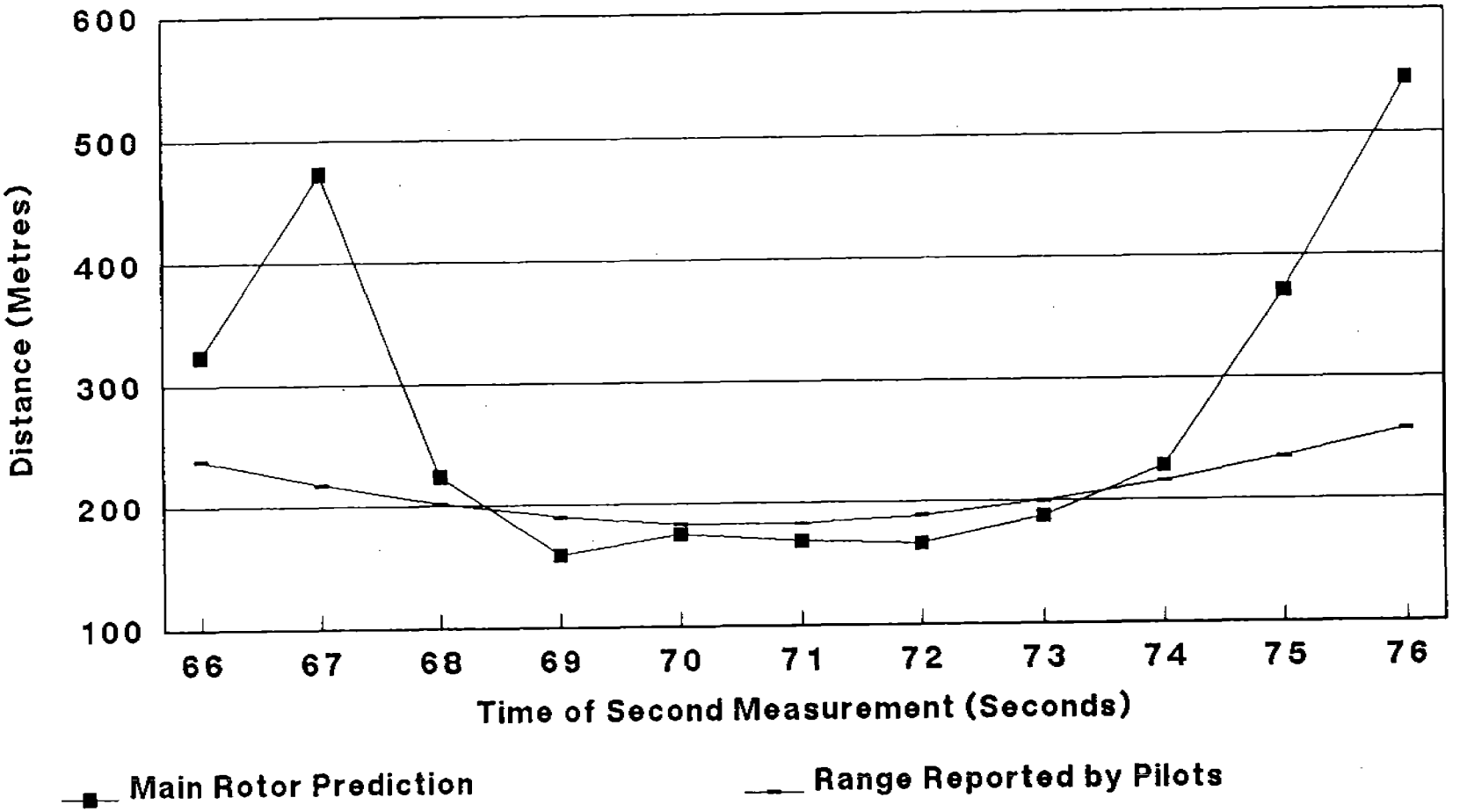


Figure 6. Prediction of range at CPA utilizing the slope of Doppler shift Run 21 - channel 1. The pilot-reported range includes the 0.52-sec shift from CPA to receiver

The Arithmetic Fourier Transform (AFT): Iterative Computation and Image Processing Applications[‡]

Donald W. Tufts and Haiguang Chen[†]

Department of Electrical Engineering
Kelley Hall
University of Rhode Island
Kingston, R. I. 02881

Abstract

A Fourier analysis method using an iterative Arithmetic Fourier Transform (AFT) is presented. It overcomes the difficulty of dense, Farey-fraction sampling which is inherent in the original AFT algorithm. This disadvantage of the AFT is turned into an advantage and dense frequency-domain samples are obtained without any additional interpolation or zero-padding. The implementation of the iterative computations is designed to preserve the advantage of the AFT for VLSI implementation by using a permuted difference coefficient structure. This iterative AFT is intended for cases in which (a) the function to be analyzed can only be sampled uniformly and at a rate close to the Nyquist rate or (b) dense frequency-domain samples are needed.

The one and two dimensional versions of the discrete cosine transform (1-D DCT) and (2-D DCT) can be simply computed using the 1-D and 2-D AFT, but dense, Farey-fraction sampling in the image domain is then required. And it also requires special computations for the marginal DCT values.

These difficulties can be overcome by the iterative 1-D or 2-D AFT. Dense samples then occur in the transform domain where they can be advantageously used for parameter estimation or the determination of a few principal components.

[‡]This research was sponsored by the SDIO/IST, managed by the Army Research Office under Contract DAAL03-86-K-0108, Donald W. Tufts, Principal Investigator.

[†]Now at Radiologic Imaging Lab, University of California, 400 Granview Dr., South San Francisco, CA 94080.

INTRODUCTION. The Arithmetic Fourier Transform (AFT) is an algorithm for accurate high speed Fourier analysis and narrow-band filtering. The arithmetic computations in the AFT can be performed in parallel. Except for a small number of scalings in one stage of the computation, only multiplications by 0, +1 and −1 are required [1]. Thus the accuracy of the AFT is limited only by the analog-to-digital conversion of the input data, not by accumulation of rounding and coefficient errors as in the Fast Fourier Transform (FFT). Furthermore, the AFT needs no storage of sine/cosine coefficients and does not requires complicated memory addressing. These properties of the AFT make it very suitable for VLSI implementation of Fourier analysis [4].

In early part of this century, a mathematician, H. Bruns, found a method for computing the Fourier series coefficients of a periodic function using Möbius inversion [2]. Later in 1945, another mathematician, Aurel Wintner, reconsidered this technique and developed an arithmetic approach to ordinary Fourier series [3]. Tufts and Sadasiv discovered the same algorithm and named it the Arithmetic Fourier Transform (AFT) [1]. They showed how parallel computations and efficient communication and control are built into the algorithm and pointed out its applications in fast Fourier analysis, narrow-band filtering, and beam-forming. Reed *et al.* have studied the Fourier analysis and signal processing using the AFT [4] and used a simple interpolation scheme to realize their extended AFT. The original Bruns' method has also been studied to provide more balanced computation for the even and odd Fourier series coefficients [5]. Boudreaux-Bartels *et al.* have analyzed the effect of sampling-time errors in the computation of the Fourier coefficients using the AFT and carried out a preliminary comparison with the method of summation by parts (SBP) [6]. Tufts *et al.* have extended the AFT to the two-dimensional case for use in image processing [7]. Fischer *et al.* have considered the analog/digital VLSI implementation of the AFT with switched capacitors [8]. The advantage of performing the AFT analysis on delta-modulation representations of functions is also being investigated [9].

Recently, a proposed method of approximately realizing the AFT by successive approximation was presented by Weiping Li [10]. His method is closely related to the least mean squares (LMS) successive approximation realization of the Discrete Fourier Transform (DFT) of Widrow *et al.* [11]. Using Weiping Li's adaptive method, only N time-domain data samples are required and about N^2 frequency-domain samples are obtained. This is in contrast with the original forward use of the AFT algorithm in which about N^2 time-domain samples are required to compute N frequency-domain samples [12]. The computations involved in this sequential AFT method are the same as those in the AFT, namely, scaling by inverse-integer factors and accumulation. The number of iterations of this sequential method depends directly on the input data length N and there are difficulties about the convergence of the approximation process to a result which is consistent with a zero-padded DFT. This can be seen from the example in Weiping Li's paper [10].

In this paper, a different iterative AFT algorithm is presented. This algorithm uses a data block of N samples to iteratively compute a set of about N^2 frequency samples. Each iteration uses the error information between the observed data and data synthesized using the original AFT algorithm [1]. If started with a properly synthesized data vector, the algorithm will con-

verge and give the AFT values at the Farey-fraction arguments which are consistent with the values given by a zero-padded DFT. Therefore, it effectively overcomes the difficulty of dense, Farey-fraction sampling by iterative use of the AFT. Dense frequency-domain samples are obtained without any interpolation or zero-padding. The implementation of this iterative method also preserves the advantage of the AFT for VLSI implementation by using a permuted difference coefficient structure (PDC) [13] to provide simple computation of the updated Fourier transform vector. PDC is equivalent to the mathematical formulation known as Summation by Parts (SBP) which is a finite difference analog to the integration by parts reformulation of an integral found in any standard calculus book [26, 23, 27]. The arithmetic computation of this iterative AFT has a high degree of parallelism and the resulting architecture is regular. Because of its simplicity, this iterative AFT method could be of interest in many applications such as phase retrieval [14, 15], two-dimensional maximum entropy power spectral estimation [17] and recursive digital filter design [18], where many Fourier transform and inverse Fourier transform calculations are required. The iterative AFT method could be naturally used with the AFT in these applications to perform the Fourier analysis efficiently.

In Section 2, the block iterative computation of the AFT is discussed. In Section 3, the determination of the minimum norm solution for the frequency-domain samples using the steepest descent method is addressed. Then in Section 4, the minimum norm solution is used to obtain the Fourier transform solution. Example of the iterative computation of an oversampled Fourier transform is presented in Section 5. In Section 6 we present illustrate application of AFT for the computation of Discrete Cosine Transform. Section 7 concludes the paper.

BLOCK-ITERATIVE COMPUTATION OF THE AFT. In order to compute N uniformly spaced time-domain samples $x[n]$ using the AFT, we require F frequency-domain samples $X[\frac{k}{m}]$, of the Fourier transform of $x[n]$, at the Farey-fraction values of $\frac{k}{m}$ [19]. The samples $x[n]$ and $X[\frac{k}{m}]$ are related by

$$X[\frac{k}{m}] = \sum_{n=1}^{N} x[n] cos(2\pi \cdot n \cdot \frac{k}{m}), \tag{1}$$

$$with \quad m = 1, \cdots, N; \quad k = 0, \cdots, m-1.$$

The Farey-fraction sequence of order N is defined as the ascending series of irreducible rational fractions between 0 and 1 (both inclusive) with denominators which do not exceed N [12]. For example, the sequence of Farey-fractions of order 5 in the interval $[0, 1]$ are

$$\frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1}.$$

The total number F of frequency-domain, Farey-fraction samples in the interval $[0, 1]$ corresponding to N time-domain samples can be estimated

$$N_f = 3\left(\frac{N}{\pi}\right)^2 + O(N \ln N). \tag{2}$$

That is, much denser frequency-domain samples are needed for the AFT than the conventional inverse discrete Fourier transform (IDFT). The dense, Farey-fraction samples in the frequency-domain are useful rational approximations. Given any real frequency value f, we can always find a "nearby" Farey fraction $\frac{k}{m}$ of order N which gives the approximating error $e_f = |f - \frac{k}{m}|$ bounded by $\frac{1}{m(N+1)}$. If $m > \frac{N}{2}$, the error compares well with the approximate approximating error bound $\frac{\pi^2}{12N^2}$ resulting from the same number of uniformly spaced DFT frequency-domain samples [19].

Formula (1) can be expressed in a matrix form as

$$\underline{\mathbf{X}} = \mathbf{C}\underline{\mathbf{x}} \tag{3}$$

where $\mathbf{C}$ is an $F \times N$ matrix whose i^{th} column is the cosine function $cos(2\pi it)$ sampled at values of t which are Farey-fraction values $\frac{k}{m}$. The vectors $\underline{\mathbf{x}}$ and $\underline{\mathbf{X}}$ are defined as

$$\underline{\mathbf{x}} = (x[1]\ x[2]\ \ldots\ x[N])^T, \tag{4}$$

$$\underline{\mathbf{X}} = \left(X[0]\ X[\frac{1}{N}]\ \ldots\ X[\frac{k}{m}]\ \ldots\ X[\frac{N-1}{N}]\right)^T \tag{5}$$

with T denoting the vector transpose operation. The elements $X[0]$ to $X[\frac{N-1}{N}]$ are arranged according to the order of the increasing Farey-fraction values $\frac{k}{m}$. Then according to the AFT algorithm by Tufts and Sadasiv [1], we can get

$$x[n] = \sum_{m=1}^{\lfloor \frac{N}{n} \rfloor} \mu[m] s[m \cdot n], \quad n = 1, \ldots, N. \tag{6}$$

where $\lfloor \frac{N}{n} \rfloor$ indicates the integer part of $\frac{N}{n}$ and $\mu[m]$ is the Möbius function [19] defined on the positive integers by

$$\mu(m) = \begin{cases} 1, & if \quad m = 1; \\ (-1)^s, & if \quad m = p_1 \cdot p_2 \cdots p_s, \ where\ p_i\ are\ distinct\ primes; \\ 0, & if \quad p^2 | m \quad for\ any\ prime\ p. \end{cases} \tag{7}$$

where the vertical bar notation $p^2|m$ means that the integer p^2 divides the integer m exactly without remainder. The function $s[n]$ of the integer argument n is defined by

$$s[n] = \frac{1}{n} \sum_{m=0}^{n-1} X[\frac{m}{n}], \quad n = 1, \ldots, N. \tag{8}$$

Because $\mu(m)$ in (7) only takes on values $+1$, -1 and 0 and $s[n]$ in (8) can be easily computed with summation and scaling, formula (6) provides a very simple way of determining $x[n]$ from samples of $X[\frac{k}{m}]$ in the AFT [1].

When given N uniformly spaced time-domain samples $x[n]$, we can determine the frequency-domain samples $X[\frac{k}{m}]$ at the Farey-fraction values $\frac{k}{m}$ by iterative use of the AFT. From (6) and (8) above, we can relate $x[n]$ and $X[\frac{k}{m}]$ by the AFT matrix $\mathbf{A}$ as

$$\underline{\mathbf{x}} = \mathbf{A}\underline{\mathbf{X}} \tag{9}$$

The AFT matrix $\mathbf{A}$ has dimensions of $N \times F$ and rank N. The AFT matrix $\mathbf{A}$ for $N = 5$ is

$$\mathbf{A} = \begin{bmatrix} -\frac{1}{30} & -\frac{1}{5} & 0 & -\frac{1}{3} & -\frac{1}{6} & -\frac{1}{2} & -\frac{1}{6} & -\frac{1}{3} & 0 & -\frac{1}{5} \\ \frac{1}{4} & 0 & -\frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 & -\frac{1}{4} & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{4} & 0 \\ \frac{1}{5} & \frac{1}{5} & 0 & 0 & \frac{1}{5} & 0 & \frac{1}{5} & 0 & 0 & \frac{1}{5} \end{bmatrix}. \tag{10}$$

$\mathbf{A}$ is sparse and its nonzero coefficients are all proper fractions with numerator 1 and denominators which are all integer numbers.

If $N > 2$, then $F > N$ and the augmented matrix $(\mathbf{A} : \underline{\mathbf{x}})$ has the rank N. There is then an infinite number of solutions of $\underline{\mathbf{X}}$ in (9) for a given $\underline{\mathbf{x}}$. The Fourier transform vector $\underline{\mathbf{X}}$ of formula (3) and the minimum norm vector are two special solutions of (9). The minimum norm vector is defined as

$$\underline{\mathbf{X}}^{\star} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\underline{\mathbf{x}} = \mathbf{M}\underline{\mathbf{x}} \tag{11}$$

where $\mathbf{M} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$ is the matrix which provides minimum norm solution. This solution, in general, is not equal to the Fourier transform solution.

The steepest descent algorithm has been widely used for solving least squares problems in adaptive signal processing [21]. It can also be used for solving our minimum norm problem for $\underline{\mathbf{X}}^{\star}$ if we properly choose the initial vector $\underline{\mathbf{X}}_0$. Then the minimum norm solution $\underline{\mathbf{X}}^{\star}$ can be used to determine the Fourier transform vector $\underline{\mathbf{X}}$. Let $\underline{\mathbf{X}}_k$ be the k^{th} approximation to $\underline{\mathbf{X}}^{\star}$ and the synthesized signal $\underline{\mathbf{x}}_k = \mathbf{A}\underline{\mathbf{X}}_k$, then the approximation error vector is given by

$$\underline{\mathbf{e}}_k = \underline{\mathbf{x}} - \underline{\mathbf{x}}_k = \underline{\mathbf{x}} - \mathbf{A}\underline{\mathbf{X}}_k. \tag{12}$$

The squared norm of the error vector is

$$E_k = \underline{\mathbf{e}}_k^T \underline{\mathbf{e}}_k. \tag{13}$$

We update the vector of frequency-domain samples by the steepest descent method

$$\underline{\mathbf{X}}_{k+1} = \underline{\mathbf{X}}_k - \alpha \cdot \nabla E_k, \tag{14}$$

where α is the step factor of the updating and

$$\nabla E_k = -2\mathbf{A}^T(\underline{\mathbf{x}} - \mathbf{A}\underline{\mathbf{X}}_k) \tag{15}$$

is the gradient of E_k. Formula (15) can be substituted into (14) to give the following two additional forms of the updating procedures:

$$\underline{\mathbf{X}}_{k+1} = (\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})\underline{\mathbf{X}}_k + 2\alpha\mathbf{A}^T\underline{\mathbf{x}} \tag{16}$$

and

$$\underline{\mathbf{X}}_{k+1} = \underline{\mathbf{X}}_k + 2\alpha \mathbf{A}^T(\underline{\mathbf{x}} - \mathbf{A}\underline{\mathbf{X}}_k) = \underline{\mathbf{X}}_k + 2\alpha \mathbf{A}^T \underline{\mathbf{e}}_k \tag{17}$$

In next section, we will show that starting with a properly chosen initial vector $\underline{\mathbf{X}}_0$, the iterative updating process defined by (14) or (16) will converge and give the minimum norm solution $\underline{\mathbf{X}}^\star$. Then the minimum norm solution $\underline{\mathbf{X}}^\star$ can be used to determine the Fourier transform vector $\underline{\mathbf{X}}$.

CONVERGENCE TO THE MINIMUM-NORM SOLUTION. Starting with an initial vector $\underline{\mathbf{X}}_0$, we can successively use formula (16) to write the k^{th} approximation $\underline{\mathbf{X}}_k$ in the following way

$$\underline{\mathbf{X}}_k = (\mathbf{I} - 2\alpha \mathbf{A}^T\mathbf{A})^k \underline{\mathbf{X}}_0 + 2\alpha \sum_{i=0}^{k-1} (\mathbf{I} - 2\alpha \mathbf{A}^T\mathbf{A})^i \mathbf{A}^T \underline{\mathbf{x}}. \tag{18}$$

In our case, if $N > 2$, $\mathbf{A}^T\mathbf{A}$, which has F rows and F columns, is only positive semidefinite. The matrix $\mathbf{A}$ has rank N and $N < F$. Using the singular value decomposition (SVD) [25], $\mathbf{A}$ can be written as

$$\mathbf{A}_{N\times F} = \mathbf{U}_{N\times N} \mathbf{\Sigma}_{N\times F} \mathbf{V}^T_{F\times F} \tag{19}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $(N \times N)$ and $(F \times F)$ orthogonal matrices, respectively, $\mathbf{\Sigma}$ is an $(N \times F)$ pseudo-diagonal matrix which has the form

$$\mathbf{\Sigma} = \left[\tilde{\mathbf{\Sigma}}_{N\times N} : \tilde{\mathbf{O}}_{N\times(F-N)}\right]. \tag{20}$$

The square $(N \times N)$ matrix $\tilde{\mathbf{\Sigma}}$ is a diagonal matrix composed of the non-zero singular values σ_i, $i = 1, 2, \ldots, N$, of the matrix $\mathbf{A}$ and $\tilde{\mathbf{O}}$ is an $(N \times (F - N))$ zero matrix. The matrix $(\mathbf{I} - 2\alpha \mathbf{A}^T\mathbf{A})^i$ is therefore

$$(\mathbf{I} - 2\alpha \mathbf{A}^T\mathbf{A})^i = \mathbf{V} \begin{bmatrix} \mathbf{D}^i & \hat{\mathbf{O}} \\ \tilde{\mathbf{O}} & \bar{\mathbf{I}} \end{bmatrix} \mathbf{V}^T, \tag{21}$$

where the $(N \times N)$ matrix $\mathbf{D}$ has the form

$$\mathbf{D} = \begin{bmatrix} 1 - 2\alpha\sigma_1^2 & 0 & \cdots & 0 & 0 \\ 0 & 1 - 2\alpha\sigma_2^2 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & \cdots & 0 & 1 - 2\alpha\sigma_N^2 \end{bmatrix}, \tag{22}$$

$\hat{\mathbf{O}}_{N\times(F-N)}$ and $\tilde{\mathbf{O}}_{(F-N)\times N}$ are zero matrices, and $\bar{\mathbf{I}}_{(F-N)\times(F-N)}$ is an $(F - N) \times (F - N)$ identity matrix. Because of $\bar{\mathbf{I}}$, the first term in (18) will not converge to the zero vector for any α. But if we choose the initial vector $\underline{\mathbf{X}}_0$ to be the zero vector, $\underline{\mathbf{0}}$, the first term will have no effect on the iterative process (18) and then we can show that the iterative process converges to the minimum norm solution.

Considering the matrix $(\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})^i\mathbf{A}^T$ in the second term of (18), we get

$$(\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})^i\mathbf{A}^T = \mathbf{V}\begin{bmatrix} \mathbf{D}^i & \hat{\mathbf{O}} \\ \bar{\mathbf{O}} & \bar{\mathbf{I}} \end{bmatrix}\begin{bmatrix} \bar{\Sigma} \\ \tilde{\mathbf{O}}^T \end{bmatrix}\mathbf{U}^T = \mathbf{V}\begin{bmatrix} \bar{\mathbf{D}}^i \\ \tilde{\mathbf{O}}^T \end{bmatrix}\mathbf{U}^T \tag{23}$$

where $\bar{\mathbf{D}}^i$ is an $(N \times N)$ diagonal matrix with the (j,j)th element being $(1 - 2\alpha\sigma_j^2)^i\sigma_j$. If we choose the step factor α in the range

$$0 < \alpha < \frac{1}{\sigma_{max}^2} \tag{24}$$

where σ_{max}^2 is the biggest eigenvalue of $\mathbf{A}^T\mathbf{A}$, then

$$\lim_{k\to\infty}\sum_{i=0}^{k}(1 - 2\alpha\sigma_j^2)^i\sigma_j = \frac{1}{2\alpha\sigma_j}, \tag{25}$$

since this is an infinite sum of a geometric series with the ratio $r = (1 - 2\alpha\sigma_i^2)$ and $|r| < 1$. Therefore the matrix $\bar{\mathbf{D}}^i$ will approach the zero matrix and we get

$$2\alpha \lim_{k\to\infty}\sum_{i=0}^{k-1}\bar{\mathbf{D}}^i = \bar{\Sigma}^{-1} \tag{26}$$

and

$$2\alpha \lim_{k\to\infty}\sum_{i=0}^{k-1}(\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})^i\mathbf{A}^T = \mathbf{V}\begin{bmatrix} \bar{\Sigma}^{-1} \\ \tilde{\mathbf{O}}^T \end{bmatrix}\mathbf{U}^T. \tag{27}$$

Since the matrix $\mathbf{A}\mathbf{A}^T$ can be written as

$$\mathbf{A}\mathbf{A}^T = \mathbf{U}\left[\bar{\Sigma} : \tilde{\mathbf{O}}\right]\mathbf{V}^T\mathbf{V}\begin{bmatrix} \bar{\Sigma} \\ \tilde{\mathbf{O}}^T \end{bmatrix}\mathbf{U}^T = \mathbf{U}\bar{\Sigma}^2\mathbf{U}^T, \tag{28}$$

and thus

$$(\mathbf{A}\mathbf{A}^T)^{-1} = \mathbf{U}(\bar{\Sigma}^2)^{-1}\mathbf{U}^T. \tag{29}$$

Therefore, the pseudoinverse matrix which provides the minimum norm solution of formula (11) is

$$\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1} = \mathbf{V}\begin{bmatrix} \bar{\Sigma}^{-1} \\ \tilde{\mathbf{O}}^T \end{bmatrix}\mathbf{U}^T. \tag{30}$$

From (27) and (30) we see that

$$2\alpha \lim_{k\to\infty}\sum_{i=0}^{k-1}(\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})^i\mathbf{A}^T\underline{\mathbf{x}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\underline{\mathbf{x}}. \tag{31}$$

That is, the minimum norm solution can be computed iteratively.

OBTAINING THE FOURIER TRANSFORM SOLUTION. The difference of the Fourier transform solution and the minimum norm solution is determined by

$$\underline{\mathbf{X}} - \underline{\mathbf{X}}^* = \mathbf{C}\underline{\mathbf{x}} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\underline{\mathbf{x}} = \mathbf{P}\underline{\mathbf{x}}, \tag{32}$$

where the matrix $\mathbf{P}$ is defined as

$$\mathbf{P} = \mathbf{C} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1} \tag{33}$$

and $\mathbf{C}$ is given in (3). Considering the initial vector given by

$$\underline{\mathbf{X}}_0 = \mathbf{P}\underline{\mathbf{x}} \tag{34}$$

we have

$$\mathbf{A}\underline{\mathbf{X}}_0 = \mathbf{A}\mathbf{P}\underline{\mathbf{x}} = \mathbf{A}(\mathbf{C} - \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1})\underline{\mathbf{x}}. \tag{35}$$

The matrix multiplication $\mathbf{AC}$ is the original AFT operation [1] and $\mathbf{AC} = \mathbf{I}$ with $\mathbf{I}$ being an $(N \times N)$ identity matrix. Thus we get

$$\mathbf{A}\underline{\mathbf{X}}_0 = (\mathbf{AC} - \mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1})\underline{\mathbf{x}} = (\mathbf{I} - \mathbf{I})\underline{\mathbf{x}} = \underline{\mathbf{0}} \tag{36}$$

for any input $\underline{\mathbf{x}}$. Therefore, if started with this initial vector, the k^{th} approximation in (18) will be

$$\underline{\mathbf{X}}_k = \underline{\mathbf{X}}_0 + 2\alpha \sum_{i=0}^{k-1}(\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})^i\mathbf{A}^T\underline{\mathbf{x}} = \mathbf{P}\underline{\mathbf{x}} + 2\alpha \sum_{i=0}^{k-1}(\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})^i\mathbf{A}^T\underline{\mathbf{x}} \tag{37}$$

and we have

$$\lim_{k\to\infty} \underline{\mathbf{X}}_k = \underline{\mathbf{X}}_0 + 2\alpha \lim_{k\to\infty} \sum_{i=0}^{k-1}(\mathbf{I} - 2\alpha\mathbf{A}^T\mathbf{A})^i\mathbf{A}^T\underline{\mathbf{x}} = \mathbf{C}\underline{\mathbf{x}}. \tag{38}$$

Thus, the convergence of the iteration process (18) will not be affected by this properly chosen initial vector, and the final result is the the Fourier transform vector $\underline{\mathbf{X}}$. In fact, from formula (36) above, $\underline{\mathbf{X}}_0$ cannot pass the AFT filter and can be computed independently in parallel with the iterative process.

Therefore, we can realize the iterative arithmetic Fourier transform using the scheme shown in Fig. 1 and the steps of the iterative Arithmetic Fourier transform algorithm can be specified as follows:

1. Specify the maximum tolerance in the squared norm E_m of the error vector $\underline{\mathbf{e}}_k$ or specify a maximum number of iterations;
2. Calculate the initial frequency-domain vector $\underline{\mathbf{X}}_0 = \mathbf{P}\underline{\mathbf{x}}$ where $\underline{\mathbf{x}}$ is the signal vector in (4);
3. Synthesize the time-domain signal vector $\underline{\mathbf{x}}_k = \mathbf{A}\underline{\mathbf{X}}_k$ using the AFT filter;
4. Calculate the error signal vector $\underline{\mathbf{e}}_k = \underline{\mathbf{x}} - \underline{\mathbf{x}}_k$ and squared norm E_k;
5. Update the frequency-domain vector $\underline{\mathbf{X}}_{k+1} = \underline{\mathbf{X}}_k + 2\alpha\mathbf{A}^T\underline{\mathbf{e}}_k$;
6. Repeat steps 3 $-$ 5 by incrementing the iterate index k until a satisfactory convergence ($E_k \leq E_m$) has been achieved or the maximum number of iterations have been completed

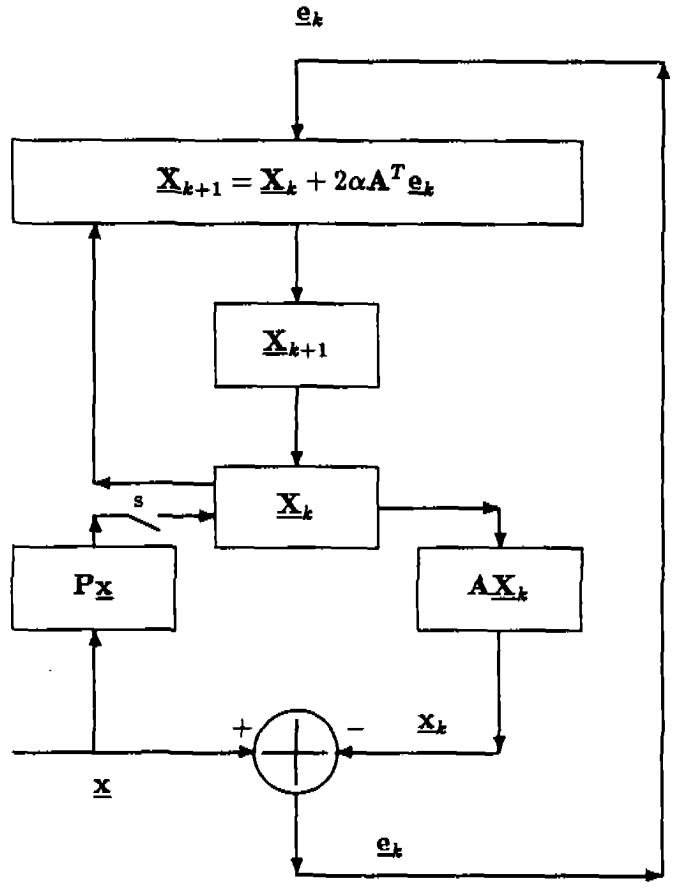


Figure 1: Block diagram of the computation of the iterative arithmetic Fourier transform

The switch s can be closed at the first step ($k = 0$) in order to obtain the initial frequency-domain vector $\underline{X}_0 = \mathbf{P}\underline{x}$. However, the computation of $\mathbf{P}\underline{x}$ can also be carried out in parallel with the iterations and added to $\underline{X}_k$ when the iterations cease or even at an intermediate time.

Alternatively, the iterations can be started with the zero vector in step 2. We have shown above that the solution component $\mathbf{P}\underline{\mathbf{x}}$ is orthogonal to the row subspace of $\mathbf{A}$. It can be computed in parallel with thee iterative computation of the minimum-norm solution (formula (11)) of the AFT equations (formula (9)). The contribution $\mathbf{P}\underline{\mathbf{x}}$ to the solution can then be added at any convenient time.

Since the computation of $\underline{\mathbf{x}}_k = \mathbf{A}\underline{\mathbf{z}}_k$ in this algorithm can be easily realized by the forward AFT and only needs multiplications by $0, +1, -1$ and a small number of scalings, most multiplications required in this iterative method arise from the computations of $\mathbf{A}^T\underline{\mathbf{e}}_k$ and $\mathbf{P}\underline{\mathbf{x}}$. Based on the properties of the AFT matrix $\mathbf{A}$ and the difference matrix $\mathbf{P}$, the computation of $\mathbf{A}^T\underline{\mathbf{e}}_k$ and $\mathbf{P}\underline{\mathbf{x}}$ can be implemented with only few multiplications by using a permuted difference coefficient (PDC) structure [13, 23, 26].

EXAMPLE 1: ITERATIVE AFT. As an example of the iterative AFT, we calculate the transform of a signal $x[n] = cos(2\pi f_1 n) + cos(2\pi f_2 n)$ with $n = 1, 2, \cdots, 10$ and $f_1 = 0.1$, $f_2 = 0.2$. There are 33 elements in the corresponding vector $\underline{\mathbf{X}}_k$ of Farey-fraction frequency-domain samples. In Fig. 2, the solid line represents the values of the continuous function $X(f)$ defined by

$$X(f) = \sum_{n=1}^{N} x[n] cos(2\pi \cdot n \cdot f). \tag{39}$$

$X[\frac{k}{m}]$ defined in formula (1) are samples of $X(f)$ at Farey fraction values $\frac{k}{m}$. The values of the iterative AFT at different iterations are shown by the asterisks. The squared time-domain error norm E_k and the squared frequency-domain error norm $E_f(k)$ are plotted in Fig. 3 as functions of the iteration number k, where the squared frequency-domain error norm is defined by

$$E_f(k) = \sum_{j,m} \left[X[\frac{j}{m}] - X_k[\frac{j}{m}] \right]^2 \tag{40}$$

$$for \;\; j = 0, 1, \cdots, m; \;\; m = 1, 2, \cdots, N.$$

The values of the squared norms in the time and frequency domains, defined by formulas (13) and (40), respectively, are different, even at the same iteration, because of the Farey-fraction sampling in the frequency domain. The initial AFT spectrum is determined by $\underline{\mathbf{X}}_0 = \mathbf{P}\underline{\mathbf{x}}$. From Fig. 2 and Fig. 3, we can see that after several iterations, the iterative AFT quickly reduces the squared error norms and the results converge to the DFT values.

COMPUTATIONAL COMPLEXITY. Since the computation of the iterative AFT is based on successive approximation, it is clear that the accuracy and the computational complexity of the algorithm depend on the iteration number k. In Table 2, we provide the number of Farey-fraction frequency-domain values N_f as a function of N, the number of elements in the time-domain vector $\underline{\mathbf{x}}$ for $N = 10$ to $N = 26$. Also tabulated are the corresponding values of N_m, N_p and N_c. These are each the number of different values of multiplication coefficients

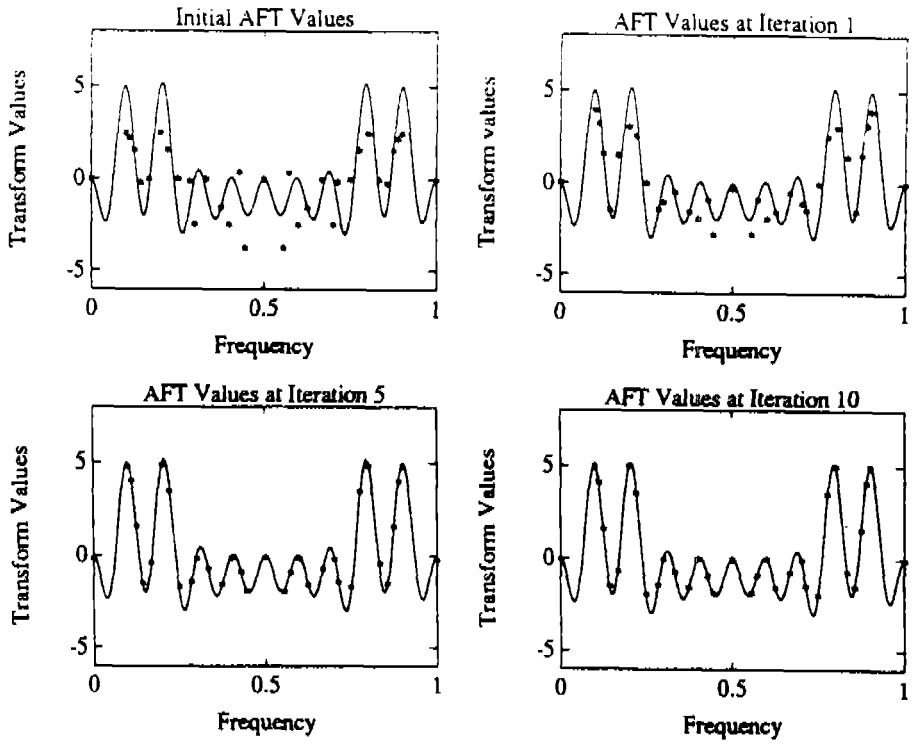


Figure 2: Transforms of iterative AFT and DFT

required for direct matrix-vector product implementation of the computations of the minimum-norm solution $\mathbf{M}\underline{\mathbf{x}}$ of formula (11), the additional solution component $\mathbf{P}\underline{\mathbf{x}}$ of formula (32), and the direct DFT computation $\mathbf{C}\underline{\mathbf{x}}$ of formula (3), respectively. Since N_f increases with N at a rate about $N^2/3$, the numbers of multiplications required in direct implementation without iteration also increase very fast with the number of elements, N, of the time-domain data vector.

The use of the iterative approach to compute $\mathbf{M}\underline{\mathbf{x}}$ reduce the number of multiplications related to N_m in Table 2. In each iteration we need $(N-1)$ scalings by integer numbers in the computation $\mathbf{A}\underline{\mathbf{X}}_{\mathbf{k}}$ and about N multiplications in the computation $\mathbf{A}^{\mathbf{T}}\underline{\mathbf{e}}_{\mathbf{k}}$. Therefore, in k iterations, about $2 \cdot k \cdot N$ multiplications are required. Since N_m varies with N at a rate of about N^2, if the iteration number k is less than $N/2$, then $2 \cdot k \cdot N < N^2$ and the iterative approach requires fewer multiplications than the direct method in the computation of $\mathbf{M}\underline{\mathbf{x}}$, in general. Some values of multiplications required in the direct computation $\mathbf{C}\underline{\mathbf{x}}$ and the iterative computation for $k = 5$ and $k = 10$ iterations are shown in Table 3, where $N_i(k)$ is the number required by the iterative AFT for k iterations. It can be seen that fewer multiplications are required in the iterative AFT. Still further reduction is possible by only computing the solution component $\mathbf{P}\underline{\mathbf{x}}$. This will be discussed below.

Theoretically, an infinite number of iterations is required to achieve the minimum time-

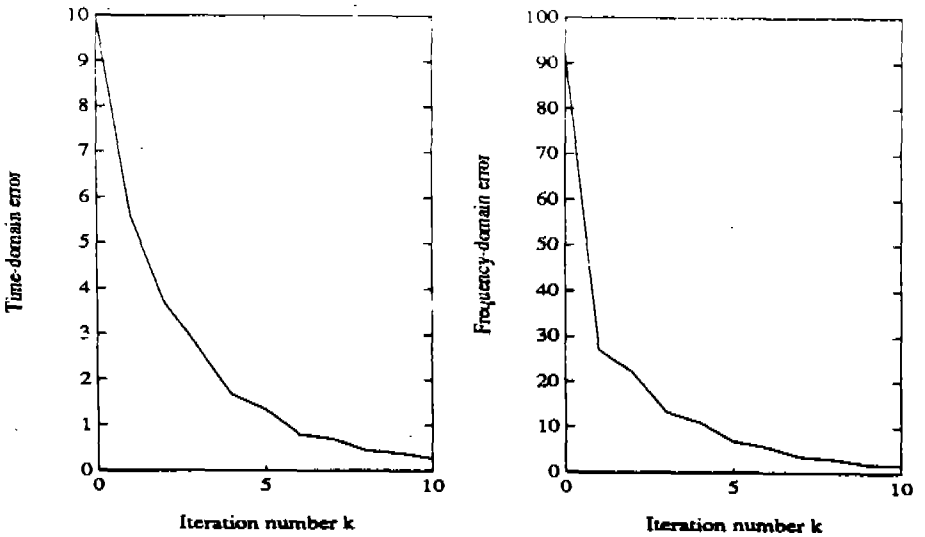


Figure 3: Squared error norms in iterative AFT

N	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
N_f	31	41	45	57	63	71	79	95	101	119	127	139	149	171	179	199	211
N_m	56	76	94	128	142	162	188	222	264	290	326	354	398	492	584	724	846
N_p	48	98	102	134	158	188	206	304	332	456	490	584	608	732	794	868	946
N_c	98	158	170	254	278	342	374	518	549	730	770	908	973	1251	1308	1596	1705

Table 1: Numbers of direct multiplication coefficients N_m, N_p and N_c required in the computations of $\mathbf{M}\underline{\mathbf{x}}$, $\mathbf{P}\underline{\mathbf{x}}$ and $\mathbf{C}\underline{\mathbf{x}}$, respectively, as functions of N and N_f, the sizes of the time-domain and frequency-domain vectors, respectively

domain error which is zero. However, in the practice of VLSI implementation of this iterative algorithm, the accuracy of the computation is limited by the input A/D conversion process, the finite word length of the internal accumulation registers, and the implementation of the scaling operations and the PDC operations. Because of this, the minimum time-domain error norm cannot be reduced to zero value. Since the convergence rate of this iterative algorithm is exponential [21], only several iterations are necessary to reach the minimum error norm and to obtain the final result of dense frequency-domain samples. Thus, the results in Table 2 and Table 3 show that for realistic ranges of the required number of iterations, the iterative AFT requires a smaller number of multiplications.

The matrix $\mathbf{P}$ resulting from the difference of $(\mathbf{C} - \mathbf{M})$ provides good transform-domain vector $\underline{\mathbf{X}}_0 = \mathbf{P}\underline{\mathbf{x}}$ by itself without adding the minimum-norm solution $\mathbf{M}\underline{\mathbf{x}}$. In Fig. 4, the values of the transform-domain vector $\underline{\mathbf{X}}_0$ for the case $N = 10$ are plotted on top each other

N	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
$N_i(5)$	148	208	222	264	298	338	366	474	512	646	690	794	828	962	1034	1118	1206
$N_i(10)$	248	318	342	394	438	488	526	644	692	836	890	1004	1048	1192	1274	1368	1466
N_c	98	158	170	254	278	342	374	518	549	730	770	908	973	1251	1308	1596	1705

Table 2: Numbers of multiplications in direct and iterative computations for $N = 10$ to $N = 26$

for individual time-domain signals $x[n] = cos(2\pi \cdot n \cdot ft)$. The test frequency ft takes 100 different values and these values are equally spaced in the frequency range $[0, 0.5]$. The values of the transform-domain vectors $\underline{\mathbf{X}}_0 = \mathbf{P}\underline{\mathbf{x}}$ are plotted versus the difference frequency (f-ft) and superposed. The matrix filtering operation $\mathbf{P}\underline{\mathbf{x}}$ thus provides a good set of closely spaced, overlapping, narrow-band filters by itself without adding the minimum-norm contribution $\mathbf{M}\underline{\mathbf{x}}$. This is not surprising because the minimum-norm frequency-domain solution $\mathbf{M}\underline{\mathbf{x}}$ will suppress spectral peaks because of its minimum-norm property. For many applications, such as initial spectrum estimation prior to parametric modeling, the initial computation using the $\mathbf{P}$ matrix alone provides sufficient accuracy.

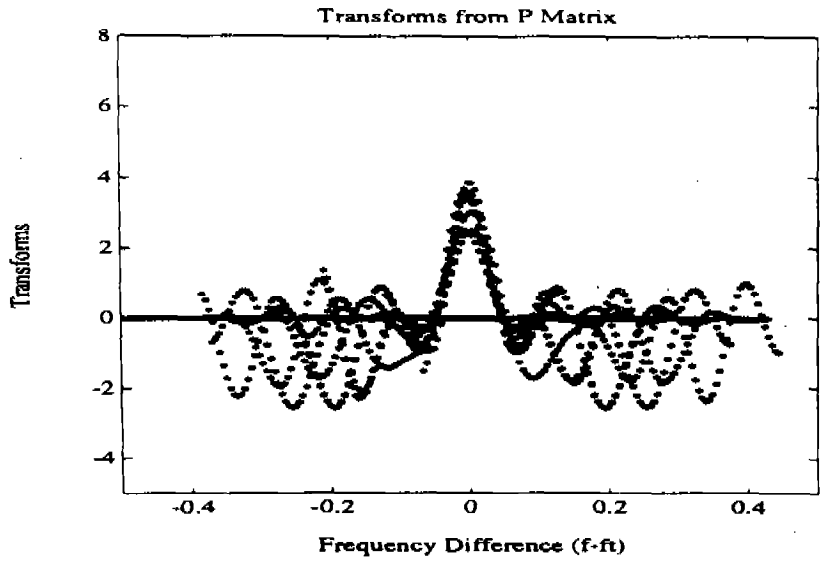


Figure 4: Superposition of frequency-domain vectors $\mathbf{P}\underline{\mathbf{x}}$ for 100 uniformly spaced values of the test frequency
ft in the Nyquist frequency range $[0, 0.5]$ plotted versus the difference frequency $(f - ft)$

The matrix $\mathbf{P}$ is also suitable for implementation with the permuted difference coefficient structure. The number of multiplications in $\mathbf{P}$ can be more effectively reduced than the matrix $\mathbf{C}$. As a result, the total number of multiplications in the iterative approach will be less than the number of multiplications in the direct implementation of $\mathbf{C}\underline{\mathbf{x}}$ when the condition $k < N/2$

is satisfied, in general. The decomposition of the computation $\mathbf{C}\underline{\mathbf{x}}$ to the computations of $\mathbf{P}\underline{\mathbf{x}}$ and $\mathbf{M}\underline{\mathbf{x}}$ and the property of $\mathbf{A}\underline{\mathbf{X}}_0$ discussed in Section 4 also allows the parallel computation of $\mathbf{P}\underline{\mathbf{x}}$ and $\mathbf{M}\underline{\mathbf{x}}$, which could provide further time saving from the direct computation of $\mathbf{C}\underline{\mathbf{x}}$.

COMPUTATION OF DISCRETE COSINE TRANSFORM. The discrete cosine transform (DCT) is an orthogonal transformation. Its basis vectors are sampled cosine functions [28]. The one dimensional DCT and inverse discrete cosine transform (IDCT) of an N point real sequence x_n are defined by

$$c_k = \frac{2}{N} e_k \sum_{n=0}^{N-1} x_n \cdot cos\left[\frac{(2n+1)k\pi}{2N}\right] \tag{41}$$

$$x_n = \sum_{k=0}^{N-1} e_k \cdot c_k \cdot cos\left[\frac{(2n+1)k\pi}{2N}\right] \tag{42}$$

$$for \quad 0 \leq n, \quad k \leq N-1,$$

respectively, where

$$e_k = \begin{cases} \frac{1}{\sqrt{2}} & if \quad k = 0; \\ 1 & otherwise. \end{cases} \tag{43}$$

The basis set of DCT is a good approximation to the eigenvectors of the class of Toeplitz matrices. It has been shown that DCT offers a higher effectiveness than the discrete Fourier transform (DFT) and performs very close to the statistically optimal Karhunen-Loève transform (KLT) when used for coding signals with Markov-1 statistics [29]. DCT can be expressed as

$$c_k = \frac{2}{N} e_k Re\{e^{-\frac{ik\pi}{2N}} \frac{1}{2N} \sum_{n=0}^{2N-1} x_n e^{-i2\pi\frac{kn}{2N}}\} \tag{44}$$

where $x_n = 0$ for $n = N,\ N+1,\ \cdots,\ 2N-1$ and $Re\{\cdot\}$ represents taking the real part of the term enclosed. Therefore, the N point DCT can be computed using $2N$ point fast Fourier transform (FFT). Some other algorithms have also been proposed for the fast computation or the simple implementation of DCT [30, 31]. In this study, we investigate the use of the iterative arithmetic Fourier transform [32] to realize DCT.

COMPUTATION OF DCT WITH THE AFT. The arithmetic Fourier transform (AFT) has been proposed for computation of the DCT by Tufts *et al.* [7]. Considering the continuous function

$$x(t) = \sum_{k=0}^{N-1} e_k \cdot c_k \cdot cos(\pi kt) = \sum_{k=0}^{N-1} e_k c_k(t) \tag{45}$$

where

$$c_k(t) = c_k \cdot cos(\pi kt) \tag{46}$$

we can see that $x(t)$ has period 2 and x_n is obtained by sampling $x(t)$ at time $t = (n + \frac{1}{2})/N$ for $n = 0, \cdots, N-1$. If there is no zero-frequency component, namely, $c_0 = 0$, we get

$$x(t) = \sum_{k=1}^{N-1} c_k \cdot cos(\pi k t) \tag{47}$$

and x_n is determined by

$$x_n = \sum_{k=1}^{N-1} c_k cos\left[\frac{(2n+1)k\pi}{2N}\right]. \tag{48}$$

Similar to the AFT algorithm, we define a set of delay-line filters

$$D_n(t) = \frac{1}{n} \sum_{m=0}^{n-1} x(t - \frac{2m}{n}). \tag{49}$$

Note that for the DCT the wider sampling interval $[0, \frac{2(N-2)}{N-1}]$ is required than the sampling interval $[0, \frac{N-1}{N}]$ in the AFT.

Substituting (7) into (9) and rearranging the order of summations, we get

$$\begin{aligned} D_n(t) &= \tfrac{1}{n} \textstyle\sum_{m=0}^{n-1} \sum_{k=1}^{N-1} c_k cos[\pi k(t - \frac{2m}{n})] \\ &= \textstyle\sum_{k=1}^{N-1} c_k \frac{1}{n} \sum_{m=0}^{n-1} cos[\pi k(t - \frac{2m}{n})] \end{aligned} \tag{50}$$

$$for \;\; n = 1, \cdots, N-1.$$

Since

$$\frac{1}{n} \sum_{m=0}^{n-1} cos[\pi k(t - \frac{2m}{n})] = \begin{cases} cos(\pi k t), & if \; k = l \cdot n \; for \; some \; integer \; l; \\ 0, & if \; n \not| \; k, \end{cases} \tag{51}$$

the output of the delay-line filter is

$$D_n(t) = \sum_{k=1}^{\lfloor (N-1)/n \rfloor} c_{nk} \cdot cos(\pi n k t) = \sum_{k=1}^{\lfloor (N-1)/n \rfloor} c_{nk}(t), \tag{52}$$

$$for \;\; n = 1, \cdots, N-1;$$

where $\lfloor (N-1)/n \rfloor$ denotes the largest integer which is less than or equal to $(N-1)/n$. Applying the Möbius inversion formula to (12), we get

$$c_k(t) = \sum_{n=1}^{\lfloor (N-1)/k \rfloor} \mu(n) D_{nk}(t), \;\; for \; k = 1, \cdots, N-1. \tag{53}$$

Sampling $c_k(t)$ at $t = 0$, we obtain the formula for the discrete cosine transform using the AFT as

$$c_k = c_k(t)|_{t=0} = \sum_{n=1}^{\lfloor (N-1)/k \rfloor} \mu(n) D_{nk}(0), \tag{54}$$

$$for \; k = 1, 2, \cdots, N-1.$$

This computation needs only additions except for a small amount of multiplications by prescribed scale factors. Also, the high degrees of regularity and parallelism of the AFT make it very suitable for VLSI implementation. From (9), we can see that, for computation of discrete cosine transform using the AFT, the dense data samples of $x(t)$ at non-equally spaced fractions $\frac{2m}{n}$ $(n = 1, \cdots, N-1; \quad m = 0, \cdots, n-1)$ are required [12]. The sampling range is increased to $[0, \frac{2(N-2)}{N-1}]$ from the original sampling interval $[\frac{1}{2N}, \frac{2N-1}{2N}]$ and there is also a zero-mean requirement for the signal x_n.

ITERATIVE COMPUTATION OF THE DCT. The iterative AFT could be used to calculate the discrete cosine transform with the same data set x_n in (1) and therefore overcome the dense, Farey-fraction sampling problem. The previous requirement that $x(t)$ has zero mean can also be eliminated. Furthermore, dense frequency-domain samples will be obtained.

Considering first the discrete-time cosine transform (DTCT) defined by

$$c(\omega) = \sum_{n=0}^{N-1} x_n cos((n + \frac{1}{2})\omega) \tag{55}$$

we can see that except the scaling factors, the DCT can be considered as samples of the DTCT with $\omega = \frac{k}{N}\pi, k = 0, 1, \cdots, N-1$. The function $c(\omega)$ has the period of 4π. Similar also to the AFT algorithm [1], we define a set of N delay-line filters which have outputs [34]

$$S(2m+1) = \frac{1}{2m+1} \sum_{k=0}^{2m} c(4\pi \frac{k}{2m+1}) = \frac{1}{2m+1} \sum_{k=0}^{2m} \hat{c}(\frac{k}{2m+1}) \tag{56}$$

with $\hat{c}(\frac{k}{2m+1}) = c(4\pi \frac{k}{2m+1})$ for $m = 0, 1, \cdots, N-1$. Substituting (55) into (56) and exchanging the orders of summations, we get

$$\begin{aligned} S(2m+1) &= \tfrac{1}{2m+1} \textstyle\sum_{k=0}^{2m} \sum_{n=0}^{N-1} x_n cos((n + \frac{1}{2}) \frac{4k\pi}{2m+1}) \\ &= \textstyle\sum_{n=0}^{N-1} x_n \frac{1}{2m+1} \sum_{k=0}^{2m} cos(\frac{2n+1}{2m+1} 2k\pi) \end{aligned} \tag{57}$$

Since

$$\frac{1}{2m+1} \sum_{k=0}^{2m} cos(\frac{2n+1}{2m+1} 2k\pi) = \begin{cases} 1, & if \ \frac{2n+1}{2m+1} = l, \ \ for \ some \ \ integer \ l, \\ 0, & otherwise; \end{cases} \tag{58}$$

we get

$$S(2m+1) = \sum_n x_n \quad for \ \ \frac{2n+1}{2m+1} = \ integer \ l. \tag{59}$$

Using the Möbius inversion formula, we obtain the formula for determining the time domain signal x_n from the outputs $S(2m+1)$ of delay-line filters

$$x_n = \sum_{l=0}^{\lfloor \frac{\frac{2N-1}{2n+1}-1}{2} \rfloor} \mu(2l+1) S((2l+1)(2n+1)) \tag{60}$$

for $n = 0, 1, \cdots, N-1$. There is no need of multiplications of cosine coefficients. The sampling instants of frequency domain samples are the Farey-fraction values of odd-number denominator. For example, the sequence of 19 sampling points for $N = 5$ are

$$0, \frac{1}{9}, \frac{1}{7}, \frac{1}{5}, \frac{2}{9}, \frac{2}{7}, \frac{1}{3}, \frac{2}{5}, \frac{3}{7}, \frac{4}{9}, \frac{5}{9}, \frac{4}{7}, \frac{3}{5}, \frac{6}{9}, \frac{5}{7}, \frac{7}{9}, \frac{4}{5}, \frac{6}{7}, \frac{8}{9}.$$

The total number F of frequency domain samples in the interval $[0, 4\pi]$ corresponding to N time domain samples can be estimated as

$$F(N) = 3(\frac{2N}{\pi})^2 + O(N\ln(N)). \tag{61}$$

That is, much more frequency domain samples are needed for determination of time domain samples using the Möbius inversion formula (60).

The matrices corresponding to operations of the delay-line filters (56) and the Möbius inversion (60) for $N = 5$ are

$$\begin{bmatrix} S(1) \\ S(3) \\ S(5) \\ S(7) \\ S(9) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 \\ \frac{1}{7} & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 \\ \frac{1}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9} \end{bmatrix} \begin{bmatrix} \hat{c}(0/9) \\ \hat{c}(1/9) \\ \hat{c}(1/7) \\ \hat{c}(1/5) \\ \hat{c}(2/9) \\ \hat{c}(2/7) \\ \hat{c}(1/3) \\ \hat{c}(2/5) \\ \hat{c}(3/7) \\ \hat{c}(4/9) \\ \hat{c}(5/9) \\ \hat{c}(4/7) \\ \hat{c}(3/5) \\ \hat{c}(2/3) \\ \hat{c}(5/7) \\ \hat{c}(7/9) \\ \hat{c}(4/5) \\ \hat{c}(6/7) \\ \hat{c}(8/9) \end{bmatrix} \tag{62}$$

and

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 & -1 & -1 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} S(1) \\ S(3) \\ S(5) \\ S(7) \\ S(9) \end{bmatrix} \tag{63}$$

respectively. The computation of x_n from $\hat{c}(\frac{k}{2m+1})$ can be expressed in a matrix form as

$$\underline{\mathbf{x}} = \mathbf{A}\underline{\hat{\mathbf{c}}}. \tag{64}$$

The vectors $\underline{\hat{c}}$ and $\underline{x}$ are defined as

$$\underline{\hat{c}} = \left[\hat{c}(0)\ \hat{c}(\frac{1}{2N-1})\ \ \hat{c}(\frac{1}{2N-3})\ \cdots\ \hat{c}(\frac{2(N-1)}{2N-1})\right]^T \tag{65}$$

and

$$\underline{x} = [x_0\ x_1\ \cdots\ x_{N-1}]^T \tag{66}$$

respectively, where T represents the vector transpose operation. The elements of $\underline{\hat{c}}$ are arranged according to the increasing order of values $(\frac{k}{2m+1})$ for $m = 1, \cdots, N-1;\ k = 0, \cdots, 2m$. Combining the delay-line filter matrix (62) and the Möbius inversion filter matrix (63), we obtain the following matrix $\mathbf{A}$ of describing the complete AFT operation for $N = 5$ as

$$\begin{bmatrix}
\frac{34}{105} & 0 & \frac{-1}{7} & \frac{-1}{5} & 0 & \frac{-1}{7} & \frac{-1}{3} & \frac{-1}{5} & \frac{-1}{7} & 0 & 0 & \frac{-1}{7} & \frac{-1}{5} & \frac{-1}{3} & \frac{-1}{7} & 0 & \frac{-1}{5} & \frac{-1}{7} & 0 \\
\frac{2}{9} & \frac{-1}{9} & 0 & 0 & \frac{-1}{9} & 0 & \frac{2}{9} & 0 & 0 & \frac{-1}{9} & \frac{-1}{9} & 0 & 0 & \frac{2}{9} & 0 & \frac{-1}{9} & 0 & 0 & \frac{-1}{9} \\
\frac{1}{5} & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 & 0 & \frac{1}{5} & 0 & 0 \\
\frac{1}{7} & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 & 0 & \frac{1}{7} & 0 \\
\frac{1}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9} & \frac{1}{9} & 0 & 0 & \frac{1}{9} & 0 & \frac{1}{9} & 0 & 0 & \frac{1}{9}
\end{bmatrix} \tag{67}$$

When given N uniformly spaced time domain samples $x_n, n = 0,\ 1,\ \cdots,\ N-1$, we can determine the frequency domain samples $\hat{c}(\frac{k}{2m+1})$ by iterative use of formula (64). We use the steepest descent algorithm for updating the frequency domain samples. The updating formula is given by

$$\underline{\hat{c}}_{j+1} = \underline{\hat{c}}_j + 2\alpha \mathbf{A}^T \underline{e}_j \tag{68}$$

where $\underline{\hat{c}}_j$ is the j^{th} approximation to $\underline{\hat{c}}$ and α is the step size of updating. The error vector $\underline{e}_j$ is defined as

$$\underline{e}_j = \underline{x} - \underline{x}_j \tag{69}$$

where $\underline{x}_j$ denotes the synthesized time domain signal using the AFT filter (64). That is,

$$\underline{x}_j = \mathbf{A}\underline{\hat{c}}_j \tag{70}$$

If we start with a zero vector $\underline{\hat{c}}_0 = \underline{0}$ and choose the step size α in the range

$$0 < \alpha < \frac{1}{\sigma_{max}^2} \tag{71}$$

where σ_{max} is the maximum singular value of the matrix $\mathbf{A}$, the process (68) will converge and give the minimum norm solution $\underline{\hat{c}}_m$ of the equation (64)

$$\underline{\hat{c}}_m = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\underline{x} = \mathbf{M}\underline{x}, \tag{72}$$

where $\mathbf{M}$ is the minimum norm solution matrix $\mathbf{M} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}$.

The frequency domain samples $\tilde{c}(\frac{k}{2m+1})$ defined by the cosine transform and the time domain samples x_n are related by

$$\tilde{c}(\frac{k}{2m+1}) = \sum_{n=0}^{N-1} x_n cos(\frac{2n+1}{2m+1} 2k\pi) \tag{73}$$

$$with \quad m = 0, 1, \cdots, N-1; \quad k = 0, 1, \cdots, 2m.$$

Writing this in matrix form, we have

$$\underline{\tilde{c}} = \mathbf{C}\underline{\mathbf{x}} \tag{74}$$

where $\mathbf{C}$ is the cosine transform matrix whose n^{th} column ($n = 0, 1, \cdots, N-1$) is the cosine function $cos((n+\frac{1}{2})\omega)$ sampled at values of $\omega = \frac{4k\pi}{2m+1}$ for $m = 0, \cdots, N-1$; $k = 0, \cdots, 2m$. In general, the cosine transform matrix $\mathbf{C}$ is not equal to the minimum norm matrix $\mathbf{M}$. We use $\mathbf{D}$ to represent the difference matrix between the cosine transform matrix $\mathbf{C}$ and the minimum norm matrix $\mathbf{M}$, namely,

$$\mathbf{D} = \mathbf{C} - \mathbf{M}. \tag{75}$$

If we start with the initial vector

$$\underline{\hat{c}}_0 = \mathbf{D}\underline{\mathbf{x}} \tag{76}$$

the convergence property of th updating process (28) will not be affected and the process will converge to the cosine transform solution (74)

$$\lim_{j \to \infty} \underline{\hat{c}}_j = \underline{\tilde{c}}. \tag{77}$$

In this algorithm, the computation of synthesizing the signal $\underline{\mathbf{x}}_j = \mathbf{A}\underline{\hat{c}}_j$ can be easily achieved by the AFT. Using the permuted difference coefficient (PDC) structure [13], the number of multiplications in the computations of $\mathbf{A}^T\underline{\mathbf{e}}_j$ and of the initial vector $\underline{\hat{c}}_0$ can be effectively reduced. The PDC structure can be implemented with random access memory (RAM) and read-only memory (ROM). Therefore, the arithmetic computations of this iterative method also has high degree of parallelism and the resulting architecture is regular. As a result of this iterative use of the AFT, the problem of dense, non-equally spaced time domain data samples has been overcomed. The dense frequency domain samples of cosine transform are obtained without any interpolation or zero-padding.

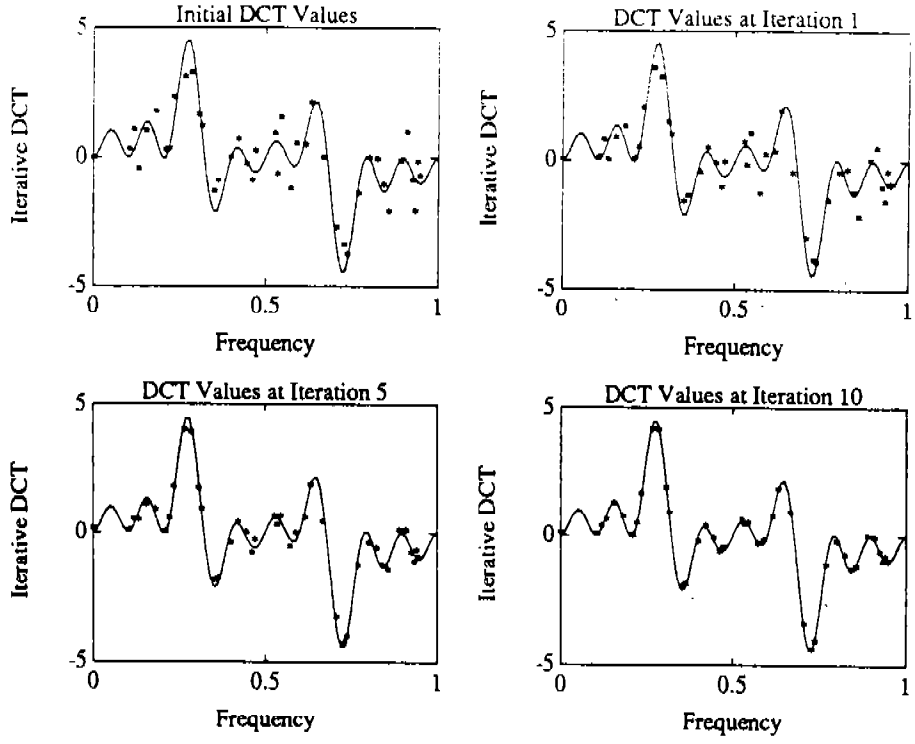


Fig. 5. DCT Spectrum and AFT Spectrums

As an example of computation of discrete cosine transform using iterative AFT, we calculate the spectrum of a signal $x_n = cos(2\pi fn)$ with $f = 0.2$ and $n = 0, 1, \cdots, 4$. By "spectrum", we mean the values of the elements of a transform domain vector, such as $\underline{\tilde{c}}$ or $\underline{\hat{c}}_j$. The DCT spectrum is obtained by direct computation of (73) and is shown in the dashed lines in these figures. The spectrums of iterative AFT at different iterations are shown in solid lines. The squared error norms are shown in Fig. 6 as functions of iteration number j. The squared time domain error norm is defined by

$$E_t(j) = \sum_{n=0}^{N-1} |x_n - x_{j,n}|^2. \tag{78}$$

The corresponding squared frequency domain error norm is defined by

$$E_f(j) = \sum_{k,m} |\hat{c}_j(\frac{k}{2m+1}) - \tilde{c}(\frac{k}{2m+1})|^2 \tag{79}$$

where frequency points are at the odd-number denominator Farey fractions $\frac{k}{2m+1}$. We can see that after several iterations, the iterative AFT quickly reduces the squared error norms and the resulting spectrums converge to the DCT spectrum.

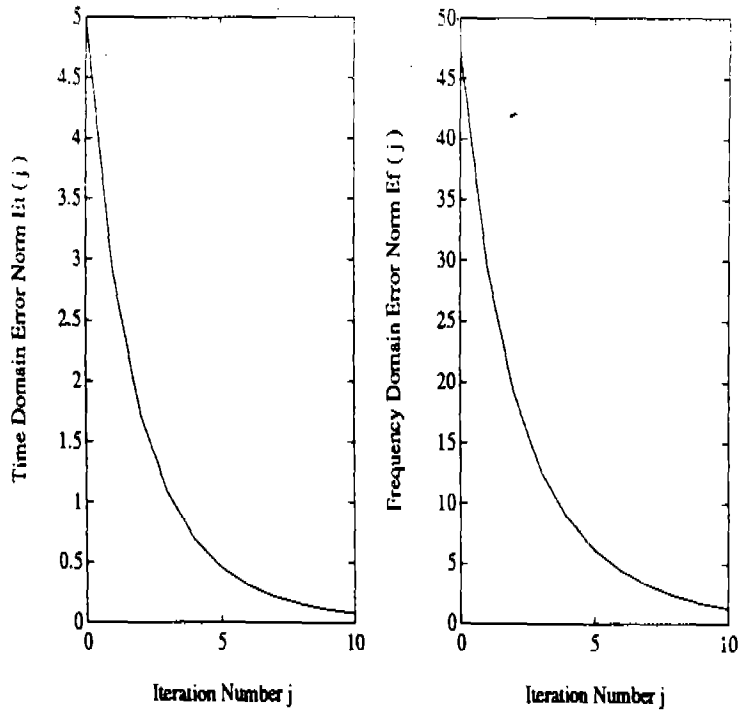


Fig. 6. Squared Error Norms of Iteration Process

The arithmetic fourier transform has been extended for 2-D applications by Tufts, Fan and Cao [7]. Two methods of computing the 2-D discrete cosine transform using AFT have been developed [33]. The first method uses the 2-D AFT to implement a simple computation of the 2-D DCT and dense samples are required. The second method is based on the iterative AFT. It overcomes the difficulty of dense, Farey-fraction sampling in the image-domain and could be used for cases in which (a) the function to be analyzed can only be sampled uniformly and at a rate close to the Nyquist rate or (b) dense transform-domain samples are needed. The 2-D inverse discrete cosine transform can be very efficiently computed from these dense, Farey-fraction transform-domain samples using the 2-D AFT. Therefore, this iterative method could be used with the AFT to form a transform and inverse transform pair and to efficiently perform the 2-D discrete cosine transform and the 2-D inverse discrete cosine transform.

CONCLUSIONS. An iterative arithmetic Fourier transform method is proposed in this paper. This method overcomes the problem of dense time-domain sampling in the original AFT and preserves its advantages for VLSI implementation and fast computation. This iterative AFT could be used with the AFT [1] in certain applications to reduce computation and efficiently perform Fourier analysis.

The application of the iterative AFT algorithm for the computation of the Discrete Cosine Transform is also presented. Further research work on using partial information about the phases or modulus of the transform (such as in problems of reconstructing a signal from the phases or modulus of its Fourer transform) to improve the convergence could be pursued.

ACKNOWLEDGEMENTS The authors thank Abhijit A. Shah for his help in the preparation of the manuscript.

References

[1] D. W. Tufts and G. Sadasiv, "The Arithmetic Fourier Transform," IEEE ASSP Magazine, Vol. 5, pp. 13-17, Jan. 1988.

[2] H. Bruns, "Grundlinien des Wissenschaftlichnen rechnens," Leipzig, 1903.

[3] Aurel Wintner, "An Arithmetical Approach to Ordinary Fourier Series," Baltimore, 1945.

[4] I. S. Reed, D. W. Tufts, T. K. Tuong, N. T. Sin, Xiaowei Yin and Xiaoli Yu, "Fourier Analysis and Signal Processing by Use of the Möbius Inversion Formula," IEEE Trans. on ASSP., Vol. 38, pp. 458-470, March 1990.

[5] M. T. Shih, I. S. Reed, T. K. Truong, E. Hendon, and D. W. Tufts, "A VLSI Architecture for Simplified Arithmetic Fourier Transform Algorithm," Submitted for Publication.

[6] G. F. Boudreaux-Bartels, D. W. Tufts, P. Dhir, G. Sadasiv and G. Fischer, "Analysis of Errors in the Computation of Fourier Coefficients Using the Arithmetic Fourier Transform (AFT) and Summation by Parts (SBP)," ICASSP, 1989.

[7] D. W. Tufts , Z. Fan and Z. Cao, "Image Processing and the Arithmetic Fourier Transform," SPIE Vol. 1058, High Speed Computing II, pp. 46-53, Jan. 15-20, 1989.

[8] G. Fischer, D. W. Tufts and G. Sadasiv, "VLSI Implementation of the Arithmetic Fourier Transform (AFT): A New Approach to High-Speed Computation for Signal Processing," IEEE ASSP Workshop on VLSI Signal Processing, Nov. 1988, Monterey, California.

[9] D. W. Tufts and G. Sadasiv, "Arithmetic Fourier Transform and Adaptive Delta Modulation: A Symbiosis for High Speed Computation," SPIE Vol. 880 High Speed Computing 1988, pp. 168-178.

[10] Weiping Li, "Fourier Analysis Using Adaptive AFT," Proceedings of ICASSP 1990, D7.8, pp. 1523-1526.

[11] Bernard Widrow, Philippe Baudrenhien, Martin Vetterli and Paul F. Titchener, "Fundamental Relations Between the LMS Algorithm and the DFT," IEEE Trans. on Circuits and Systems, Vol. CAS-34, No. 7, pp. 814-820, July 1987.

[12] D. W. Tufts, "A Note on the Computational Complexity of the Arithmetic Fourier Transform," IEEE Transactions on ASSP, Vol. 37, No. 7, pp. 1147-1148, July 1989.

[13] Kenji Nakayama, "Permuted Difference Coefficient Realization of FIR Digital Filters," IEEE Trans. on ASSP, Vol. ASSP-30, No. 2, April 1982, pp. 269-278.

[14] J. R. Fienup, "Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint," J. Opt. Soc. Am. Vol. 4, No. 1, pp. 118-123, Jan. 1987.

[15] H. A. Hauptman, "The phase problem of X-ray crystallography," Physics Today, pp. 24-29, Nov. 1989.

[16] J. R. Fienup and C. C. Wackerman, "Phase-retrieval stagnation problems and solutions," J. Opt. Soc. Am. A. Vol. 3, No. 11, pp. 1897-1907, Nov. 1986.

[17] J. S. Lim and N. A. Malik, "A new algorithm for two-dimensional maximum entropy power spectrum estimation," IEEE Trans. on ASSP., Vol. 29, No. 3, pp. 401-413, June 1981.

[18] M. P. Ekstrom, R. E. Twogood, and J. W. Woods, "Two-dimensional recursive filter design - a spectral factorization approach," IEEE Trans. on ASSP. Vol. 28, pp. 16-26, Feb. 1980.

[19] M. R. Schroeder,*Number Theory in Science and Communication*, 2^{nd} edition, Springer-Verlag, pp. 77, 1986.

[20] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Washington, DC, National Bareau of Standards, 1968.

[21] J. M. McCool and B. Widrow, "Principles and Applications of Adaptive Filters: A Tutorial Review," IEEE ISCAS 1980, pp. 1143-1157.

[22] D. W. Tufts, D. W. Rorabacher, and W. E. Mosier, "Designing simple effective digital filters," IEEE Trans. on Audio and Electroacoustics, Vol. 18, No. 2, pp. 142-158, June 1970.

[23] G. F. Boudreaux-Bartels and T. W. Parks, "Discrete Fourier transform using summation by parts," ICASSP-87, Dallas, TX, April 6-9, 1987.

[24] J. Lee and Y. Chen, "A new method for the design of two-dimensional recursive digital filters," IEEE Trans. ASSP. Vol. 36, No. 4, pp. 589-598, April 1988.

[25] J. Westlake, *A Handbook of Matrix Inversion and Solution of Linear Equations*, Wiley, New York, 1968.

[26] K. S. Miller, "An Introduction to the Calculus of Finite Differences and Difference Equations," Henry Hort and Co., New York, 1960.

[27] J. H. Rosenbaum and G. F. Boudreaux-Bartels, "Rapid Convergence of Some Seismic Processing Algorithms," Geophysics, Vol. 46, No. 12, pp. 1667-1672, Dec. 1981.

[28] N. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing*, Springer-Verlag, New York, 1975.

[29] Massih Hamidi and Judea Pearl, "Comparison of the Cosine and Fourier Transform of Markov-1 Signals," IEEE Trans. ASSP., pp. 428-429, October 1976.

[30] W. H. Chen, "A fast computational algorithm for the discrete cosine transform," IEEE Trans. Commun., Vol. 25, pp. 1004-1009, Sept. 1977.

[31] M. T. Sun, L. Wu and M. L. Liou, "A concurrent architecture for VLSI implementation of discrete cosine transform," IEEE Trans. Circuits and Systems, Vol. 34, pp. 992-994, August 1987.

[32] D. W. Tufts and H. Chen, "Iterative Realization of the Arithmetic Fourier Transform," submitted to IEEE Transactions on ASSP, Sept. 1990.

[33] H. Chen and D. W. Tufts, "Computation of the 2-D Discrete Cosine Transform using the 2-D AFT and 2-D Iterative AFT," submitted to IEEE Transactions on Circuits and Systems for Video Technology, Jan. 1991.

[34] Weiping Li, "Data Compression Using the Discrete-Time Cosine Transform and Möbius Inversion," Submitted to ICASSP 1991.

Combinatorial Aspects of the Hilbert Scheme

Alyson A. Reeves *

Department of Mathematics, Cornell University, Ithaca, NY 14853

ABSTRACT.

The Hilbert Scheme is a fundamental object of study in Algebraic Geometry, as it parametrizes all algebras of the form $k[x_0, \ldots, x_n]/I$ having certain properties in common. In this paper I give a description of such algebras, what properties algebras on the same Hilbert Scheme have in common, and some general properties of the Hilbert Scheme itself. I also indicate how computers become involved in solving these problems for particular algebras and particular Hilbert Schemes.

Throughout, let k be a field of characteristic 0, for example $k = \mathbf{C}$, the complex numbers. Consider the polynomial ring $k[x]$. We can consider this as a homogeneous (meaning all monomials in a given polynomial are of the same degree), graded k-algebra, graded by degrees. We then write $k[x] = \bigoplus_{i=0}^{\infty} kx^i$. If we let $f(t) = dim_k(k[x])_t$, then $f(i) = 1$ for all i. f is the Hilbert function of $k[x]$. For $A = k[x, y] = \bigoplus_{i=0}^{\infty} A_i$, $A_i = kx^i \oplus kx^{i-1}y \oplus \cdots \oplus ky^i$ has dimension $i + 1$, so $f(t) = t + 1$ for $k[x, y]$. In general, for a graded k-algebra A, the **Hilbert function** is defined to be $f(t) = dim_k A_t$.

Let $I = \{$collection of polynomials$\}$ whose zero set we would like to study. For instance, suppose $I = (x^2 - xy)$. We can compute $f_{(A/I)}(t)$, where $A/I = k[x, y]/(x^2 - xy)$, by noting that $x^2 - xy = 0$. This allows us to replace xy with x^2, cutting down on the number of generators in each dimension. Thus $A_i = kx^i \oplus ky^i$ for $i > 0$, and $A_0 = k$ (as always). Hence,

$$f_{(A/I)}(t) = \begin{cases} 1 & \text{for } t = 0 \\ 2 & \text{else} \end{cases}.$$

$A/I = k[x, y]/(x^2)$ has

$$f_{(A/I)}(t) = \begin{cases} 1 & \text{for } t = 0 \\ 2 & \text{else} \end{cases},$$

as well.

Note that in the case of $k[x]$, $f(t) = 1$ is a polynomial, as is $f(t) = t + 1$ for $k[x, y]$. In the last two cases, $f(t)$ can be expressed as the polynomial $f(t) = 2$ for $t > 0$. In fact,

* *Partly supported by the U.S. Army Research Office through ACSyAM, MSI of Cornell University.*

the Hilbert function is always expressible as a polynomial $p(t)$ for large enough t. The polynomial $p(t)$ is called the **Hilbert polynomial.**

For those familiar with projective geometry, $k[x]$ corresponds to projective 0-space, $k[x,y]$ corresponds to the projective line, and $k[x,y]/(x^2 - xy)$ and $k[x,y]/(x^2)$ correspond to two points on the projective line. For these examples, this information is easily obtained from the given algebras, but for more complicated examples, the only method available to determine the dimension and degree of the zeros of a collection of polynomials is to compute the Hilbert polynomial, and then read off this information from the leading term of the polynomial. Fortunately, the Hilbert polynomial can be computed on a computer using, say, the program Macaulay.

Note: Two algebras having the same Hilbert polynomial define zero sets whose dimensions and degrees are the same.

Question: Can we classify all algebras A/I (where $A = k[x_0, \ldots, x_n]$) having the same Hilbert polynomial?

The answer is yes, and there are, in fact, a variety of ways to do so, one of which is by means of the Hilbert scheme $Hilb^{p(z)}_{\mathbf{P}^n}$ (see [1]) , where $p(z)$ is the Hilbert polynomial and $\mathbf{P}^n$ corresponds to $k[x_0, \ldots, x_n]$. Each point of the Hilbert scheme corresponds to a particular algebra A/I with Hilbert polynomial $p(z)$, and $A = k[x_0, \ldots, x_n]$.

For $p(z) = \binom{n+z}{z}$ (for example, for $k[x]$, $n = 0$ and $p(z) = 1$, and for $k[x,y]$, $n = 1$, and $p(z) = z + 1$), the Hilbert scheme $Hilb^{p(z)}_{\mathbf{P}^n}$ is a single point. Letting $n = 1$ and $p(z) = 2$, we saw above that $k[x_0, x_1]/(x_0^2 - x_0x_1)$ and $k[x_0, x_1]/(x_0^2)$ are two algebras having $p(z) = 2$. These correspond to the two pts. $\{(x_0 - 0x_1), (x_0 - x_1)\}$ and $\{(x_0 - 0x_1), (x_0 - 0x_1)\}$ respectively. If we look at all possible sets of two points $\{(x_0 - ax_1), (x_0 - bx_1)\}$ for all possible a, b, we see that each set of two points can be described by the two values a and b. Letting a and/or b take on the value of infinity, we see that the possible sets of two points are parametrized by $\mathbf{P}^2$, the projective plane. In fact, for n points in $\mathbf{P}^1$, $Hilb^n_{\mathbf{P}^1} \cong \mathbf{P}^n$. In general, however, the scheme will be far more complicated. In particular, it may not consist of just a single component.

The problem of determining exactly what the Hilbert scheme looks like for general projective n-space, and general Hilbert polynomials is quite difficult in practice. More tractable are problems involving the determination of certain of its characteristics. For example,

Problem: Determine the component structure of the Hilbert scheme, that is, the number of *irreducible* components, their dimensions, their intersections and, if possible, a general description of the types of algebras (zero sets) on each one.

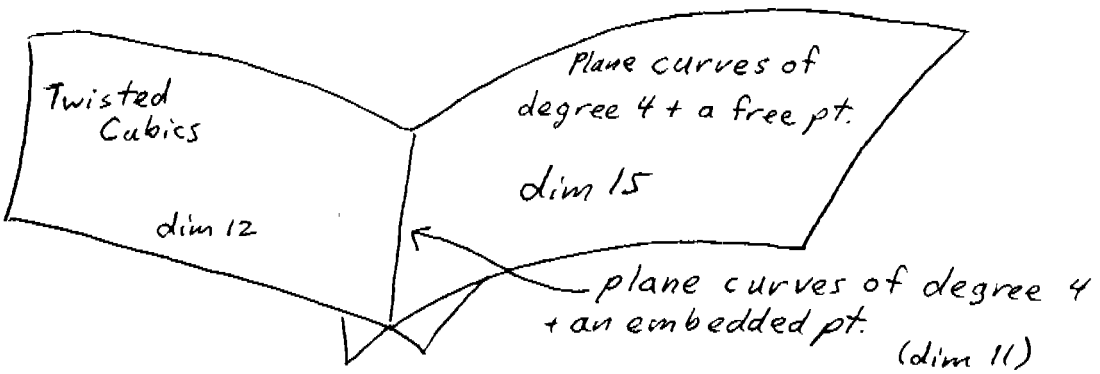
For the Hilbert scheme $Hilb^{3t+1}_{\mathbf{P}^3}$ corresponding to certain curves of degree three in projective 3-space, we can get a good idea of the algebras (curves) it parametrizes by noting:

1) It has two components, one of dimension 12, one of dimension 15.

2) These components intersect and their intersection is of dimension 11.

3) After a change of coordinates, each point on the component of dimension 12 corresponds to a twisted cubic curve, i.e. parametrically, the image of the map $t \mapsto (t, t^2, t^3)$.

4) Similarly we can describe the points on the component of dimension 15 as being plane curves of degree 3 with a point in $\mathbf{P}^3$.



Although it is generally believed that anything bad that can occur on a scheme will occur on *some* Hilbert scheme, the following facts hold for all Hilbert schemes:

1) Every Hilbert scheme is a closed subspace of a Grassmanian.

2) Every Hilbert scheme is connected.

3) No Hilbert scheme "sprawls" too much, i.e. the maximum "distance" (measured in components) between two components is directly related to the dimension of the sets of zeros being parametrized.

This last fact was discovered computationally with the help of the program Macaulay, and it is one of the main results in my dissertation. Though its truth was ascertained by examples, it is a purely theoretical result and has a purely theoretical proof. Likewise, Macaulay, a computer algebra system for algebraic geometry written by Dave Bayer and Mike Stillman, employs Gröbner basis theory to compute many of the structures found in Algebraic Geometry and Computational Algebra. This fascinating interplay between theory and computation, made possible by Gröbner basis theory and programs like Macaulay, has revitalized the field of Computational Algebraic Geometry, and is certain to be a rich source of examples and results for many years to come.

Reference

1. E. Sernesi, *Topics on Families of Projective Schemes*, **Queen's Papers in Pure and Applied Mathematics, No. 73,** Queen's University, Kingston, Ontario, Canada, K7L 3N6, 1986

USING GROEBNER BASES TO DETERMINE THE NATURE OF FIELD EXTENSIONS*

Moss E Sweedler
ACSyAM, MSI
Cornell University
Ithaca NY 14853

ABSTRACT. Suppose the field of fractions of a polynomial ring modulo a prime ideal contains an element c and a finitely generated subfield K. Groebner basis techniques are presented which determine if c is algebraic or transcendental over K. If c is algebraic over K, a minimal polynomial for c over K is found. The minimal polynomial tells whether c lies in K. What makes everything work is the reduction to questions about finitely generated algebras and the use of Buchberger theory with tag variables.

INTRODUCTION. Frequently, fields arise as fields of fractions of integral domain quotients of polynomial rings. Suppose the polynomial ring is $k[X_1,...,X_n]$, sometimes denoted $k[X]$. k is a field. Let $I = \langle f_1,...,f_m \rangle$ be a prime ideal in $k[X]$ and let L be the quotient ring $k[X]/I$. We shall indicate "field of fractions" of an integral domain by putting parentheses around the integral domain. E.g. (L) denotes the field of fractions of L. Elements of (L) can be written as fractions $\underline{a}/\underline{b}$ where $\underline{b}$ is non-zero. Here $a,b \in k[X]$ and we underline to indicate the image of "a" and "b" in $k[X]/I$. Let B be a subalgebra of (L) which is generated by $\underline{c}_1/\underline{d}_1,...,\underline{c}_s/\underline{d}_s$ where the $1/\underline{d}_i$'s are non-zero. Consider the questions:

1. Is $\underline{a}/\underline{b}$ algebraic over (B)?
2. If so, find a minimal polynomial for $\underline{a}/\underline{b}$ over (B)?
3. Is (L) an algebraic extension of (B)?
4. If so, find the index: [(L) , (B)] ?
5. If not, find the transcendence degree of (L) over (B) .

The bare-bones algorithms for solving these problems are presented. The answers to questions three through five are not simply iterations of the techniques used to answer questions one and two. In particular, only one Groebner basis calculation is needed to answer questions three through five. Further explanation and verification that the algorithms are correct will appear elsewhere.

CREDITS. The results described herein are a natural outgrowth and extension of [Shannon87]. Tag variables play a role here which builds on their role in [Shannon87] and [Shannon88] and is complimentary to the seminal role of tag variables in [Spear77]. This paper deals, in part, with transcendence degree which is related to dimension. See [Kredel88] for other work on dimension using Buchberger theory.

*Supported by the U.S. Army Research Office.

ALGORITHM I. Introduce additional variables: $Y, Z_1,...,Z_s, S, T_1,...,T_s$ and form the polynomial ring: $k[X_1,...,X_n,Y,Z_1,...,Z_s,S,T_1,...,T_s]$, which may be abbreviated: $k[X,Y,Z,S,T]$. Choose any term order on $k[X,Y,Z,S,T]$ with the properties:

* Each X_i, Y and Z_i is greater than any monomial in $k[S,T]$.
* S is greater than any monomial in $k[T]$.

S and the T_i's are tag variables because they tag their image under the ring map: $\pi : k[X,Y,Z,S,T] \to (L)$ determined by:

$$X_i \to \underline{X}_i, \quad Y \to 1/\underline{b}, \quad Z_i \to 1/\underline{d}_i, \quad S \to \underline{a}/\underline{b}, \quad T_i \to \underline{c}_i/\underline{d}_i$$

Alternatively π is described by:

$$h(X,Y,Z,S,T) \to h(\underline{X}_1,...,\underline{X}_n,1/\underline{b},1/\underline{d}_1,...,1/\underline{d}_s,\underline{a}/\underline{b},\underline{c}_1/\underline{d}_1,...,\underline{c}_s/\underline{d}_s)$$

With respect to the term order, find a Groebner basis G for Ker π. This may be done by the Buchberger algorithm starting with the generating set for Ker π:

$$\{ f_i \} \cup \{ bY - 1 \} \cup \{ d_iZ_i - 1 \} \cup \{ a - bS \} \cup \{ c_i - d_iT_i \}$$

Let G_T denote $G \cap k[T]$. Let G_S denote the subset of $G \cap k[S,T]$ consisting of polynomials whose lead term is not divisible by the lead term of a polynomial in G_T. The first two questions can now be answered.

$\underline{a}/\underline{b}$ is transcendental (B) if and only if G_S is empty. If G_S is not empty, choose $h(S,T)$ in G_S of minimal S degree. $h(S,\underline{c}_1/\underline{d}_1,...,\underline{c}_s/\underline{d}_s)$ considered as a polynomial in (B)[S] is a minimal polynomial for $\underline{a}/\underline{b}$ over (B). Hence, the S degree of h equals the index: $[(B)[\underline{a}/\underline{b}], (B)]$.

We go into no details beyond the following. Since (L) is generated *as a field* by the $\underline{X}_i$'s, the images of the tag variables, $\underline{a}/\underline{b}$ and the $\underline{c}_i/\underline{d}_i$'s, can be expressed as *rational functions* in the $\underline{X}_i$'s. Additional main variables Y and the Z_i's and their images have been selected so that the images of the tag variables can be expressed as polynomials - not just rational functions - in the images of the main variables.

ALGORITHM II. Introduce additional variables: $X_{n+1},...,X_{n+s}, T_1,...,T_s$ and form the polynomial ring: $k[X_1,...,X_n,X_{n+1},...,X_{n+s},T_1,...,T_s]$, which may be abbreviated: $k[X,T]$. Choose any term order on $k[X,T]$ with the property:

* Each X_i is greater than any monomial in $k[X_{i+1},...,X_{n+s},T]$.

The T_i's are tag variables because they tag their image under the ring map: $\pi : k[X,T] \to (L)$, determined by: $X_i \to \underline{X}_i$, $X_{n+i} \to 1/\underline{d}_i$, $T_i \to \underline{c}_i/\underline{d}_i$. Alternatively π is described by: $h(X,T) \to h(\underline{X}_1,...,\underline{X}_n,1/\underline{d}_1,...,1/\underline{d}_s,\underline{c}_1/\underline{d}_1,...,\underline{c}_s/\underline{d}_s)$. With respect to the term order, find a Groebner basis G for Ker π. This may be done by the Buchberger algorithm starting with the following generating set for Ker π:
$\{ f_i \} \cup \{ d_iX_{n+i} - 1 \} \cup \{ c_i - d_iT_i \}$. Let G_{n+s} denote the subset of $G \cap k[X_{n+s},T]$ consisting of polynomials whose lead term is not divisible by the lead term of polynomials in $G \cap k[T]$. Similarly, for $1 \le i < n + s$, let G_i denote the subset of $G \cap k[X_i,...,X_{n+s},T]$ consisting of polynomials whose lead term is not divisible by the

lead term of polynomials in $G \cap k[X_{i+1},...,X_{n+s},T]$. For $i = 1,...,n+s$ if G_i is not empty, choose $h_i(X_i,...,X_{n+s},T)$ in G_i of minimal X_i degree. Let E_i be this minimal X_i degree of h_i. The X_i's play pivotal main variable / tag variable roles. X_i is a main variable with respect to G_j with $i < j$ and X_i is a tag variable with respect to G_j with $j \leq i$. Questions three through five can now be answered.

If all the G_i's are non-empty then (L) is algebraic over (B). In this case the index: $[(L),(B)]$ equals the product of the E_i's. If not, (L) is transcendental over (B) of transcendence degree equal to the number of empty G_i's.

REFERENCES.

Buchberger, B. (1965). An algorithm for finding a basis for the residue class ring of a zero-dimensional polynomial ideal, Dissertation, Universitaet Innsbruck, Institut fuer Mathematik.

Buchberger, B. (1970). An algorithmic criterion for the solvability of algebraic systems of equations. Aequationes Mathematicae 4/3, 374-383.

Buchberger, B. (1976). A theoretical basis for the reduction of polynomials to canonical forms. ACM Sigsam Bull. 10/3 19-29 1976 & ACM Sigsam Bull. 10/4, 19-24.

Buchberger, B. (1979). A criterion for detecting unnecessary reductions in the construction of Groebner bases. Proc. of EUROSAM 79, Lect. Notes in Computer Science 72, Springer, 3-21.

Buchberger, B. (1984). A critical-pair/completion algorithm for finitely generated ideals in rings. Decision Problems and Complexity. (Proc. of the Symposium "Rekursive Kombinatorik", Muenster, 1983.) E. Boerger, G. Hasenjaeger, D. Roedding, eds. Springer Lecture Notes in Computer Science, 171, page 137.

Buchberger, B. (1985). Groebner bases: an algorithmic method in polynomial ideal theory. Multidimensional Systems Theory. N. K. Boese ed. D. Reidel Pub Co., 184-232.

Kredel, H. and Weispfenning, V. (1988). Computing dimension and independent sets for polynomial ideals. Special Volume of the JSC on the computational aspects of commutative algebra. Vol. 6, 1988.

Shannon, D. and Sweedler, M. (1988). Using Groebner bases to determine algebra membership, split surjective algebra homomorphisms and determine birational equivalence. J. Symbolic Computation, 6, 267-273.

Shannon, D. and Sweedler, M. (1987). Using Groebner bases to determine the algebraic or transcendental nature of field extensions within the field of rational functions. Preprint.

Spear, D. (1977). A constructive approach to commutative ring theory. Proceedings 1977 MACSYMA User's Conference, 369-376.

ANALYTIC SOLUTION OF THE PERIOD FOUR QUADRATIC RECURSION POLYNOMIAL

Harry J. Auvermann
U. S. Army Atmospheric Sciences Laboratory
White Sands Missile Range, New Mexico 88002-5501

ABSTRACT. This paper is concerned with stable points of iterates of the function $F(z,d) = d - z^2$. The number of these stable points bifurcates successively as the real parameter d varies from -1/4 to 2. The number of stable points of a particular bifurcation is termed the period. Period one stable points are roots of the polynomial that result from substituting z for F(z,d) in the above. Two applications of F(z,d) produce a fourth order polynomial. Period two stable points are roots of this polynomial that are easily obtained. Four applications of F(z,d) produce a sixteenth order polynomial. Period four stable points are the roots of this polynomial. Four of these roots are known from analysis of the lower iterates. Solution of a twelfth order polynomial then determines the period four stable points. A general analytic solution method to recursion polynomials of this type has been given previously. This paper presents an alternate method for obtaining the analytical closed form expressions for the period four roots as a function of the parameter d.

INTRODUCTION. Transition from order to disorder, similar to the transition of a fluid from laminar to turbulent flow, has been observed in mathematical expressions such as one-dimensional maps, an example being the recursion expression

$$z_{k+1} = d - z_k^2. \tag{1}$$

The parameter d in the mathematical process corresponds to the Reynolds number in the fluid flow process. Corresponding to the random-like samples of the local velocity in the flow are the iterates z_k of the mathematical process. Stable points are repeating numbers in the sequence z_k. The condition where each point is the same stable point is analyzed by substitution of z_k for z_{k+1} on the left hand side of equation (1) and solving for the roots of the resulting polynomial. Bifurcation occurs here in the sense that for larger values of d the stable points repeat every second iteration. This sequence of two is termed a period two limit cycle. The former case is termed a period one limit cycle.

In this paper, attention will be limited to periods one, two, and four limit cycles. However, bifurcation continues to happen as d increases. The values where bifurcation occurs (called thresholds) become closer and closer together. If d is made sufficiently close to some ultimate value, called the accumulation point, an arbitrary high number of stable points make up the

Approved for public release. Distribution unlimited.

limit cycle (Feigenbaum, 1978). Stable points occur in isolated intervals when d is greater than the accumulation point (Berge', 1984, p 202). Between these isolated intervals are intervals of chaos similar to fully developed turbulence. That is, the iteration sequence never repeats itself and the values depend upon the starting point. A noise like iterate sequence from equation (1) in the chaotic regime is shown in figure 1. This similarity between iterate sequences and random processes is the reason for the intense interest in the mathematics of one-dimensional maps and limit cycles. Understanding limit cycles can perhaps be translated into an understanding of the transition of physical systems from order to chaos.

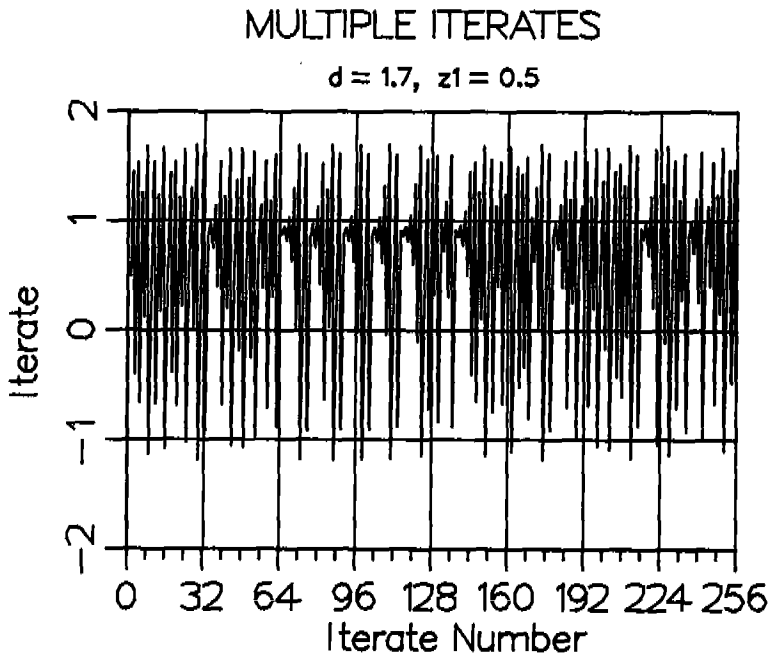


Figure 1. Iterates of the quadratic recursion relation.

Feigenbaum (1983) found that high period limit cycles have two associated universal numbers. Universal means that these numbers do not depend upon the details of the recursion function used, which means that investigation of the two universal numbers can be effected using any member, such as equation (1), of the allowed class of functions if the investigation is carried far enough. Being quadratic (second degree) in the iteration variable z, the phenomenon is referred to as quadratic bifurcation. Stable points and thresholds determined for one quadratic recursion function may be used directly to determine those for another quadratic recursion function by the use of a linear transformation.

The present state of quadratic bifurcation is covered in the literature (Guckenheimer, 1979). Abel (1829) has given a method by which solutions of polynomials of this type can be reduced to the solution of polynomials whose order is the same as the period. Netto (1898) has shown how the method of Lagrange resolvants can be used to solve for period three and period four roots.

The new results reported here are an alternate method for obtaining the roots of a period four quadratic recursion polynomial. This method is simpler but, of course, not general as are the methods of Abel (1829) and Netto (1898). Much of the algebra necessary to show the results presented here has been left out.

The notation to be used is described below. The symbol N has been adopted from previous work (Feigenbaum, 1978, p 50) and given the name bifurcation index. It will be employed as an identifying label for polynomials, stable points, and thresholds. The symbol n is used for the period of a particular bifurcation. The period-index relationship is

$$n = 2^N. \tag{2}$$

In this paper N = 0, 1, 2 is the range of indices considered. The following list contains the essential elements of the remaining notation used.

P(z,N) = stable point polynomial for the Nth bifurcation

H(z,N) = factor of P(z,N) [P(z,N) = P(z,N -1)H(z,N)]

Z(d,N,m) = mth (m = 1, 2,...2^n) stable point of the Nth bifurcation

Q(g) = a polynomial in the variable g

D(N) = threshold of d where the period changes from 2^{N-1} to 2^N

EXPRESSIONS ASSOCIATED WITH THE LOWER PERIODS. In this section, the expressions for the polynomials, stable points, and the thresholds of the lower indices will be developed. The first step is to write out the corresponding stable point polynomials.

For a beginning value of the variable, z_0, and a given parameter, a series of iterates z_k is produced by repeated application of

equation (1). If d is greater than D(0), z_k approaches a fixed point Z(d,0,m) as k increases. This stability occurs when z_{k+1} is equal to z_k in equation (1). The values of z that satisfy this condition are the roots of the index zero polynomial

$$P(z,0) = z^2 + z - d, \tag{3}$$

where the serial number k has been dropped for writing economy. The index 1 polynomial is obtained by developing the expression for the iterate 2 later in the sequence. Hence,

$$P(z,1) = (z^2 - d)^2 + z - d. \tag{4}$$

From equations (3) and (4), one has

$$P(z,1) = P(z,0)[P(z,0) - 2z + 1] = 0; \tag{5}$$

$$P(z,1) = P(z,0)H(z,1). \tag{6}$$

Equations (5) and (6) serve to define H(z,1), which is termed here the index 1 primitive polynomial. The index 2 polynomial is obtained in a similar manner by developing the expression for the iterate four later in the sequence. Hence,

$$P(z,2) = \{[(z^2 - d)^2 - d]^2 - d\}^2 + z - d. \tag{7}$$

From equations (4) and (7),

$$P(z,2) = P(z,1)[P^3(z,1) - 4zP^2(z,1) + 2(3z^2 - d)P(z,1) - 4z(z^2 - d) + 1 = 0; \tag{8}$$

$$P(z,2) = P(z,1)H(z,2). \tag{9}$$

Equations (8) and (9) serve to define H(z,2), the index 2 primitive polynomial, and show that P(z,1) is a factor of P(z,2).

Roots of polynomial P(z,0) of equation (3) are given by

$$Z(d,0,1) = 1/2[-1 + (1 + 4d)^{1/2}]; \tag{10}$$

and

$$Z(d,0,2) = 1/2[-1 - (1 + 4d)^{1/2}]. \quad (11)$$

Z(d,0,1) is a stable fixed point and Z(d,0,2) is an unstable fixed point (Feigenbaum, 1983)). From equation (6) one sees that two roots of the index 1 polynomial P(z,1) are the same as the roots of P(z,0). The remaining index 1 roots are obtained by applying the quadratic formula to H(z,1) from equation (5). These roots are

$$Z(d,1,3) = 1/2[1 + (4d - 3)^{1/2}]; \quad (12)$$

$$Z(d,1,4) = 1/2[1 - (4d - 3)^{1/2}]. \quad (13)$$

The threshold can now be determined. For equations (10) and (11) to be real, the radical must be nonnegative. This condition on d defines the index zero threshold.

$$D(0) = -1/4. \quad (14)$$

The value of d at which equations (12) and (13) become real is the index 1 threshold

$$D(1) = 3/4. \quad (15)$$

The index 2 threshold is

$$D(2) = 5/4. \quad (16)$$

The condition that a radical in the index 2 stable point expressions of the following section be real gives this result.

INDEX 2 RESULTS. In this section the expressions for the roots of the index 2 primitive polynomial H(z,2) will be developed. It is shown in its expanded form in equation (16a).

$$\begin{aligned} H(z,2) = {} & z^{12} - 6dz^{10} - z^9 + (15d^2 - 3d)z^8 + 4dz^7 - (20d^3 - 12d^2 \\ & - 1)z^6 - (6d^2 - 2d)z^5 + (15d^4 - 18d^3 + 3d^2 - 4d)z^4 \\ & + (4d^3 - 4d^2 - 1)z^3 - (6d^5 - 12d^4 + 6d^3 - 5d^2 + d)z^2 \\ & - (d^4 - 2d^3 + d^2 - 2d)z + (d^6 - 3d^5 + 3d^4 - 3d^3 + 2d^2 + 1). \end{aligned} \quad (16a)$$

Figure 2 shows the variation of H(z,2) as a function of z. The first clue to how the roots of H(z,2) are constructed came from following the graph of the polynomial as the parameter is increased. A recent paper (Godwin, 1984) examines this point in detail. Suppose the value of the polynomial is greater than zero for all z of interest when d is less than D(2). As d is increased past D(2), the minima of the period four polynomial approach the zero line, touch this line at the points corresponding to the stable points of the period two polynomial, and then proceed downward so that the graph now crosses zero on either side of the period two stable points. This behavior is illustrated in figure 3. Just after the threshold has been reached, the two new roots are very near each other and are approximately equidistant from the period two stable points.

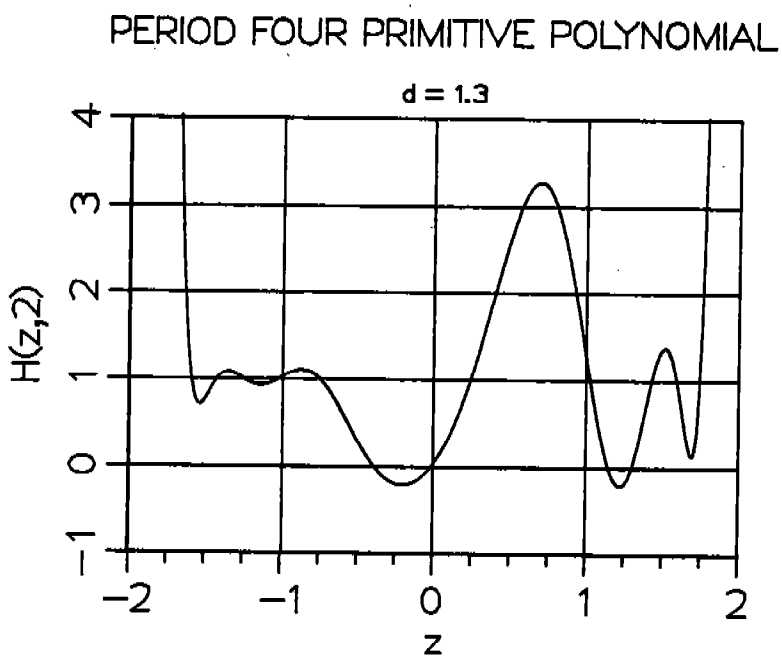


Figure 2. Period Four Polynomial showing the four stable points.

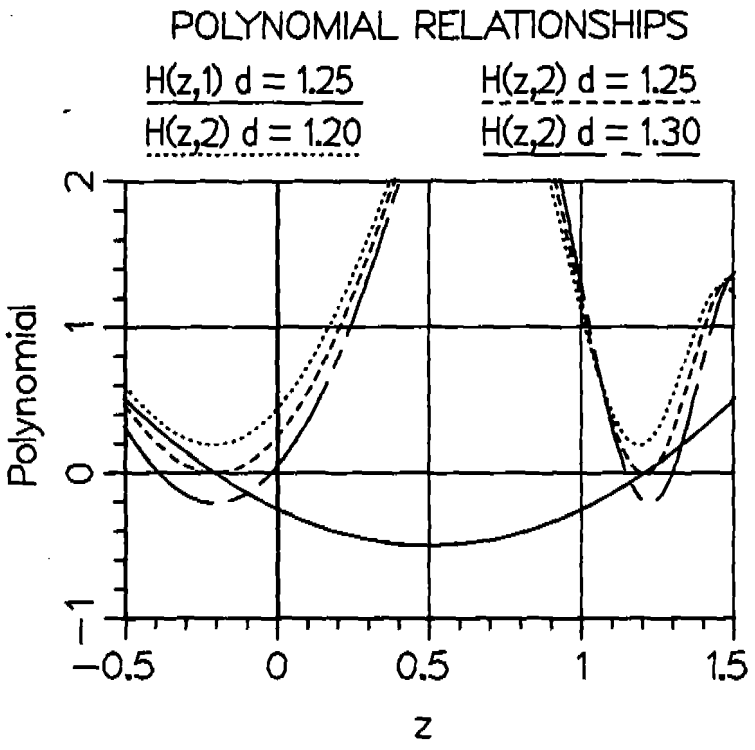


Figure 3. Period two and period four relationship at period four threshold.

The second clue came from observation of the numerical roots of P(z,2). Looking only at the real roots (for d = 1.5) and identifying them as follows

$$R1 = Z(1.5,2,5) = +1.45160150, \tag{17}$$

$$R2 = Z(1.5,2,6) = -0.60714691, \tag{18}$$

$$R3 = Z(1.5,2,7) = +1.13137260, \tag{19}$$

$$R4 = Z(1.5,2,8) = +0.21999598, \tag{20}$$

one observes that

$$(R1 + R3)(R2 + R4) = -1.00000083. \tag{21}$$

A similar result occurred for the other root sets. This similarity suggests that R1 and R3 are related to each other in some way

and that R2 and R4 are related to each other in some way. It also suggests that the two sums are negative reciprocals. Combining the two clues suggested representing the first root set in general by the following

$$Z(d,2,5) = +G + H; \tag{22}$$

$$Z(d,2,6) = -I - J; \tag{23}$$

$$Z(d,2,7) = +G - H; \tag{24}$$

$$Z(d,2,8) = -I + J; \tag{25}$$

where G, H, I, J are essentially another variable set. Applying equation (1) to the right hand sides of equations (22) through (25) results in

$$-I - J = d - G^2 - 2GH - H^2. \tag{26}$$

$$+G - H = d - I^2 - 2IJ - J^2. \tag{27}$$

$$-I + J = d - G^2 + 2GH - H^2. \tag{28}$$

$$+G + H = d - I^2 + 2IJ - J^2. \tag{29}$$

Combining appropriately, one finds that

$$1 = +4GI. \tag{30}$$

This shows that the result suggested numerically in equation (21) is in fact true. Combining another way and substituting G = g/2, one finds

$$Q(g) = g^6 - 4dg^4 - 4g^3 + 4dg^2 - 1 = 0. \tag{31}$$

It is found numerically that +2G and -2I satisfy Q(g). Noting that Q(g) will contain like information for root sets Z(d,2,9) through Z(d,2,12) and for root sets Z(d,2,13) through Z(d,2,16), the following hypothesis is advanced

$$Q(g) = (g + 2I_1)(g - 2G_1)(g + 2I_2)(g - 2G_2)$$
$$(g + 2I_3)(g - 2G_3) = 0. \tag{32}$$

In the above, G_1, I_1 belong to the first set; G_2, I_2 belong to the second set; and G_3,I_3 belong to the third set. Combining monomials by twos and using equation (30) result in

$$Q(g) = (g^2 + 2K_1g - 1)(g^2 + 2K_2g - 1)$$
$$(g^2 + 2K_3g - 1) = 0 \qquad (33)$$

$$K_i = I_i - G_i \qquad i = 1,2,3 \qquad (34)$$

Carrying out the multiplication, comparing coefficients with equation (31), eliminating K_2, K_3, and letting $K_1 = K = I - G$, the result is

$$K^3 + (1/4)(3 - 4d)K + (1/2) = 0. \qquad (35)$$

The solution for equation (35) is known (Abramowitz and Stegun, 1970). The real root is

$$K = (S1 + S2) \qquad (36)$$

$$S1 = \{[1/4][-1 + (1 - (4d/3 - 1)^3/4)^{1/2}]\}^{1/3} \qquad (37)$$

$$S2 = \{[1/4][-1 - (1 - (4d/3 - 1)^3/4)^{1/2}]\}^{1/3} \qquad (38)$$

Only this real root will be written out. Substituting from equation (30) for I in equation (34), we obtain

$$G = [-K + (K^2 + 1)^{1/2}]/2. \qquad (39)$$

The positive radical was used so that evaluation of equation (39) yields the values in equations (17) through (20). Using equations (34) and (26) through (29), one finds

$$I = [+K + (K^2 + 1)^{1/2}]/2 \qquad (40)$$

$$H = (d - G^2 + I)^{1/2} \qquad (41)$$

$$J = (d - I^2 - G)^{1/2} \qquad (42)$$

where the positive radical is to be used. The full expression for the roots may now be obtained from equations (22) through (25). The first one of these is

$$Z(d,2,5) = (1/2)\{[-K + (K^2 + 1)^{1/2}] + [4d - 1 + 2K - 2K^2 + 2(K + 1)(K^2 + 1)^{1/2}]^{1/2}\}$$

K is given by equation (36) and S1 and S2 are given by equations (37) and (38). Roots Z(d,2,6) through Z(d,2,8) are generated by appropriately combining G, H, I, J from equations (39), (40), (41), and (42) according to equations (23) through (25). Roots Z(d,2,9) through Z(d,2,16) are obtained from K_2, K_3 in a like fashion where K_2 is the second root of equation (35) and K_3 is the third root of equation (35) (Abramowitz and Stegun, 1970).

We are now able to show how the index 2 threshold may be obtained from a condition on the index 2 stable point expressions. The condition will relate to some radical passing from imaginary through zero to real as d passes through D(2). Searching the expressions for the relevant radical, we find the expressions for H, J are the proper ones. From equations (41) and (42) we write the conditions for H, J to be zero simultaneously

$$d - G^2 + I = 0 \qquad (44)$$

$$d - I^2 - G = 0 \qquad (45)$$

Subtracting equation (45) from equation (44) and factoring, we get

$$(G + I)[1 - (G - I)] = 0. \qquad (46)$$

For the bracket to be zero, (G - I) must be unity, giving the condition on d in equation (16).

CONCLUDING REMARKS. The author is indebted to Dr. D. M. Giarrusso, then a member of the Mathematical Sciences Institute at Cornell, now at Saint Lawrence University, for the location of the early work (Abel, 1829, and Netto, 1898). The solution method reported in this paper is independent of Lagrange, but of course gives the same expressions for the roots. Attempts to apply the present method to the index 3 polynomial have resulted in some simplification but have not produced root expressions.

REFERENCES

Abel, N. H., 1829, "Mémoire sur une Classe Particulière D'Équations Résolubles Algébriquement," Journal für die Reine and Angewendte Mathematik, Crelle, Berlin.

Abramowitz, M., and I. A. Stegun, ed., 1970, Handbook of Mathematical Functions, National Bureau of Standards AMS 55, U. S. Government Printing Office, Washington, D. C.

Berge', P., Yves Pomeau, and C. Vidal, 1984, Order within Chaos, John Wiley and Sons, New York.

Feigenbaum, M. J., 1978, "Quantitative Universality for a Class of Nonlinear Transformations," J. Stat. Phy., 19:25-52.

Feigenbaum, M. J., 1983, "Universal Behavior in Nonlinear Systems," Nonlinear Dynamics and Turbulence, eds. G.I. Barenblatt, G. Iooss, and D. D. Joseph, Pitman Advanced Publishing Program, Boston, London, and Melbourne.

Godwin, A. N., 1984, "The precise determination of Maxwell sets for cuspoid catastrophes," Int. J. Math. Educ. Sci., Technol., 15:167.

Guckenheimer, J., 1979, "The Bifurcation of Quadratic Functions," Bifurcation Theory and Applications in Scientific Disciplines, Annals of the New York Academy of Sciences, vol 316, eds. O. Gurel and O. E. Rossler, New York Academy of Sciences, New York.

Netto, Eugen, 1898, Verlesungen uber Algebren, Teubner, Leipzig.

Beyond Rolle's Theorem

Bruce Anderson

November 12, 1991

Abstract

The concept and existence of higher-order Rolle's theorems are introduced, and a summary of results for polynomials of degree up to 5 follows. The general topic at hand is what kinds of restrictions there are on the placement of roots of various derivatives. The surprising result is that Rolle's theorem is not the only such restriction.

Introduction. Much recent work has been concerned with the signs of polynomials and their derivatives. For example, Ben-Or-Kozen-Reif [1] construct an algorithm for counting the number of real zeros of a polynomial with a prescribed sign sequence corresponding to the signs of the derivatives of ascending order. Coste-Roy and Sweedler [2] have also pointed to a way to determine, based on the sign sequence of the polynomials at two different points, which point lies to the right of the other. Their algorithm rests on Rolle's theorem, and gives a method for determining whether a list of sign sequences is consistent - i.e. whether it is at all possible for a single polynomial to generate this list of sign sequences.

My work centers on something of a stronger question: Is being consistent, in the way defined by Coste-Roy and Sweedler, the only such restriction on polynomials, and on differentiable functions in general?

Rolle's Theorem. I will phrase this question in a different way: Given the relative position of the zeros of a polynomial and the zeros of successive derivatives, is it possible to find a polynomial which has these zeros and no others?

For example, is it possible to find a polynomial (or even a real differentiable function), with three real roots $a < b < c$, and whose derivative has exactly two real roots between a and b, and none between b and c? (See figure 1)

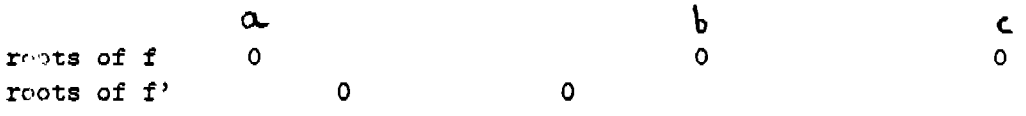


Figure 1

The answer is obviously no, for this would violate the well-known Rolle's theorem, which states that between any two real roots of a differentiable function, its derivative must have a root somewhere in the interval of the original two real roots. (Fig. 2)

Figure 2

We can iterate Rolle's theorem, so that given a sufficiently differentiable function with, say, 5 real roots, its derivative must have at least 4 real roots, one inside each of the 4 intervals defined by the original 5 roots. Likewise, the second derivative must have 3 roots, the third derivative must have 2 roots, the fourth derivative must have at least 1 root, and the fifth derivative is not guaranteed to have any roots as a result of simply iterating Rolle's theorem. (See figure 3).

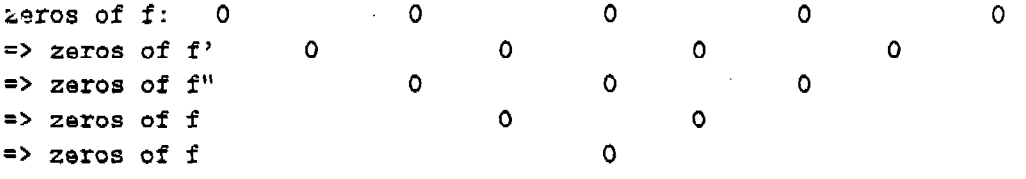


Figure 3

The algorithm of Sweedler and Coste-Roy provides a nice way to determine whether the iterated Rolle's theorem has been violated. However, so long as the iterated Rolle's theorem is satisfied, this method won't point to any inconsistencies. I initially began my research under the belief that iterated Rolle's theorem was the only such restriction on differentiable functions, thinking that if there were others, they would be known by now. Nevertheless, it has emerged that there are other restrictions, in fact quite a few (possibly infinite). I have not yet been able to determine one global restriction or a general pattern of restrictions yet. These restrictions which do not follow from iterated Rolle's theorem may be considered a higher order

Rolle's theorem, because the placement of the roots of various derivatives of f relative to one another guarantees a root of a derivative whose existence is not already guaranteed by iterated Rolle's theorem. The simplest such higher-order Rolle's Theorem I have found is described below.

A Higher Order Rolle's Theorem. Let f be a real function which (on an open interval) is differentiable five times, and has at least five roots. Suppose the first two roots of f" lie to the left of the second and third roots of f, respectively. Suppose the first root of $f^{(3)}$ lies to the right of the second root of f'. Then $f^{(5)}$ must have a root in the open interval between the first and fifth root of f.

The hypothesis of the theorem is illustrated in figure 4.

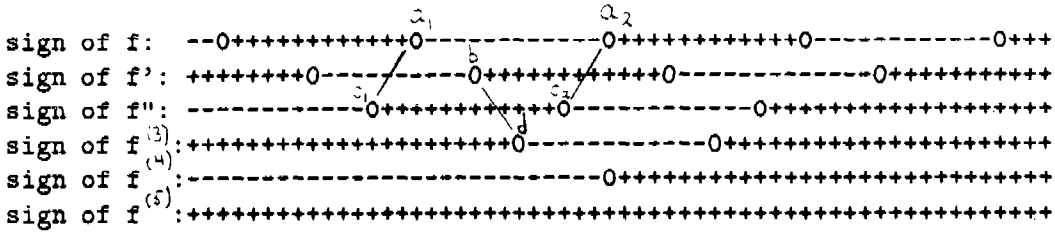


Figure 4

In particular, $c_1 < a_1$, $b < d$, and $c_2 < a_2$ In our higher order Rolle's Theorem, we are guaranteed a root of $f^{(5)}$ somewhere in the interval (a_1, c_2).

The proof of the theorem above is based on Taylor's remainder formula. One expands $f(d+\epsilon)$ and $f(d-\epsilon)$ about d. A bit of symbol pushing leads to the conclusion that d lies closer to a_2 than a_1. In a similar manner, expanding $f^{(2)}(d+\epsilon)$ and $f^{(2)}(d-\epsilon)$ will lead us to conclude that d lies closer to c_1 than c_2. One can check that d can not both lie closer to c_1 than c_2 *and* closer to a_2 than a_1 while insisting that

$$c_1 < a_1 < d < c_2 < a_2$$

as originally assumed. This implies the existence of a zero of $f^{(5)}$.

The discovery of this theorem sprang out of a systematic investigation of the special case of polynomials (as opposed to general differentiable functions). Polynomials are much easier to handle, and discovering a restriction on where a polynomials's roots may lie often leads to a general higher order Rolles theorem such as the one described above. Therefore the rest of our discussion will relate only to polynomials, keeping in mind their importance in generalizing to general differentiable functions.

The Case of Polynomials. Given a polynomial of degree n with n real roots, iterated Rolle's theorem will guarantee that all the roots of all the derivatives will be real, and will specify certain intervals in which they must fall. Again, what other restrictions are there which can not be deduced from Rolle's theorem?

I have been able to characterize all such restrictions on polynomials of degree 2,3,4, and 5. I will go through the cases of degree 2,3, and 4 in some detail, but only summarize the results for degree 5.

Degree 2. The degree 2 case is quite simple. Given a second degree polynomial with two real roots a and b, by Rolle's theorem its first derivative will of course have a root between a and b. (In fact it will lie exactly half way between a and b.) $p(x) = (x-1)(x+1)$ is an example of such a polynomial, so all possibilities satisfying Rolle's theorem (in this case there is only one possibility) are constructable. (See Fig. 5)

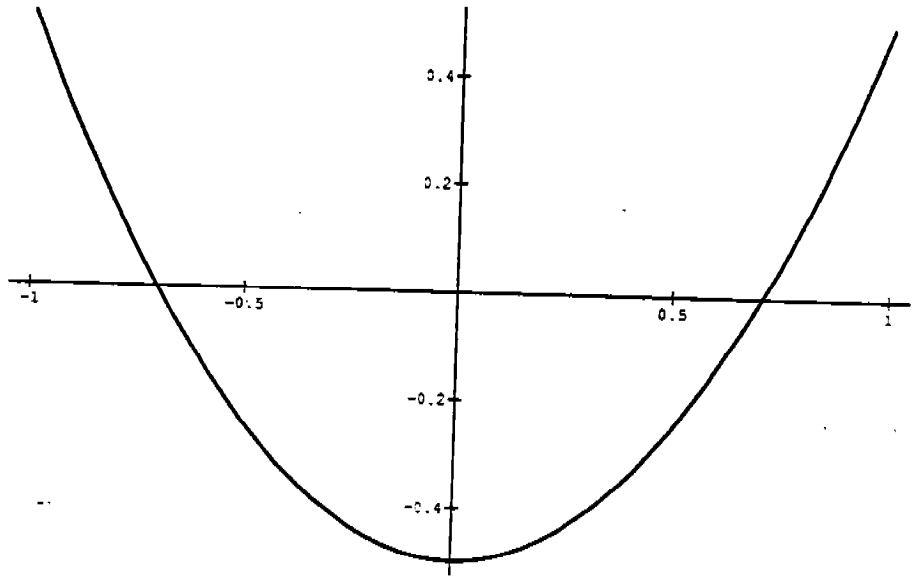


Figure 5

Degree 3. For the degree 3 case there are two possibile arrangements of roots which satisfy Rolle's theorem. Given a cubic polynomial with three real roots a, b, and c, its derivative must have two real roots, one between a and b, the other between b and c. Now the second derivative must have one real root, but it may lie to the left or to the right of b, hence the conclusion

that there are two possibilities. And in fact, these two possibilities are constructable, as indicated in figure 6.

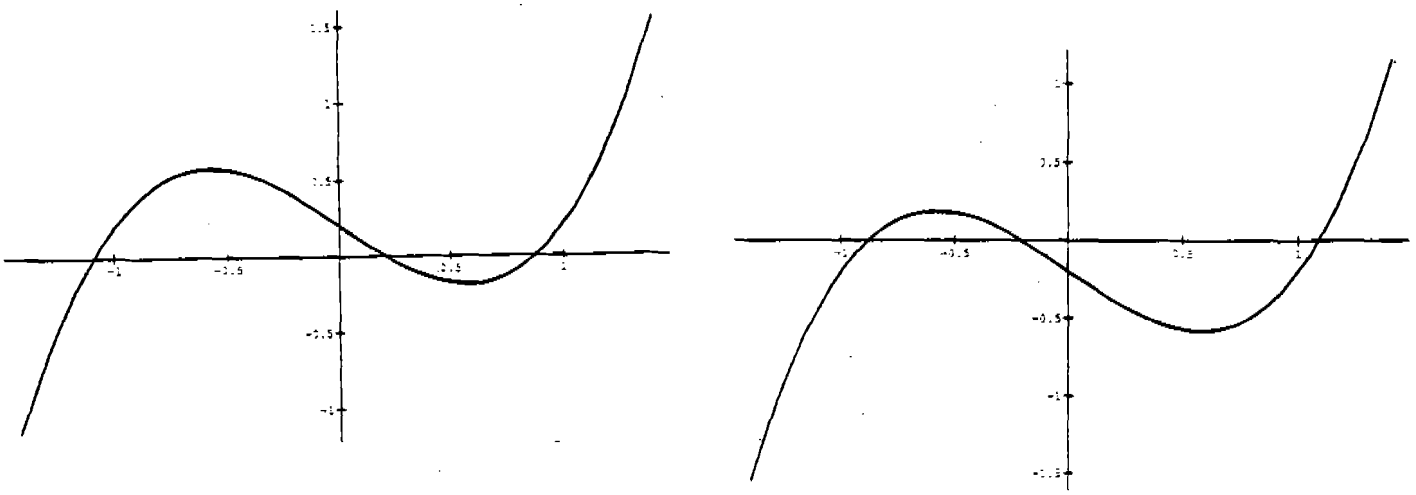


Figure 6

Degree 4. In the fourth degree case, things begin to get somewhat complicated. There turn out to be twelve possible arrangements of the roots which satisfy Rolle's theorem. They are listed in table 1.

{0 1 2 3 0 1 2 0 1 0}
{0 1 2 0 3 1 2 0 1 0}
{0 1 2 0 1 3 2 0 1 0}
{0 1 2 3 0 1 0 2 1 0}
{0 1 2 0 3 1 0 2 1 0}
{0 1 2 0 1 3 0 2 1 0}
{0 1 2 0 1 0 3 2 1 0}
{0 1 0 2 3 1 2 0 1 0}
{0 1 0 2 1 3 2 0 1 0}
{0 1 0 2 3 1 0 2 1 0}
{0 1 0 2 1 3 0 2 1 0}
{0 1 0 2 1 0 3 2 1 0}

Table 1

Here the notation "0102132010" means "first the polynomial has a zero, then the first derivative, then the polynomial, then the second derivative, then the first derivative, etc." Note that between every pair of 0's there is a 1, between every pair of 1's there is a 2, etc. This is simply due to the iterated Rolle's theorem. However, only ten of these twelve possible arrangements are actually constructable; in other words, two arrangements of the roots which satisfy Rolle's theorem can not actually occur in reality. To illustrate why these two arrangements are not achieved, it is helpful to construct the following graph. [1] By "modding out" affine transformations of the polynomial (affine transformations will not affect the relative arrangement of roots), we can begin by assuming that the first and last roots are given by -1 and 1, respectively, so that in order to completely determine the polynomial (up to affine transformations) we need only specify the inner two roots, a and b. Here a and b vary between -1 and 1 and $a < b$. For each particular choice for a and b we can numerically determine which of the twelve "legal" arrangments is achieved. By doing this at every possible choice of a and b (at least within some numerical approximation) we can then construct the "map", which is shown in figure 7.

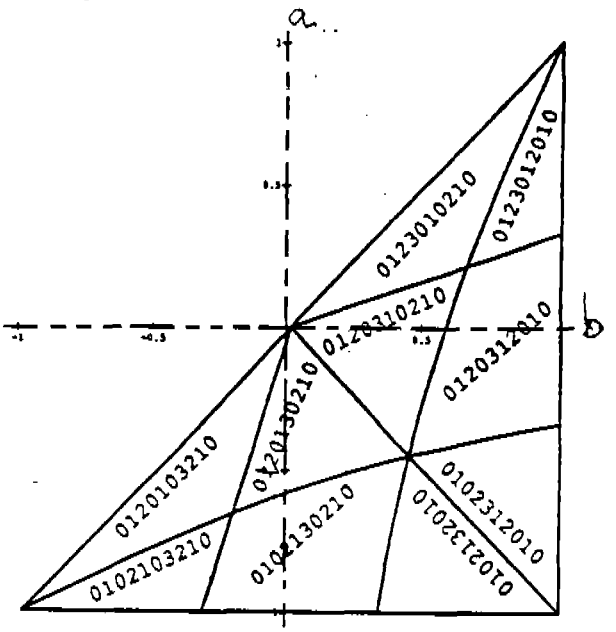


Figure 7

[1] The idea for this graph was suggested by Carl de Boor at the University of Wisconsin, Madison.

Note that only ten regions show up on the map. These are the constructable cases. The root arrangements which are missing are "0120132010" and "0102310210". Note that one is just the reverse of the other. The curves which separate one region from the others correspond to values of a and b where two roots "line up", so that on one side of the curve one of the roots is always to the left of the other, while on the other side of the curve the reverse is true. Figure 8 shows on what side of the curves one would have to be in order for "0120132010" to be constructable.

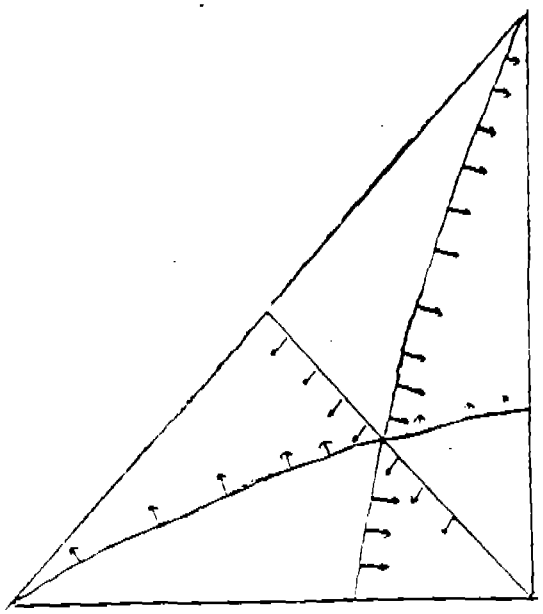


Figure 8

Clearly, there are no points on the correct side of all three curves, and this illustrates why it is not constructable. It was the non-constructability of these two arrangements of the roots that led me to the simplest higher-order Rolle's theorem described above.

Degree 5. Not surprisingly, the complexity of the situation increases rapidly as we increase the degree. In the fifth degree case, there are 286 possible arrangements of the roots satisfying Rolle's theorem. However, only 116 of them can actually be realized. Thus, 170 are not constructable. To prove that an arrangement can be realized, one needs to merely construct a polynomial which generates the correct sequence of roots. To prove that a polynomial is not realizable, however, is somewhat more difficult. For the fifth degree case, one can isolate six "rules" which completely explain all the unconstructable polynomials of fifth degree, and a proof of these six rules has been found. Some of them easily extend to a higher-order Rolle's theorem as in the fourth degree case, while others lead to no obvious such extension.

References

[1] M. Ben-Or, D. Kozen, and J. Reif. The complexity of elementary algebra and geometry. *J. Comp. and Sys. Sci.*, 32:251-264, 1986.

[2] M. Coste, M.F. Roy 1988, Thom's lemma, the coding of real algebraic numbers and the computation of the topology of semi-algebraic sets. *J. Sym. Comp.* 5 121-129.

Iterative Methods and Finite Difference Schemes for Incompressible Flow

John C. Strikwerda*
Department of Computer Sciences and
Center for the Mathematical Sciences

Dongho Shin
Department of Mathematics

University of Wisconsin-Madison
Madison, WI 53706.

Abstract. We consider several methods for solving the linear equations arising from finite difference discretizations of the Stokes equations. The two best methods, one presented here for the first time, apparently, and a second, presented by Bramble and Pasciak, are shown to have computational effort that grows slowly with the number of grid points. The methods work with second-order accurate discretizations. Computational results are shown for both the Stokes and incompressible Navier-Stokes at low Reynolds number.

1. Introduction.

The steady-state Stokes equations in R^d are

$$\begin{aligned} \nabla^2 \vec{u} - \vec{\nabla} p &= \vec{f} \\ \vec{\nabla} \cdot \vec{u} &= g \end{aligned} \qquad \text{in} \qquad \Omega \subset R^d. \tag{1.1}$$

In almost all applications the function g in the second equation of (1.1) is zero, but the methods discussed here do not require that g be zero, so we include this slightly more general case. We consider only the the Dirichlet boundary condition

$$\vec{u} = \vec{b} \qquad \text{on} \qquad \partial\Omega.$$

The methods discussed here should be easy to extend to other boundary conditions. The velocity $\vec{u}$ is a vector of dimension d and the pressure p is a scalar. The domain of our

* The work of this author was supported by the U.S. Army Research Office under grants DAAL03-87-K-0028 and DAAL03-91-G-0094.

computational examples is the unit square in R^2. In current research we are using these methods on more general domains. For a discussion of the mathematical theory related to the Stokes equations see the book by Temam [16].

Let A_h, G_h and D_h be the matrices generated by discretizations of the differential operators $(-\nabla^2), \vec{\nabla}$ and $(-\vec{\nabla}\cdot)$ respectively. The discretization of (1.1) may be written as

$$\begin{pmatrix} A_h & G_h \\ D_h & 0 \end{pmatrix} \begin{pmatrix} u_h \\ p_h \end{pmatrix} = \begin{pmatrix} f_h \\ g_h \end{pmatrix}. \tag{1.2}$$

In some formulations of the difference equations, e.g., staggered grids and finite element formulations, the matrix D_h is the transpose of G_h, i.e., $D_h = G_h^T$. However, in many cases this is not true, either because of boundary conditions or because of the difference schemes. In this paper we do not assume that $D_h = G_h^T$. Note that

$$\det \begin{pmatrix} A_h & G_h \\ D_h & 0 \end{pmatrix} = \det(A_h^{-1}) \det(-Q_h)$$

where

$$Q_h = D_h A_h^{-1} G_h.$$

Hence (1.2) is solvable if A_h and Q_h are invertible.

The methods we discuss here are based on the operator Q_h. We note that

$$u_h = A_h^{-1}(f_h - G_h p_h) \tag{1.3}$$

from the first row in (1.2). Using the second row, we have

$$D_h A_h^{-1}(f_h - G_h p_h) = g_h.$$

Thus (1.2) can be solved by first solving

$$Q_h p_h = h_h \tag{1.4}$$

for p_h where

$$h_h = D_h A_h^{-1} f_h - g_h.$$

After p_h is obtained, u_h can be recovered from (1.3). The operator Q_h is the Schur complement of the system (1.2).

The operator Q_h often has several rather desirable properties. As we show in the next section, Q_h is close to being a symmetric, positive definite operator. Moreover, in

many cases the eigenvalues of Q_h can be bounded independently of the mesh spacing. In these cases, one can use the conjugate gradient method to solve (1.4), and the number of conjugate gradient iterations required to solve (1.4) should be relatively independent of the grid parameters. We call the iterative method based on solving (1.4) by the conjugate gradient method the pressure equation method, and refer to it as the PE method.

The PE method requires that A_h needs to be inverted in each iteration of the conjugate gradient method. This must be done efficiently in order for the overall method to be efficient. Multigrid methods or preconditioned conjugate gradient methods are two possible methods. The price for inversion of A_h would be essentially independent of the grid size when the multigrid method is used, and would grow slowly if a preconditioned conjugate gradient method were used.

The Uzawa iterative method, see [1], can be viewed as solving equation (1.4) by a fixed point iteration. This method can be written as,

$$\begin{aligned} A_h u_h^{\nu+1} &= -G_h p_h^{\nu} + f_h \\ p_h^{\nu+1} &= p_h^{\nu} - \gamma(D_h u_h^{\nu+1} - g_h). \end{aligned} \tag{1.5}$$

The method converges for γ in some interval $(0, \bar{\gamma})$ depending on the scaling of the operators.

A potential disadvantage of these methods is the necessity of inverting A_h at each iteration. There have been a number of iterative methods that avoid the inversion of the operator A_h as required by the Uzawa method. We describe only a few here. For other related methods see [8], and [1].

Bramble and Pasciak [3] proposed an iterative method using a preconditioned conjugate gradient method to solve finite element approximations to the Stokes equations. To avoid the inversion of A_h, Bramble and Pasciak used a preconditioner A_{h0}^{-1}. With the preconditioner, (1.2) is transformed to

$$M_h \begin{pmatrix} u_h \\ p_h \end{pmatrix} = \hat{f}_h \tag{1.6}$$

where

$$M_h = \begin{pmatrix} A_{h0}^{-1} A_h & A_{h0}^{-1} G_h \\ D_h A_{h0}^{-1}(A_h - A_{h0}) & D_h A_{h0}^{-1} G_h \end{pmatrix} \quad \text{and} \quad \hat{f}_h = \begin{pmatrix} A_{h0}^{-1} f_h \\ D_h A_{h0}^{-1} f_h - g_h \end{pmatrix}.$$

They assumed that $G_h^* = D_h$ and

$$0 < ((A_h - A_{h0})u_h, u_h) \leq \alpha(A_h u_h, u_h) \tag{1.7}$$

for all $u_h \neq 0$ and for some α with $0 < \alpha < 1$. If (1.7) is satisfied, then M_h is symmetric and positive definite with the inner product

$$\left[\begin{pmatrix} u_h \\ p_h \end{pmatrix}, \begin{pmatrix} v_h \\ q_h \end{pmatrix}\right] = (A_h u_h, v_h) - (A_{h0} u_h, v_h) + (p_h, q_h)$$

where $(\cdot,\cdot)$ is the usual inner product in the discrete space.

Under an assumption equivalent to the inf-sup condition (see [2]) which implies that the condition number $\kappa(M_h)$ of Q_h is bounded by a constant independent of h, they showed that

$$0 < C_1 \left\| \begin{pmatrix} u_h \\ p_h \end{pmatrix} \right\|^2 \leq \left[M_h \begin{pmatrix} u_h \\ p_h \end{pmatrix}, \begin{pmatrix} u_h \\ p_h \end{pmatrix} \right] \leq C_2 \left\| \begin{pmatrix} u_h \\ p_h \end{pmatrix} \right\|^2 \tag{1.8}$$

for some positive constants C_1 and C_2 and for all $(u_h, p_h)^t$. This implies that $\kappa(M_h)$ is bounded by a constant independent of h and the conjugate gradient is a good method to solve (1.6). We refer to the above iterative method suggested by Bramble and Pasciak as the BP method in more detail.

Strikwerda [14] avoided the inversion of A_h by using one step of successive-over-relaxation. If A_h is written as

$$A_h = \Delta_h - L_h - U_h$$

where Δ_h is the diagonal of A_h and L_h and U_h are strictly lower and upper triangular matrices respectively, then the method introduced in [14] is

$$\begin{aligned} u_h^{\nu+1} &= u_h^{\nu} - \omega \Delta_h^{-1} (\Delta_h u_h^{\nu} - L_h u_h^{\nu+1} - U_h u_h^{\nu} + G_h p_h^{\nu} - f_h) \\ p_h^{\nu+1} &= p_h^{\nu} - \gamma (D_h u_h^{\nu+1} - g_h). \end{aligned} \tag{1.9}$$

We refer to this method as the SOR method.

The number of iterations required by the SOR method is, at best, proportional to h^{-1} where h denotes mesh size [14], and this requires a great amount of time to get a solution for small mesh sizes. For example, Strikwerda and Scarbnick [12] pointed out that the SOR method was quite slow when they used domain decomposition methods. An advantage of the SOR method is the relative simplicity of coding the algorithm. We include this method in our study as a representative of iterative methods that use either time-marching or SOR-like methods to solve the Stokes equations, see Roach [11]. Although there is a great variety of such methods, they all take a number of time steps or iterations that is proportional to h^{-1} at best.

The PE method is the fastest of the methods we compare here. Both the PE and the BP method have work that is proportional to the number of grid points, but the PE

method is faster. In part this is because the PE method needs to invert A_h just once in each conjugate gradient step, while the BP method needs to operate with A_{h0}^{-1} twice. The other reason is that the inner product used in the BP method requires considerable work to compute. This extra work cancels out the advantage of using the preconditioner. The exact comparison of efficiency is done in the section 4.

The PE method doesn't require parameters. This is a significant advantage over the SOR method for which good values of the parameters ω and γ in the SOR method can be hard to find. The BP method also needs a scaling parameter in the preconditioning and, in our experiments, the method was very sensitive to the scaling parameter. In the two subsequent sections, we discuss the PE method and the BP method.

2. Analysis of the Pressure Equation Method.

To analyze the PE method, we first examine the analogous problem for partial differential equations. Define the operator Q for p in $L^2(\Omega)/R$ as

$$Qp := \vec{\nabla} \cdot \vec{\psi}$$

where

$$\nabla^2 \vec{\psi} = \vec{\nabla} p \qquad \text{with} \qquad \vec{\psi}|_{\partial\Omega} = 0.$$

Q can be expressed symbolically as $(-\vec{\nabla}\cdot)(-\nabla^2)^{-1}(\vec{\nabla})$. Crozier [7] has proved the following theorem, see also [9].

Theorem 2.1. *If Ω is a connected, bounded domain in R^2 with smooth boundary, then the operator Q is a bounded, positive definite operator, with bounded inverse, on $L^2(\Omega)/R$.*

The norm of Q is actually bounded by 1. So the above theorem can be expressed mathematically as

$$0 < C||p||^2 \leq (Qp, p) \leq ||p||^2 \tag{2.1}$$

for some positive constant C and for all p. Moreover the operator Q is self-adjoint.

Even more can be said about the eigenvalues of Q. The eigenvalue 1 occurs with infinite multiplicity. This is on the orthogonal compliment of the harmonic functions in $L^2(\Omega)$. We conjecture, based on some evidence, that the rest of the eigenvalues are clustered around one-half.

Conjecture. *The operator Q has the eigenvalue 1 with an infinite multiplicity, and the remaining eigenvalues have a cluster point at 1/2 with no other cluster point.*

If Q_h is a consistent and regular finite difference approximation to Q, then one can expect that Q_h is positive definite and has its condition number bounded by a constant independent of h.

If one uses the usual central difference scheme for D_h and G_h, then Q_h is symmetric. However, if central difference formulas are used for D_h and G_h then the scheme is not regular, see [13], and Q_h will either be singular or be nearly singular.

If the regularized central difference scheme (see [13]) is used for D_h and G_h, then the symmetry of Q_h is lost. However Q_h is close to being symmetric. As our numerical solutions show, the ordinary conjugate gradient method works very well.

The following is the conjugate gradient method we used to find the pressure p_h, see [15]. Let (u_h^0, p_h^0) be an initial solution with u_h^0 having the true boundary values. Let $s_h^0 = r_h^0 = h_h - Q_h p_h^0$ where s_h^ν and r_h^ν denote the search vectors and residual vectors, respectively. Define $q_h^0 = Q_h r_h^0$. The conjugate gradient method for the PE method is

$$
\begin{aligned}
p_h^{\nu+1} &= p_h^\nu + \alpha_\nu s_h^\nu \\
r_h^{\nu+1} &= r_h^\nu - \alpha_\nu q_h^\nu \\
s_h^{\nu+1} &= r_h^{\nu+1} + \beta_\nu s_h^\nu \\
q_h^{\nu+1} &= Q_h r_h^{\nu+1} + \beta_\nu q_h^\nu \\
\alpha_\nu &= \frac{(r_h^\nu, r_h^\nu)}{(s_h^\nu, q_h^\nu)} \\
\beta_\nu &= \frac{(r_h^{\nu+1}, r_h^{\nu+1})}{(r_h^\nu, r_h^\nu)}
\end{aligned}
$$

When A_h is inverted, the boundary values must be assigned to obtain a unique solution. The residual vector r_h in the conjugate gradient method is defined to be $h_h - Q_h p_h$ and initially $r_h^0 = D_h A_h^{-1}(f_h - G_h p_h^0) - g_h$. The first row in the equation (1.2) implies that the boundary values of $A_h^{-1}(f_h - G_h p_h^0)$ have to be the boundary values of u_h, the velocity field of the solution. But, in later steps, when one needs to evaluate $Q_h r_h$, the zero boundary values should be used for A_h^{-1}.

The multigrid process using V-cycles was used to invert A_h. The ordinary Gauss-Seidel iteration was used as the smoother. The number of relaxations in each node of the multigrid was 2. Injection was used to go to a coarser level and interpolation was used to go to a finer level. The residual was computed just before the injection process and at the end of the V-cycles. For the multigrid terminology, refer to [6].

3. The Bramble-Pasciak Method.

The conjugate gradient method applied to (1.6) is defined as the following, refer to [3] for details. Let z_h^0 be an initial approximation to the solution pair $(u_h^0, p_h^0)^t$ with the true boundary values assigned for u_h^0. With $s_h^0 = r_h^0 = \hat{f}_h - M_h z_h^0$, define for $\nu \geq 0$

$$\begin{aligned} \alpha_\nu &= \frac{[r_h^\nu, s_h^\nu]}{[M_h s_h^\nu, s_h^\nu]}, \\ z_h^{\nu+1} &= z_h^\nu + \alpha_\nu s_h^\nu, \\ r_h^{\nu+1} &= \hat{f}_h - M_h z_h^{\nu+1}, \\ \beta_\nu &= -\frac{[M_h s_h^\nu, r_h^{\nu+1}]}{[M_h s_h^\nu, s_h^\nu]}, \\ s_h^{\nu+1} &= r_h^{\nu+1} + \beta_\nu s_h^\nu. \end{aligned} \tag{3.1}$$

Note that, from [15],

$$\alpha_\nu = \frac{[r_h^\nu, r_h^\nu]}{[M_h s_h^\nu, s_h^\nu]} \quad \text{and} \quad \beta_\nu = \frac{[r_h^{\nu+1}, r_h^{\nu+1}]}{[r_h^\nu, r_h^\nu]}. \tag{3.2}$$

Since M_h is positive-definite, (3.2) shows that α_ν and β_ν are nonnegative. This fact can be used to test a good candidates for A_{h0}^{-1}. One possible choice for A_{h0}^{-1} is to let it be one V-cycle for solving

$$A_h u_h = f_h \tag{3.3}$$

when the boundary values of u_h are specified. However this choice of A_{h0}^{-1} may not satisfy (1.7). A better choice is to take

$$A_{h0}^{-1} = \sigma A_{h1}^{-1}$$

where A_{h1}^{-1} is one V-cycle for solving (3.3) and σ is a scaling factor. If σ is chosen improperly, then there is a chance for M_h to be indefinite. This is detected in computation by checking on the positivity of α_ν and β_ν. By changing the value σ, one is able to find a A_{h0}^{-1} satisfying (1.7).

The parameter σ is not hard to find since it is larger than and close to 1 by the following argument. Since $A_{h1}^{-1} A_h \approx I_h$, $\sigma A_{h1}^{-1} A_h \approx I_h$ also for σ near 1. Note $A_h - A_{h1} \approx 0$. To get $A_h - \sigma^{-1} A_{h1} > 0$, σ needs to be larger than and close to 1.

The following comments explain how we implemented the BP method. Some care must be taken to insure good efficiency. From (1.6) and the definition of M_h, the residual vector is

$$r_h = \begin{pmatrix} A_{h0}^{-1}(f_h - A_h u_h - G_h p_h) \\ D_h A_{h0}^{-1}(f_h - A_h u_h - G_h p_h) + D_h u_h - g_h \end{pmatrix}.$$

To compute r_h, first set and save the vector

$$w_h := f_h - A_h u_h - G_h p_h \tag{3.4}$$

for later use. Next the system

$$A_{h0}\hat{w}_h = w_h \tag{3.5}$$

is solved for $\hat{w}_h$ with zero boundary condition, we then have

$$r_h = \begin{pmatrix} r_I \\ r_{II} \end{pmatrix} := \begin{pmatrix} \hat{w}_h \\ D_h(\hat{w}_h + u_h) - g_h \end{pmatrix}. \tag{3.6}$$

In this way the initial residual r_h^0 is computed. Also set $s_h^0 = r_h^0$.

In subsequent iterations, the inner product $[r_h, s_h]$ is computed as

$$\begin{aligned} [r_h, s_h] &= ((A_h - A_{h0})r_I, s_I) + (r_{II}, s_{II}) \\ &= (A_h r_I - w_h, s_I) + (r_{II}, s_{II}). \end{aligned} \tag{3.7}$$

where $s_h = (s_I, s_{II})^t$. The last expression is used to compute $[r_h, s_h]$. Note that A_{h0} is not used explicitly.

$$\begin{aligned} \left[M_h \begin{pmatrix} s_I \\ s_{II} \end{pmatrix}, \begin{pmatrix} s_I \\ s_{II} \end{pmatrix}\right] &= (A_h A_{h0}^{-1}(A_h s_I + G_h s_{II}) - (A_h s_I + G_h s_{II}), s_I) \\ &\quad + (D_h A_{h0}^{-1}(A_h - A_{h0})s_I + D_h A_{h0}^{-1} G_h s_{II}, s_{II}). \end{aligned}$$

To simplify this expression, we set

$$t_h := A_h s_I + G_h s_{II} \tag{3.8}$$

and solve

$$A_{h0}\hat{t}_h = t_h \tag{3.9}$$

for $\hat{t}_h$ with zero boundary condition. If m_I and m_{II} are defined to be $A_h\hat{t}_h - t_h$ and $D_h(\hat{t}_h - s_I)$ respectively, then

$$\left[M_h \begin{pmatrix} s_I \\ s_{II} \end{pmatrix}, \begin{pmatrix} s_I \\ s_{II} \end{pmatrix}\right] = (m_I, s_I) + (m_{II}, s_{II}). \tag{3.10}$$

If the vector $(m_I, m_{II})^t$ is saved, then $[M_h s_h, r_h]$ is computed as

$$(m_I, r_I) + (m_{II}, r_{II}). \tag{3.11}$$

In this whole process, we need to evaluate A_{h0}^{-1}, and never need to evaluate A_{h0} itself. The special forms of α_ν and β_ν in (3.1) were chosen to be easily computable.

4. Analysis of Efficiency.

In this section we estimate the total number of significant operations, which we designate as TSO, for each iterative method. We use these estimates to compare the efficiency of each of these methods. We take as a representative case the Stokes equations on a square in R^2 or cube in R^3. If $N+1$ is the number of grid points in a coordinate direction in R^d, then $(N-1)^d$ is the number of interior grid points. TSO_S, TSO_P and TSO_B are the TSO for the SOR method, the PE method, and the BP Method, respectively. $Iter_S, Iter_P$, and $Iter_B$ are defined similarly.

Let N_A, N_G, and N_D be the number of multiplications per grid point to apply A_h, G_h and D_h, respectively. If $u_h = (u_1, \ldots, u_d)^t$, then

$$(A_h u_h)_{l,m} = ((\nabla_h^2 u_1)_{l,m}, \ldots, (\nabla_h^2 u_d)_{l,m})^t. \tag{4.1}$$

We used the usual second-order accurate discrete Laplacian for ∇_h^2. Since A_h involves d scalar Laplacians, $N_A \approx (2d+1)d$. The regularized central differencing was used to find any first derivative with respect to any direction, and this needs 4 points to evaluate. Each of $(G_h p_h)_{l,m}$ and $(D_h u_h)_{l,m}$ needs d first derivatives to be evaluated, so $N_G \approx 4d$ and $N_D \approx 4d$. We consider our "cost" to be the number of multiplications required.

Lemma 4.1. $TSO_S \approx Iter_S \cdot d(2d+9) \cdot (N-1)^d$.

Proof. From (1.9),

$$\begin{aligned} TSO_S &\approx Iter_S \cdot (N_A + N_G + N_D) \cdot (N-1)^d \\ &\approx Iter_S \cdot (2d^2 + d + 8d) \cdot (N-1)^d \\ &\approx Iter_S \cdot (2d^2 + 9d) \cdot (N-1)^d . \spadesuit \end{aligned}$$

Lemma 4.2. *One V-cycle for the scalar second-order Laplacian costs approximately* $N_V(N-1)^d$ *where* $N_V = 2^d(2^d-1)^{-1}(10d+6)$.

Proof. Going down along a V-cycle, we do 2 smoothing processes, 1 residual finding, and 1 injection at each level. On the way up, we do 2 smoothing processes and 1 interpolation at each level. So, in a V-cycle, altogether 4 smoothing processes, 1 residual finding, 1 injection and 1 interpolation at each level are needed. On the finest level, smoothing costs $(2d+1)(N-1)^d$, computing the residual is about the same, injection and interpolation together cost at most $(N-1)^d$ operations.

Thus one V-cycle costs

$$(5(2d+1)+1)\cdot(N-1)^d\cdot\left(1+\frac{1}{2^d}+(\frac{1}{2^d})^d+\ldots+(\frac{1}{2^d})^{\#\text{ of levels}}\right)$$

where d is the dimension of our domain. The above number is approximately

$$\begin{aligned}&\frac{1}{1-2^{-d}}\cdot(10d+6)\cdot(N-1)^d\\ =&\frac{2^d}{2^d-1}(10d+6)(N-1)^d.\spadesuit\end{aligned}$$

Lemma 4.3. *$TSO_P \approx Iter_P \cdot d(8+\bar{v}N_V)\cdot(N-1)^d$, where $\bar{v}$ is the average number of V-cycles required per iteration.*

Proof. One needs to apply the matrix Q_h in each conjugate gradient iteration. From (4.1), we see that A_h^{-1} consists of d multigrid operations. So, we have by Lemma 4.2,

$$\begin{aligned}TSO_P &\approx Iter_P\cdot(N_G+N_D+d\bar{v}\cdot N_V)\cdot(N-1)^d\\ &\approx Iter_P\cdot(8d+d\bar{v}N_V)\cdot(N-1)^d\spadesuit\end{aligned}$$

Lemma 4.4. *$TSO_B \approx Iter_B\cdot 2d(4d+10+N_V)\cdot(N-1)^d$.*

Proof. In each iteration, the main effort is in finding r_h, $[r_h,s_h]$ and $[M_hs_h,s_h]$ from (3.1). By Lemma 4.2 and the equations from (3.4) to (3.6), the cost to get r_h is

$$(N_A+N_G+d\cdot N_V+N_D)(N-1)^d.$$

Evaluating $[r_h,s_h]$ costs

$$N_A\cdot(N-1)^d$$

by (3.7). The cost of evaluating $[M_hs_h,s_h]$ is

$$(2N_A+N_G+dN_V+N_D)(N-1)^d$$

by the equations from (3.8) to (1.6).

Adding these costs, we obtain

$$\begin{aligned}TSO_B &\approx Iter_B\cdot(4N_A+2N_G+2N_D+2dN_V)\cdot(N-1)^d\\ &\approx Iter_B\cdot(8d^2+4d+16d+2dN_V)\cdot(N-1)^d\\ &\approx Iter_B\cdot(8d^2+20d+2dN_V)\cdot(N-1)^d.\spadesuit\end{aligned}$$

By (1.8) and (2.1), $Iter_P$ and $Iter_B$ are bounded by some constants not depending on mesh size. Moreover, $Iter_S$ is proportional to N at best. For the test case considered in section 6 we find, for $N = 64$ and $d = 2$, $Iter_S \approx 8(N-1)$, $Iter_P = 12$, and $Iter_B = 17$. Also, $\bar{v}$ was about 2 for the PE method. So, $TSO_S \approx 208(N-1)^3$, $TSO_P \approx 1856(N-1)^2$, and $TSO_B \approx 3581(N-1)^2$.

We see that the PE is the fastest method, with the BP method being about twice as much work. The SOR method is 7 times as much work as the PE method for the one case considered here and is even less efficient as N increases. The numerical results in section 6 also show that based on CPU time, for this test case, the PE method is more than 7 times faster and the BP method is about 4 times faster than the SOR method, agreeing with our analysis.

5. The Numerical Experiments.

For the numerical experiment, we used the Stokes equations of the form

$$\nabla^2 u - \frac{\partial p}{\partial x} = -2\pi^2 \sin \pi x \sin \pi y + \pi \sin \pi x \sin \pi y,$$

$$\nabla^2 v - \frac{\partial p}{\partial y} = -2\pi^2 \cos \pi x \cos \pi y - \pi \cos \pi x \cos \pi y,$$

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0,$$

on $0 < x, y < 1$ with u and v specified on the boundary.

The exact solution is given by

$$u = \sin \pi x \sin \pi y,$$

$$v = \cos \pi x \cos \pi y,$$

$$p = \cos \pi x \sin \pi y.$$

The discretization used a uniform grid with the same number of grid points in each direction. The second order accurate five-point Laplacian was used to approximate ∇^2 for all the iterative methods.

We employed, for all the iterative methods, the regularized central difference (see [13])

given by

$$\frac{\partial p}{\partial x} \approx \delta_{x0} p_h - \frac{h^2}{6}\delta_{x-}\delta_{x+}^2 p_h,$$
$$\frac{\partial p}{\partial y} \approx \delta_{y0} p_h - \frac{h^2}{6}\delta_{y-}\delta_{y+}^2 p_h,$$
$$\frac{\partial u}{\partial x} \approx \delta_{x0} u_h - \frac{h^2}{6}\delta_{x+}\delta_{x-}^2 u_h,$$
$$\frac{\partial v}{\partial y} \approx \delta_{y0} v_h - \frac{h^2}{6}\delta_{y+}\delta_{y-}^2 v_h,$$

where h is the grid spacing and δ_{x0}, δ_{x+}, and δ_{x-} are the centered, forward, and backward difference operators in the x-direction. The operators δ_{y0}, δ_{y+} and δ_{y-} are defined similarly for the y-direction.

To obtain the pressure on the boundary, we used the quadratic interpolation, e.g.,

$$p_{0m} = 2p_{1m} - p_{2m},$$

for all the iterative methods.

The SOR method was stopped when the quantities

$$||u_h^{n+1} - u_h^n||, \quad ||v_h^{n+1} - v_h^n||, \quad ||p_h^{n+1} - p_h^n|| \tag{5.1}$$

were all less than $5 \cdot 10^{-5}, 10^{-4}$ and $2 \cdot 10^{-4}$ for mesh sizes 1/32, 1/48 and 1/64 respectively. These values were chosen because the quantities in (5.1) could not be made much smaller than these values. We did not investigate why these quantities could not be made smaller, but presume that it is due to the use of single precision arithemetic. As will be seen, the use of higher precision would not alter our conclusions. The norms of u_h and v_h in (5.1) were the discrete L^2 norms, and the norm for p_h was the L^2 norm in its quotient space (see [14]). The relaxation parameters ω and γ were given by

$$\omega = 2/(1 + c_0 h), \qquad \gamma = c_1 h$$

where c_0=3.14 and c_1=4.5. See [13], [14] for more details.

The PE method was stopped when the residual was less than 10^{-6}. In each conjugate gradient iteration of the PE method, the multigrid process using V-cycles was used to invert A_h. We found that to achieve good overall accuracy it was only necessary to do enough V-cycles to reduce the residual in the L^2 norm to less than 10^{-4}. Each multigrid process to solve $A_h u_h = f_h$ for u_h was stopped when either the number of V-cycles was

4 or the residual error was less than 10^{-4}. The maximum number of V-cycles was chosen to be 4 since the residual error didn't change significantly after 4 V-cycles. Because the reduction factor of the error is small in the multigrid process, more than 4 V-cycles would rarely be needed. With these stopping criteria, the average number of V-cycles needed in each conjugate gradient iteration was 2.

The BP method was stopped when the residuals were less than $3 \cdot 10^{-4}, 10^{-4}$ and $3 \cdot 10^{-5}$ for mesh sizes 1/32, 1/48 and 1/64 respectively. These values were chosen since, similar to the SOR method, the residuals decreased to values a little bit smaller than these values, but could not be made much smaller. Again, this is probably due to the precision of the computer arithmetic. In the BP method, several values were run for σ, the value of 1.2 worked well.

6. Test Results.

Tables 1, 2, and 3 show the errors for the PE method, the BP method and the SOR method. The column labeled "*time*" shows the CPU time required for the total computation.

Table 1

Errors and CPU time for the PE method.

N	*iter*	u	p	*time*
32	12	6.46(−5)	2.71(−3)	1.617
48	12	2.35(−5)	1.25(−3)	4.347
64	12	1.38(−5)	6.91(−4)	8.362

Table 2

Errors and CPU time for the BP method.

N	*iter*	u	p	*time*
32	14	6.34(−5)	3.04(−3)	2.843
48	16	2.19(−5)	1.36(−3)	8.558
64	17	1.19(−5)	7.83(−4)	17.162

By comparing CPU times, one can see that the PE method is most efficient, and the BP method takes about twice as much effort, and the SOR method is least efficient, taking

Table 3

Errors and CPU time for the SOR method.

N	*iter*	u	p	*time*
32	275	6.33(−5)	2.74(−3)	8.546
48	399	2.48(−5)	1.26(−3)	28.150
64	511	1.39(−5)	6.98(−4)	65.214

about 7 times as much time as the PE method. Note that the number of iterations taken by the PE method and the BP method are essentially independent of mesh size, which supports (1.7) and (2.1).

The next table, Table 4 shows the accuracy of the PE method, the BP method, and the SOR method. The order of accuracy was obtained from the formula $\log(\text{error}(h_1)/\text{error}(h_2))/\log(h_2/h_1)$ where h_1, h_2 are mesh sizes with $h_1 < h_2$. All numerical solutions show that they are second-order accurate.

Table 4

Order of accuracy for the computed solutions.

	PE		BP		SOR	
N_1, N_2	u	p	u	p	u	p
64, 48	2.1	2.0	1.9	2.1	2.1	1.9
64, 32	2.2	2.0	2.2	2.0	2.4	2.0
48, 32	2.3	1.9	2.5	1.9	2.6	2.0

7. Navier-Stokes Equations.

The steady-state Navier-Stokes equations in R^d are of the form

$$\begin{aligned} -R^{-1}\nabla^2\vec{u} + (\vec{u}\cdot\vec{\nabla})\vec{u} + \vec{\nabla}p &= \vec{f}, \\ \vec{\nabla}\cdot\vec{u} &= g \qquad \text{in} \qquad \Omega \subset R^d \end{aligned} \tag{7.1}$$

where R is the Reynolds number. We consider the Dirichlet boundary condition

$$\vec{u} = \vec{b} \qquad \text{on} \qquad \partial\Omega.$$

There are several possible extensions of the PE method from the Stokes equations to the Navier-Stokes equations, depending on how one linearizes the first equation in (7.1).

To apply the PE method efficiently to (7.1), we used the following algorithm which worked for R up to about 100.

(1) Start with an initial solution $\vec{u}^0, p^0$.

(2) Given the solution $\vec{u}^\nu$, let

$$\vec{d}^\nu := (\vec{u}^\nu \cdot \vec{\nabla}_h)\vec{u}^\nu,$$
$$\vec{f}_1^\nu := \vec{f} - \vec{d}^\nu.$$

where $\vec{\nabla}_h$ is a finite discretization of $\vec{\nabla}$, then (7.1) can be expressed as

$$\begin{aligned} -R^{-1}\nabla^2\vec{u} + \vec{\nabla}p &= \vec{f}_1^\nu, \\ \vec{\nabla}\cdot\vec{u} &= g. \end{aligned} \tag{7.2}$$

(3) The system (7.2) gives an equation for pressure p which is

$$Q_h p_h = h_h^\nu \tag{7.3}$$

where the function h_h^ν is generated by $\vec{f}_1^\nu$ and g_h. Apply the PE method to (7.3), i.e., do several conjugate gradient iterations to update $p^{\nu+1}$ from p^ν.

(4) Let

$$\vec{f}_2^\nu = \vec{f} - \vec{\nabla}p^{\nu+1},$$

then the first equation in (7.1) is the so-called convection diffusion equation

$$-R^{-1}\nabla^2\vec{u} + (\vec{u}\cdot\vec{\nabla})\vec{u} = \vec{f}_2^\nu. \tag{7.4}$$

To update $\vec{u}^{\nu+1}$, solve (7.4) for $\vec{u}$. We discuss the solution procedure later. Go to step (2).

For our numerical experiment, we used the Navier-Stokes equations of the form

$$\begin{aligned} -R^{-1}\nabla^2 u + uu_x + vu_y + p_x &= f_1 \\ -R^{-1}\nabla^2 v + uv_x + vv_y + p_y &= f_2 \\ u_x + v_y &= 0 \end{aligned}$$

on $0 < x, y < 1$ where

$$f_1 = 2R^{-1}\pi^2 \sin\pi x \sin\pi y + 0.5\pi\sin(2\pi x) - \pi \sin\pi x \sin\pi y$$
$$f_2 = 2R^{-1}\pi^2 \cos\pi x \cos\pi y - 0.5\pi\sin(2\pi y) + \pi \cos\pi x \cos\pi y.$$

The values of u and v are specified on the boundary.

The exact solution is given by

$$u = \sin \pi x \sin \pi y,$$
$$v = \cos \pi x \cos \pi y,$$
$$p = \cos \pi x \sin \pi y.$$

Because of the nonlinearity of (7.4), the Full Approximation Scheme (FAS) was used for multigrid solver. See [5] for a description of FAS. Moreover the full weighting was used in the fine-to-coarse transfers of both the solution and the residual functions. To employ a stable discretization, upwind differencing was used for the first derivatives in (7.4) when the mesh size h was larger than $2/RU$ where U is the maximum value of $\vec{u}$ on the domain, see [10]. Otherwise, the central differencing was used to get the overall second-order accuracy. In [4], the authors mentioned that it is better to employ upwind differencing only in the relaxation sweeps, central differencing in the residual transfers, but we obtained the best numerical solution when the same differencing was used in both relaxation sweeps and residual transfers. Also, the computation of $\vec{f}_2^{\nu}$ at coarser levels used upwind differencing.

Table 5 and Table 6 show the error and accuracy of the solution when R is 30. Notice that the method is second-order accurate.

Table 5

Errors for $R = 30$.

N	u	p
32	7.34(−4)	3.65(−4)
48	2.01(−4)	1.55(−4)
64	9.35(−5)	8.94(−5)

Table 6

Accuracy of the solution for $R = 30$.

N_1, N_2	u	p
64, 48	2.7	1.9
64, 32	3.0	2.0
48, 32	3.2	2.1

8. Conclusions.

The pressure equation method has been shown to be an efficient numerical method for solving the steady Stokes equations. Since the work is essentially proportional to the number of grid points, the efficiency of this method is exceptional. We have also shown that the method advocated by Bramble and Pasciak is not as efficient for the finite difference schemes used here.

The pressure equation method has been extended to the Navier-Stokes equations for low Reynolds numbers. Research is continuing on improving this method. Work is also being done on applying the method to time-dependent problems and using the method with domain decomposition.

REFERENCES

[1] K. Arrow, L. Hurwitz and H. Uzawa, *Studies in Nonlinear Programming*, Stanford University Press, Stanford, 1958.

[2] A. K. Aziz & I. Babuška, *"Survey lectures on the mathematical foundations of the finite element method, Part I"*, in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations*, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 1-362.

[3] J. H. Bramble and J. E. Pasciak, *"A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems"*, Math. Comp., 50 (1988), pp. 1-18.

[4] A. Brandt and N. Dinar, *"Multigrid solutions to elliptic flow problems"*, in *Numerical Methods for Partial Differential Equations*, S. V. Parter, ed., Academic Press, Inc., New York, 1979, pp. 53-148.

[5] A. Brandt, *"Guide to multigrid development"*, in *Multigrid Methods*, W. Hackbusch and U. Trottenberg, ed., Springer-Verlag, New York, NY, 1981, pp. 220-312.

[6] W. L. Briggs, *Multigrid Tutorial*, Lancaster Press, Lancaster, Pennsylvania, 1987.

[7] M. Crozier, *Approximation et methodes iteratives de resolution d'inequations variationnelles et de problemes non lineares*, IRIA cahier no 12., 1974.

[8] M. Fortin and R. Glowinski, *Resolution Numerique de Problèmes aux Limites par des Methodes de Langrangien Augment*, 1981.

[9] V. Girault and P. A. Raviart, *Finite Element Approximation of the Navier-Stokes Equations*, Lecture Notes in Mathematics, 749, Springer-Verlag, New York, NY, 1979.

[10] W. Hackbusch, *Multi-Grid Methods and Applications*, Springer-Verlag, New York, NY, 1980.

[11] P. Roach, *Computational Fluid Dynamics*, Hermosa, Albuquerque, NM, 1972.

[12] J. C. Strikwerda and C. D. Scarbnick, *"A domain decomposition method for incompressible viscous flow"*, SIAM J. Sci. Stat. Comput., to appear (1991).

[13] J. C. Strikwerda, *"Finite difference methods for the Stokes and Navier-Stokes equations"*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 56-68.

[14] J. C. Strikwerda, *"An iterative method for solving finite difference approximations to the Stokes equations"*, SIAM J. Numer. Anal., 21 (1984), pp. 447-458.

[15] J. C. Strikwerda, *Finite Difference Schemes and Partial Differential Equations*, Wadsworth & Brooks/Cole, Pacific Grove, CA, 1989.

[16] R. Temam, *Navier-Stokes Equations*. Elsevier Science Publishing Company, Inc., New York, NY, 1984.

NUMERICAL SIMULATION OF SABOT DISCARD AERODYNAMICS USING COMPUTATIONAL FLUID DYNAMICS

Michael J. Nusca
Free Flight Aerodynamics Branch, Launch and Flight Division
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005

Abstract. Computational fluid dynamics (CFD) solutions of the three-dimensional Navier-Stokes equations have been applied to sabot discard aerodynamics for gun-launched, saboted, armor-piercing projectiles. The portion of the launch cycle which involves strong aerodynamic interference between the projectile and discarding sabot (carrier) components has been investigated. Three sabot components were located symmetrically at various positions near the projectile and at angle of attack. The complex system of shock/boundary-layer interactions between multiple bodies (projectile and sabots), during the discard sequence, has been numerically simulated. Computed and measured surface pressures compare favorably for Mach number 4.5 and Reynolds number six million per meter. Comparison of symmetric sabot discard trajectories predicted using CFD and the AVCO sabot design code are shown.

Introduction. Currently, the most widely utilized design for kinetic energy, antitank applications is the gun-launched, fin-stabilized, long-rod projectile. The cross-sectional diameter of the rod is smaller than the diameter of the gun bore. Fins span the area between the rod and the gun tube. Therefore, a sabot (or carrier) is required to reduce in-bore balloting of the projectile. Once free of the gun tube the sabot must be discarded in order to permit unconstrained, low-drag flight to the target. The sabot is divided into three or four components along axial planes. For smooth bore gun tubes, these components separate from the projectile under the action of elastic and aerodynamic loads. Figure 1 shows a photograph and shadowgraph of typical sabot discard during free flight.

It has been demonstrated[1] that aerodynamic interference generated by the sabot components can be a significant source of projectile launch disturbance leading to unacceptable loss of accuracy at the target. Perturbations to the projectile's trajectory are magnified by geometric asymmetry in the discard pattern and by extended periods during launch when the sabot components are in close proximity to the projectile. A detailed understanding of the three-dimensional shock/boundary-layer interference flowfield between the sabot and the projectile (see Fig 1b) is not available.

An extensive experimental program to investigate the aerodynamics of sabot discard has been conducted.[2] During these tests, a projectile and three sabot components were sting-mounted in the NASA Langley Unitary Plan wind tunnel facility 4 x 4 ft test section. The model configuration included a stationary cone-cylinder projectile (without fins) at zero angle-of-attack and three 120° included-angle sabot components located symmetrically around the projectile. Figure 2a shows a schematic (cross-section) of the wind tunnel model (one sabot shown). The cylinder section of the projectile was 50.8mm in diameter; the projectile had a length-to-diameter ratio of 10.5 and a 30° included-angle conical nose. Fifty static pressure taps were positioned on the surface between the 120° planes of symmetry,

with four taps on the conical section. The sabot had cylindrical inner and outer surfaces of radii 25.4 and 76.2mm, respectively, with the a leading edge chamfer of 40°. Fifty static pressure taps were located on the inner and outer surfaces. The test Mach number and Reynolds number were 4.5 and 6.6 million per meter, respectively. A typical flight Reynolds number of 89 million per meter could not be reproduced in the tunnel; unfortunately, test results showed regions of shock/boundary-layer interaction, separated flow and other viscous phenomena.

Initial analytical work for sabot discard aerodynamics relied on the Newtonian flow approximation and empirical aerodynamic interaction analyses; for example the AVCO code.[3,4] These assumptions make discard computations tractable and in some cases represent accurate approximations. However, it is apparent that the multiple shock/expansion interaction flowfield between the projectile and sabot petals is an essential part of the analysis. The initial version of the AVCO code[3] evaluated the aerodynamic loadings on the sabot segments using Newtonian theory and a subsonic/supersonic inlet model; pressure forces on each surface of the segments, including sabot sides, were obtained separately and summed to provide results for total force and moments (excluding shear stress components). The code assumed that the aerodynamic coefficients for the projectile were known. Although the sabot separation process is initially dominated by aerodynamic interaction, the code assumed one-dimensional flow between the bodies. Recent versions[4] include an integrated flow element approach utilizing local shock/expansion procedures based on sabot surface pressures measured during wind tunnel tests.[2] These test data are used to determine pressure levels on certain sabot locations with linear variations assumed between these points. As a result, the code includes the effects of pressure pulses on the bodies caused by impinging and reflecting shock waves. When the sabot petals are not in close proximity to the projectile, Newtonian flow theory is used. In some cases, however, these code improvements produced overestimates of the discard process in contrast to initial code predictions. Consideration is limited to a general sabot configuration which is bounded radially by two cylindrical surfaces and axially by two conical surfaces.

This paper describes computational fluid dynamics (CFD) solutions applied to the three-dimensional (3D) Navier-Stokes equations for symmetric sabot discard. During symmetric discard multiple sabot components are assumed to follow identical trajectories away from the projectile, and the projectile is assumed to be at zero angle-of-attack. As shown in Figure 2b, the computational domain can therefore be limited to a smaller portion of the entire flowfield around the configuration; this reduces computational grid size, computer memory, and computer run time. For three sabot components this domain spans a 60° sector from sabot midplane to symmetry plane between neighboring sabot components. For asymmetric discard the computational domain would be greatly expanded (i.e. a full 360° sector) with a corresponding increase in computer requirements. The portion of the launch cycle which involves strong aerodynamic interference between the projectile and the sabot components is examined. Thus, simulations are performed for small vertical separation of the sabot from the projectile surface, $\Delta y/D \leq 1$ (D = projectile rod diameter = 1 cal. or 50.8mm in Fig. 2a) and sabot angle of attack $\alpha \leq 10°$. Previous work described code validation with wind tunnel results.[5,6] A four-stage sabot discard sequence was numerically simulated for the wind tunnel model configuration[7,8] This simulation has been extended to ten stages and resultant aerodynamic forces and moments computed from the flowfield. The

symmetric sabot discard trajectory can then be simulated and compared to results obtained using the AVCO code. The flowfield for a M865 projectile/sabot has also been simulated.

Computational Approach. CFD can be used to simulate the compressible flowfield around aerodynamic bodies by solving the 3D Reynolds-averaged Navier-Stokes (RANS) equations. The USA-PG3 code was developed by Chakravarthy[9,10] The RANS equations are written using a perfect gas assumption. Both laminar and turbulent flows can be investigated thus, a turbulence model[11] is required for closure. In addition, backflow regions can be present thus, a backflow turbulence model[12] is included. The equations are transformed into conservation law form and discretized using finite-volume approximations. The USA-PG3 code uses a class of numerical algorithms termed total variational diminishing (TVD). The resulting set of equations is solved using an implicit, factored, time-stepping algorithm. The solution takes place on a computational grid that is generated around the configuration in zones; zonal boundaries are transparent to the flowfield.

Equations of Motion. The RANS equations for 3D flow are written in the following conservation form. The dependent variables u, v, w, and e are mass-averaged.

$$\frac{\partial W}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} + \frac{\partial H}{\partial z} = 0 \tag{1}$$

$$W = (\rho \quad \rho u \quad \rho v \quad \rho w \quad \rho e)$$

$$F = \begin{pmatrix} \rho u \\ \rho u^2 - \sigma_{xx} \\ \rho uv - \tau_{xy} \\ \rho uw - \tau_{xz} \\ \rho ue + \dot{q}_x - \sigma_{xx}u - \tau_{xy}v - \tau_{xz}w \end{pmatrix}$$

Arrays G and H are similar in form to array F (see Ref. 5). Normal stress (σ), shear stress (τ), heat transfer ($\dot{q}$) and energy (e) are defined elsewhere.[5] The laminar and eddy viscosities, μ and μ_t, are implicitly divided by the reference Reynolds number (Re). The flow medium is assumed to be a perfect gas satisfying the equation of state $p = \rho \Re T$. A power law[13] is used to relate molecular viscosity, μ, to temperature. The laminar and turbulent Prandtl numbers, Pr and Pr_t, are assumed constant with values of 0.72 and 0.9 respectively. The ratio of specific heats, γ, is also assumed constant. Assuming a time-invariant grid and using the transformation of coordinates implied by $\tau = t$, $\xi = \xi(x, y, z)$, $\eta = \eta(x, y, z)$ and $\zeta = \zeta(x, y, z)$, Equation 1 can be recast into conservation form where ξ, η and ζ are the new independent variables and x_ξ, x_η, x_ζ, y_ξ, y_η, y_ζ, z_ξ, z_η and z_ζ are the nine transformation coefficients obtained numerically from the mapping procedure. The transformed time variable is represented by τ.

The shock/boundary-layer interference flowfield between projectile and sabots can include regions of recirculating flow. To improve the predictive capability of separated flows using RANS codes a new turbulence model has been recently developed by Goldberg.[12] The new model is based on experimental observations of detached flows and allows turbulence due to local shear effects to be taken into account in addition to wall-shear contributions. The velocity scale function, which is normally $y\omega$, is modified as $(y - y_e)\omega$ (for $y \geq y_e$). Here, ω is the magnitude of the local vorticity and y_e is the location away from the wall where the

vorticity first diminishes to a small fraction of its local maximum magnitude. From this location onward the length scale is given by $y_{max} - y_e$. The model prescribes turbulence kinetic energy and dissipation analytically within backflows. A formula for the eddy viscosity (μ_t) within backflows is derived and used for the RANS equations when calculations are done inside separation bubbles. Outside of them, another turbulence model[11] supplies the values of eddy viscosity.

Computational Algorithm. The spatial discretization technique for the equations of motion must successfully capture the complex physics of interacting projectile/sabot flowfields. The TVD formulation for the convective terms along with a special treatment of the dissipative terms (Eq. 1) provides an appropriate simulation. In recent years, TVD formulations have been constructed for shock-capturing finite-difference methods.[9,10] Near large gradients in the solution (extrema) TVD algorithms automatically reduce to first-order accurate discretizations locally while away from extrema they can be constructed to be of higher-order accuracy. This local effect restricts the maximum global accuracy possible for TVD algorithms to third order for steady-state solutions. TVD methods manifest many properties desirable in numerical solution procedures. By design they avoid numerical oscillations and "expansion shocks" while at the same time being of higher-order accuracy. TVD formulations are also based on the principle of discrete or numerical conservation which is the numerical analog of physical conservation of mass, momentum, and energy. Thus, TVD algorithms can "capture" flowfield discontinuities (e.g. shock waves) with high resolution. At a fundamental level they are based on upwind algorithms; therefore, they closely simulate the signal propagation properties of hyperbolic equations. Algorithms based on the TVD formulation are completely defined. In contrast, algorithms based solely on central differences involve global dissipation terms for stability and have one or more coefficients that must be judiciously chosen to achieve desirable results. Any conventional time discretization method suitable for the Navier-Stokes equations can be used together with this space discretization methodology; for example, approximate factorization and relaxation techniques.

Computational Grid. Numerical simulation of the interacting flowfield about projectile/sabot combinations is complicated by the non-axisymmetric, multiple-wall geometry. The computational domain is divided into zones of simple geometric shape. In each zone an algebraic grid is generated with grid clustering near walls and high flow gradient regions. The computational method is constructed such that each zone is considered an independent module interacting with other zones before or after the information corresponding to each zone is updated one cycle. Zonal boundaries are transparent to the flowfield. A typical 6-zone grid used for computations described in this paper is designed as follows (see Ref. 5): grid zone 1 covers the projectile from nose to base, zone 2 covers the area between zone 1 and the inner surface of the sabot, zone 4 covers the area between the outer surface of the sabot and the uppermost extent of the computational domain, zones 5 and 6 cover the projectile and sabot base regions, respectively. Zones 1 thru 6, excluding zone 3, extend from $\phi = 0$ to 60° in the azimuthal direction. Grid zone 3 covers the area between the sabot and the azimuthal extent of the computational domain. The entire 6-zone grid consists of 300,000 nodes and requires 10 million words of memory on a CRAY-2 supercomputer. Converged solutions require about 10 CPU hours.

Results. Figure 3 shows the measured[2] and computed pressure distributions over the projectile and sabot surface in the pitch plane; the pitch plane (Fig. 2b) bisects the azimuthal

planform of the sabot. Three sabot components are modeled with sabot bases aligned with the projectile base, $\Delta x/D = 0$, projectile surface and sabot inner surface vertically separated by $\Delta y/D = .75$, and the sabots at zero angle-of-attack. Laminar boundary layer modeling was employed; turbulent solutions are described elsewhere.[5,6] Computed pressures on the projectile surface agree favorably with the magnitude and location of a measured pressure peak ($x/D \simeq 4.22$) as well as elevated pressures preceding this peak, $2 \leq x/D \leq 4.22$. The location of this pressure peak corresponds to the termination of a low speed flow region on the projectile. Downstream of the pressure peak the agreement between computation and measurement is also favorable. On the inner surface of the sabot, numerical simulation adequately predicts the pressure level and trend on the sabot slant surface, $2.75 \leq x/D \leq 3.94$. Pressure levels on the rest of the sabot section agree with measurements including a pressure rise at $x/D \simeq 5.5$.

References 5-8 describe further results obtained for the wind tunnel model. For cases where the sabot petals are close to the projectile ($\Delta y/D \leq .75$) a low speed ($M \leq 1$) recirculating flow pocket forms between the projectile and the beveled section of the sabot petals. This causes a strong oblique shock on the projectile surface where the pocket forms and a high pressure pulse where the pocket terminates. As the sabot petals discard, a normal shock, formed at the leading edge of the sabot, becomes an oblique shock that intersects the projectile surface in a regular reflection. Inviscid flow simulations require significantly less computer time by excluding the viscous terms in the Navier-Stokes equations. However, the inviscid simulation predicts lower pressures on the projectile and sabot than measured or predicted by laminar and turbulent simulations. Turbulent calculations are similar to laminar for the low Reynolds number wind tunnel data. Comparison of CFD predictions with projectile surface data measured azimuthally about the projectile agree with the trend but not the magnitude these pressures (in particular the pressure peak, as shown in Figure 3, reduces as measured azimuthally about the body). Azimuthal grid refinement increases the level of agreement. Computations for the 2D/axisymmetric equivalent of three sabot petals (i.e. petals joined into a concentric tube with the projectile centerline) are computationally inexpensive but result in flowfields that are very different from the 3D case.

Figure 4a thru 4j show computed laminar, steady-state, pressure contours in the pitch plane for the forward part of the projectile/sabot configuration and ten stages of the programmed discard sequence. Three horizontal lines extending from $x/D = 0$ to 7.03 are zonal grid boundaries. Large flow gradients (e.g. shock waves) are indicated by clustering of pressure contour lines. Pressure levels are the same for Figs 4b-4j, $1 \leq P/P_\infty \leq 40, \Delta P/P_\infty = .5$, and for Fig 4a. $1 \leq P/P_\infty \leq 100, \Delta P/P_\infty = 1$ due to higher stagnation pressures.

The programmed discard sequence shown in Figs 4a-4j covers four vertical displacements of the sabot inner surface with respect to the projectile surface ($\Delta y/D$) and six sabot angles of attack (with respect to the projectile). The projectile was assumed to be at zero yaw with respect to the freestream and the Mach number was constant as 4.5. Since the time during which the sabot petals and projectile are in close proximity is usually short (about 2 ms or 1.5 meters from the gun), the assumption of constant Mach number is not unreasonable. This quasi-steady, programmed simulation ignores the flow time dynamics and does not link the aerodynamic forces to the sabot motion. However, such a simulation serves as a prelude to computations that utilize coupling of unsteady aerodynamics and rigid-body motion.

As seen in Figs 4a-4j, the sabot generates a strong series of shock waves, beginning as a detached nearly-normal shock that intersects the projectile surface as a strong oblique shock, and ending as an attached oblique shock that intersect the projectile surface in a regular reflection. Flow between the sabot inner surface and the projectile surface begins as a choked nearly-uniform high pressure field with transition into reflected shocks (from sabot back to projectile) that become more pronounced. Beginning with Fig. 4e, a low pressure bubble develops on the sabot inner surface extending from $x/D = 3.94$ to the next shock impingement on the sabot surface. Combined with the high pressure on the sabot beveled section ($2.75 \leq x/D \leq 3.94$) this low pressure region provides a force couple that promotes sabot discard.

Using the simulated sabot discard sequence described above, the corresponding aerodynamic forces (lift and drag) and pitching moment can be computed. This is accomplished by integrating the sabot surface pressure and shear stress distributions for each stage of the discard sequence. The sabot mass properties are used to compute vertical and horizontal accelerations which are assembled in a table as functions of sabot $\Delta y/D$ and α. A modified point-mass trajectory model is used to compute the sabot center of gravity (CG) location as a function of time using double-interpolation from values in the table. Figure 5 shows a comparison between the sabot CG location (both in the axial and radial directions) computed using the AVCO semi-empirical code and the present simulation using CFD. The present predictions match the AVCO values for early times, but diverge later in the simulated discard event. In the AVCO simulation sabot discard progresses faster than predicted using the current method. The relatively good agreement for early times in the discard event may be a result of the sabot/projectile interference methods included in the AVCO code. Reasons for discrepancies in the predictions at later times are still under investigation. One possibility is that the Newtonian theory used to predict aerodynamic forces when the sabot is not in close proximity to the projectile, results is lift and drag values that are larger than predicted using CFD. In comparing the AVCO prediction to that using CFD, several points should be noted. Both methods used the same sabot geometry and mass properties, freestream flow conditions and assumed a symmetric discard. Both methods are quasi-steady in nature, using a database of steady aerodynamic force predictions to simulate a dynamic event. However, the source of the aerodynamic data is very different between the codes (see Introduction for a discussion of the AVCO code). By virtue of the rapid aerodynamic methods incorporated into the AVCO code, a much larger aerodynamic force and moment database is available. The trajectory time-integration step for the AVCO code was much smaller that that used in the present study.

Figure 6a shows the projectile/sabot configuration of the Army M865 anti-tank round. The configuration has been altered somewhat in order to simplify computational grid generation. These alterations are also illustrated in Figure 6a. The sabot was located .75 calibers above the projectile (1 caliber = 38mm) and at zero angle-of-attack. A simulated sabot discard sequence like that used for the wind tunnel model is in progress. Figure 6b shows the laminar flow pressure contours for the M865. The Reynolds number for this flow is 6.6 million per meter. The freestream Mach is 4.5.

Conclusions and Future Work. CFD solutions of the 3D Navier-Stokes equations have been applied to the aerodynamics of symmetric sabot discard. A steady simulated sabot discard sequence using fixed sabot locations (with respect to the projectile) reveals

shock/shock and shock/boundary-layer interactions in the flowfield. The freestream Mach number was 4.5 and laminar boundary layer modeling was employed for Re 6.6 million per meter. Numerical simulations have also been performed using Re of 89 million per meter and flows with turbulence modeling.[5] The steady-state approach that uses predetermined sabot positions has lead to enhanced understanding of the discard event, serving as a prelude to computations that utilize coupling of unsteady aerodynamics and rigid-body motion. A technique for the integration of surface pressures and shear stress was developed for the wind tunnel model sabot. A more general method is being developed to determine the aerodynamic forces and moments acting on the M865 sabot.

Numerical mesh generation for the solution of complex flowfields about realistic projectile/sabot configurations may be greatly simplified by the use of unstructured (i.e. finite-element like) grids. Figure 7 shows the planar view (i.e. slice thru the pitch plane of the projectile/sabot) of a typical unstructured grid for the Army M829 sabot. Solution of the Euler equations on unstructured grids is being accomplished by Chakravarthy[14]. Work on unstructured grids and moving grid zones will eventually lead to a more realistic simulation of the discard event.

Acknowledgement. Dr. E.M. Schmidt, Chief, Fluid Physics Branch, Launch and Flight Division, US Army BRL has substantially supported this work.

References

1. Schmidt, E.M. and Shear D.D., "Aerodynamic Interference During Sabot Discard," *Journal of Spacecraft and Rockets*, AIAA, Vol. 15, No. 3, May-June 1978, pp. 162-167.

2. Schmidt, E.M., "Wind-Tunnel Measurements of Sabot-Discard Aerodynamics," *Journal of Spacecraft and Rockets*, AIAA, Vol. 18, No. 3, May-June 1981, pp. 235-240.

3. Crimi, P., and Siegelman, D., "Analysis of Mechanical and Gasdynamic Loadings During Sabot Discard from Gun-Launched Projectiles," US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, ARBRL-CR-341, June 1977.

4. Siegelman, D., Wang, J., and Crimi, P., "Computation of Sabot Discard," US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, ARBRL-CR-505, Feb. 1983.

5. Nusca, M.J., "Computational Fluid Dynamics Application to the Aerodynamics of Symmetric Sabot Discard," Technical Report BRL-TR-3167, US Army Ballistic Research Laboratory, Aberdeen Proving Ground, MD, Oct. 1990.

6. Nusca, M.J., "Computational Fluid Dynamics Application to the Aerodynamics of Symmetric Sabot Discard," AIAA-90-3096, *Proceedings of the 8th AIAA Applied Aerodynamics Conference*, Portland OR, August, 1990.

7. Nusca, M.J., "Numerical Simulation of Sabot Discard Aerodynamics Using Computational Fluid Dynamics," *Proceedings of 1991 Simulation Multiconference, Ballistics Simulation II,* Society for Computer Simulation International, New Orleans, LA, April 1991.

8. Nusca, M.J., "Numerical Simulation of Sabot Discard Aerodynamics Using Computational Fluid Dynamics," *Proceedings of 1991 Summer Computer Simulation Conference,* Society for Computer Simulation International, Baltimore, MD, July 1991.

9. Chakravarthy, S.R., Szema, K.Y., Goldberg, U.C., Gorski, J.J., and Osher, S., "Application of a New Class of High Accuracy TVD Schemes to the Navier-Stokes Equations," AIAA-85-0165, *Proceedings of the 23rd AIAA Aerospace Sciences Meeting*, Reno NV, Jan. 1985.

10. Chakravarthy, S.R., Szema, K.Y., and Haney, J.W., "Unified Nose to Tail Computational Method for Hypersonic Vehicle Applications," AIAA-88-2564, *Proceedings of the 6th AIAA Applied Aerodynamics Conference*, Williamsburg VA, June, 1988.

11. Baldwin, B.S., and Lomax, H., "Thin Layer Approximation and Algebraic Model for Separated Turbulent Flows." AIAA-78-257, *Proceedings of the 16th AIAA Aerospace Sciences Meeting*, Huntsville AL, Jan. 1978.

12. Goldberg, U.C., "Separated Flow Treatment with a New Turbulence Model," *AIAA Journal*, Vol. 24, No. 10, Oct. 1986, pp. 1711-1713.

13. Mazor, G., Ben-Dor, G., and Igra, O., "A Simple and Accurate Expression for the Viscosity of Nonpolar Diatomic Gases up to 10,000 K," *AIAA Journal*, Vol. 23, No. 4, April 1985, pp. 636-638.

14. Chakravarthy, S.R., Szema, K.Y., and Chen, C.L., "A Universe Series Code for Inviscid CFD with Space Shuttle Applications Using Unstructured Grid," AIAA-91-3340, *Proceedings of the 9th AIAA Applied Aerodynamics Conference*, Baltimore MD, Sept. 1991.

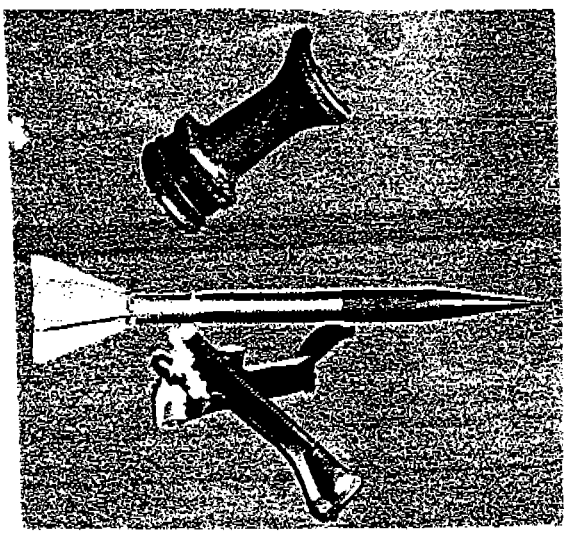


Fig. 1a. Photograph of typical kinetic energy long-rod projectile in free flight during three-petal sabot discard.

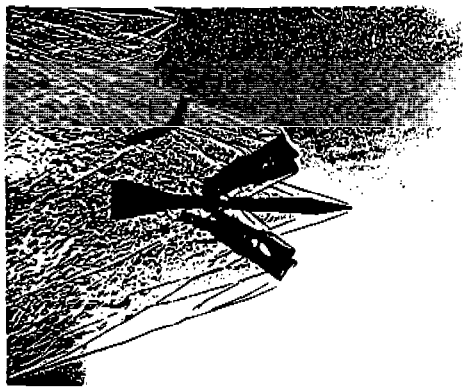


Fig. 1b. Shadowgraph of typical kinetic energy long-rod projectile in free flight during four-petal sabot discard.

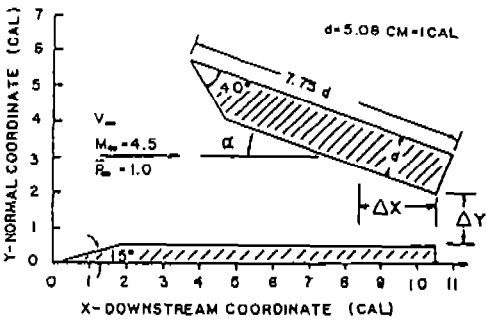


Fig. 2a. Schematic of wind tunnel model in the pitch plane ($\phi = 0, 180°$).

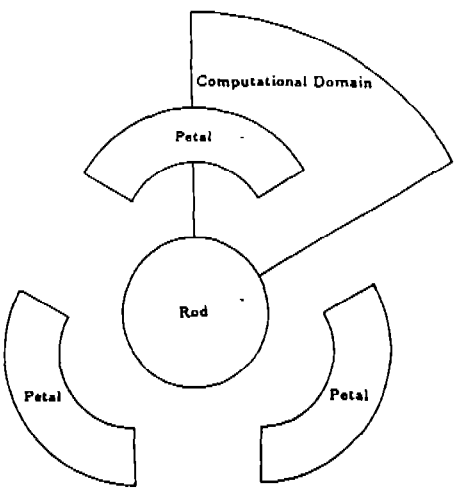


Fig. 2b. Schematic of symmetric sabot discard (rear-view) showing computational domain.

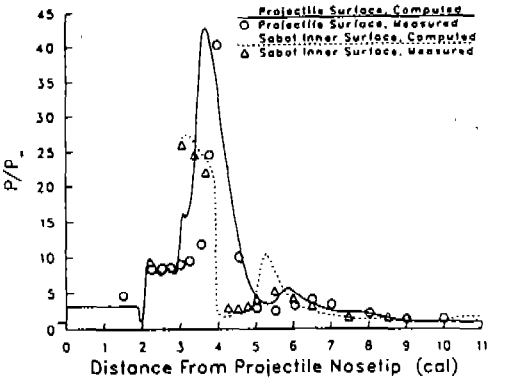


Fig. 3. Laminar flow pressure distributions for projectile and sabot surfaces in the pitch plane ($\phi = 0, 180°$), $\Delta x/D = 0$, $\Delta y/D = .75$, $\alpha = 0°$.

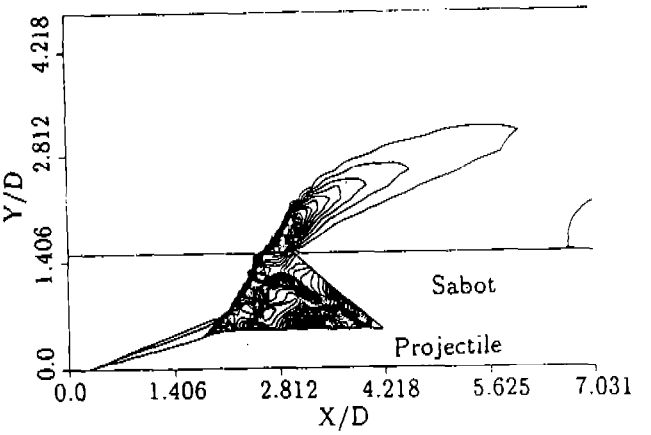


Fig. 4a. Laminar flow pressure contours in the pitch plane ($\phi = 0, 180°$) for $\Delta x/D = 0$, $\Delta y/D = 0.0$, $\alpha = 0°$.

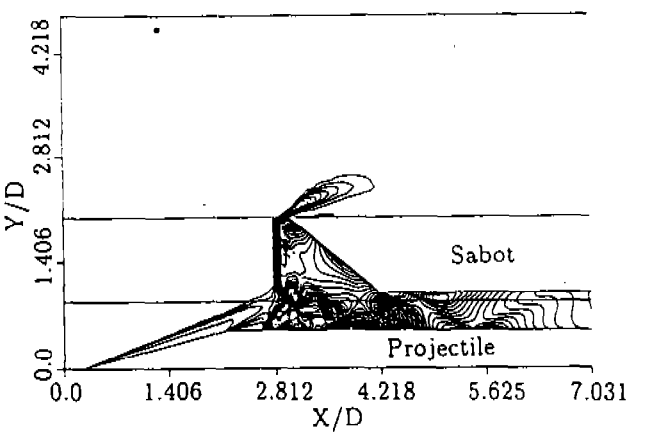


Fig. 4b. $\Delta y/D = .50$, $\alpha = 0°$.

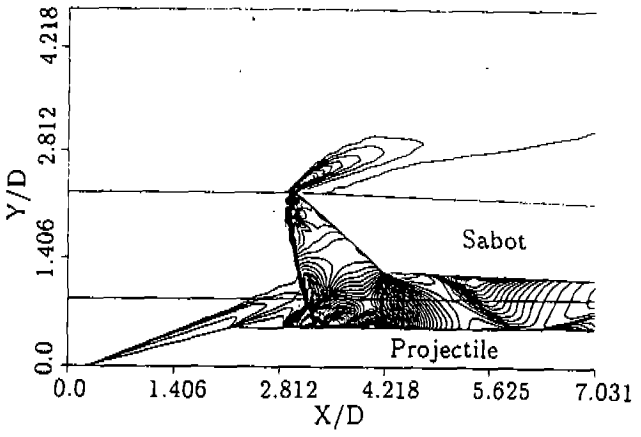


Fig. 4c. $\Delta y/D = .50$, $\alpha = 2^o$.

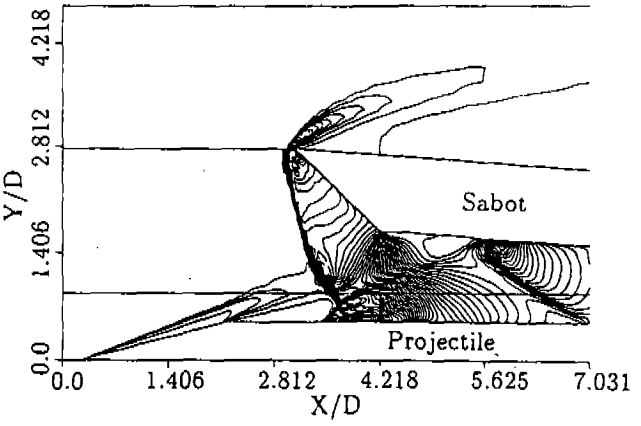


Fig. 4f. $\Delta y/D = .75$, $\alpha = 4^o$.

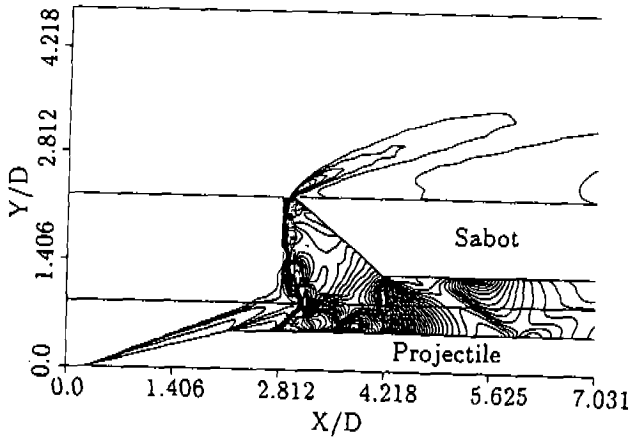


Fig. 4d. $\Delta y/D = .75$, $\alpha = 0^o$.

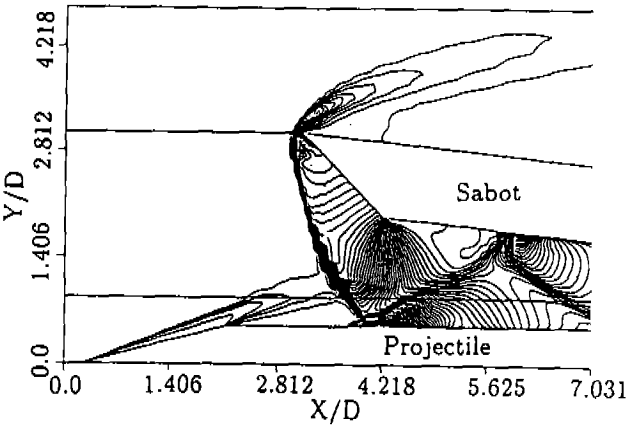


Fig. 4g. $\Delta y/D = .75$, $\alpha = 6^o$.

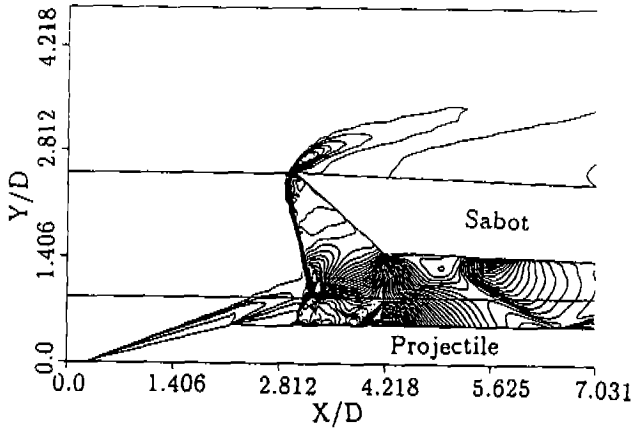


Fig. 4e. $\Delta y/D = .75$, $\alpha = 2^o$.

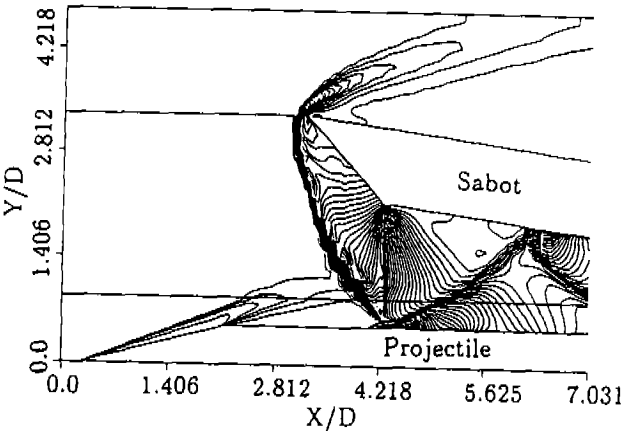


Fig. 4h. $\Delta y/D = .75$, $\alpha = 8^o$.

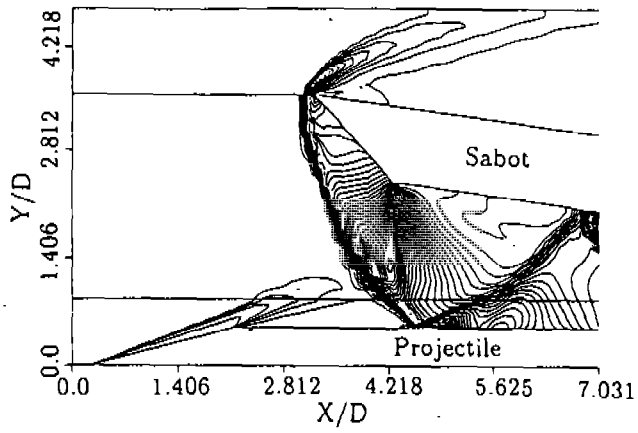


Fig. 4i. $\Delta y/D = 1.0$, $\alpha = 8°$.

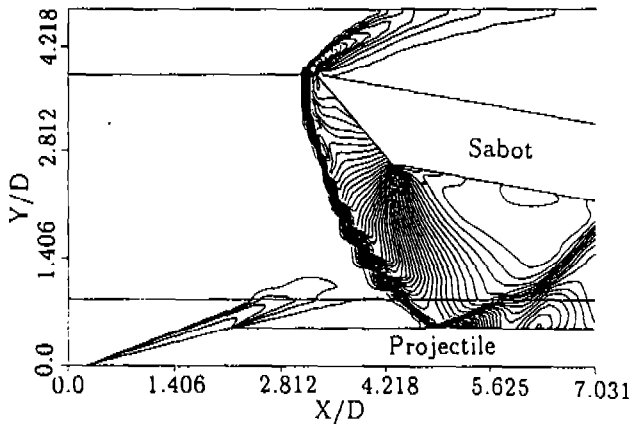


Fig. 4j. $\Delta y/D = 1.0$, $\alpha = 10°$.

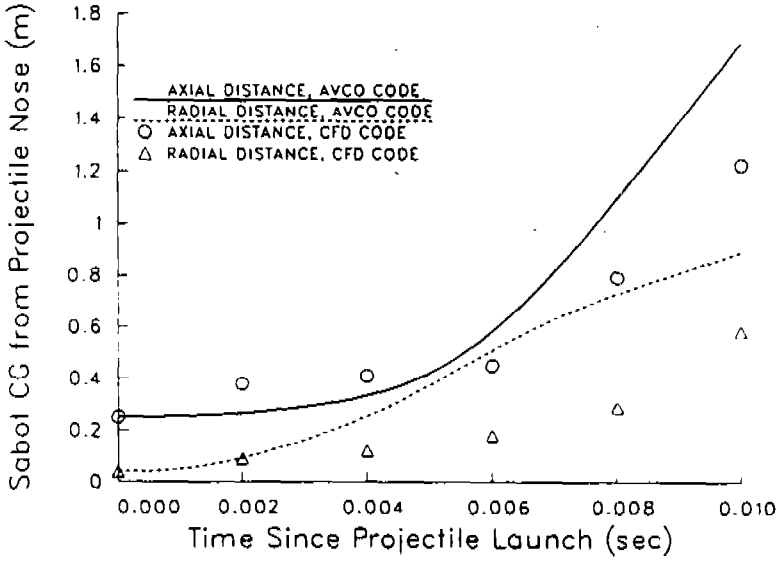


Fig. 5. Trajectory of sabot center of mass computed using AVCO design code and present simulation using CFD.

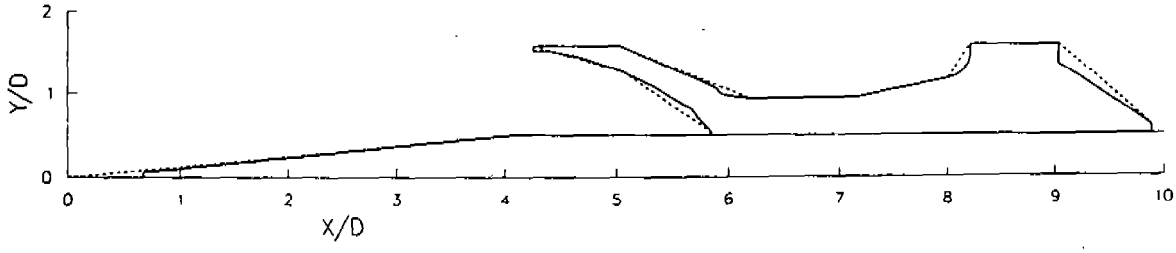


Fig. 6a. M865 projectile/sabot configuration. Solid line is actual geometry. Dashed line is computational geometry.

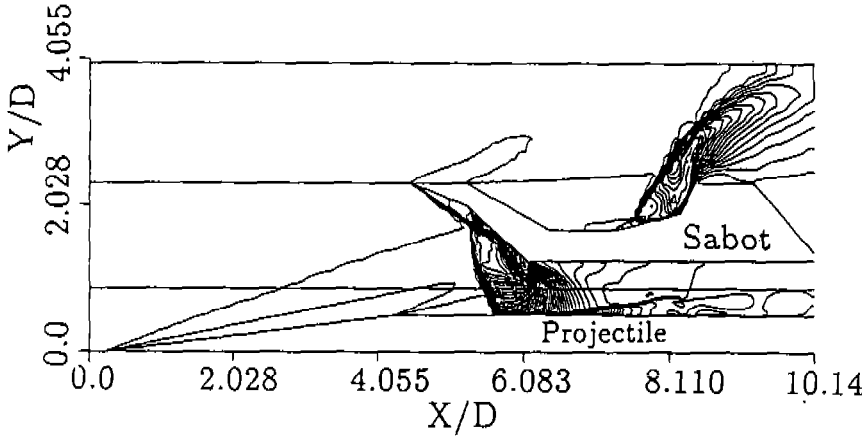


Fig. 6b. Laminar flow pressure contours in the pitch plane ($\phi = 0, 180°$) for M865 sabot, $\Delta x/D = .957$, $\Delta y/D = .75$, $\alpha = 0°$.

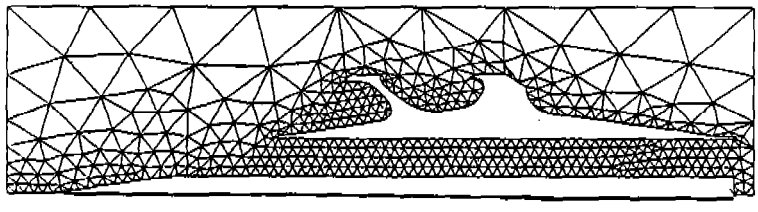


Fig. 7. Unstructured grid (pitch plane view) for M829 projectile/sabot.

Various Finite Difference Schemes for Transient Three Dimensional Heat Conduction

Rao Yalamanchili and Surya R. Yalamanchili*
Light Armament Division
Close Combat Armaments Center
U.S. Army ARDEC
Picatinny Arsenal, NJ 07806-5000

ABSTRACT. The motivation for this task comes from the needs of future hypervelocity projectile surrounded by asymmetric flow due to angle of attack and/or fins in case of kinetic energy projectile. In either case, unsteady and three-dimensional effects, large and nonuniform heat fluxes, tedious and repetetive number crunching capabilities of supercomputers dictate optimum numerical techniques and predictive critical time steps for successful and practical solutions. Finite element modeling is ideal whenever there is geometrical complexity, coatings, composite and multi materials. However, classical finite element technique yields a particular equation. There may be some finite difference schemes superior to classical finite element technique. Therefore, various finite difference schemes are derived and their characteristics are discussed applicable to transient three dimensional heat conduction problems.

INTRODUCTION. Carslaw and Jaeger [1] summarized numerous analytical solutions for rectangular regions, cylinders, and spheres under a variety of initial and boundary conditions. However, if the body has an irregular shape, such as projectile or gun barrel with rifling inside and variable outside diameter, the possibility of obtaining an analytical solution is negligible and in such circumstances one has to rely on numerical methods. Different numerical methods have been used for the solution of transient heat conduction problems. The most popular numerical methods are based on finite element and finite difference techniques. Recently, boundary element techniques are also introduced. Originally, the finite element method (FEM) was introduced as a method of direct structural analysis. Wilson and Nickel [2] applied the finite element method in conjuction with a variational principle derived by Gurtin [3] to solve transient heat conduction problems. This method has many advantages over

---

College of Engineering, Rutgers University, New Brunswick, NJ

other numerical approaches. The FEM is completely general with respect to geometry and material properties. Complex bodies composed of many different anisotropic materials are easily represented. Temperature or heat flux boundary conditions may be specified at any point within the finite element system. Mathematically , it could be shown that the method converges to the exact solution as the number of elements is increased. However, limited use is found to solve transient heat conduction problems with radiation boundary conditions because of difficulties associated with nonlinearity created by radiative heating or cooling phenomena.

Two categories of finite difference equations (FDM) have been employed: The explicit finite difference equations (the temperature at time t is expressed in terms of the temperatures at one time interval, Δt, earlier) and the implicit finite difference equations. They represent a direct approximation approach to the partial differential equation type of formulation. The finite element analysis belongs to the class of implicit technique in finite difference methods. Indeed, Yalamanchili [4,5] proved that finite element and finite difference methods belong to the class of method of weighted residuals, in particular, Galerkin and Collocation methods respectively for transient two dimensional heat conduction problems.

Numerical approximations to solutions of the heat flow equation in three space dimensions may be obtained by the step-wise solution of an associated difference equation. It is the intent of this paper to develop several difference equations that may contain from a minimum of 7 nodes to a maximum of 27 nodal points available in a typical three dimensional element. Of course, the accuracy of these finite difference equations vary by orders of magnitude. However, it is straight forward to generate a system of algebraic equations and to express it in a matrix form for any chosen finite difference scheme. Proper numbering of nodes is essential in order to obtain a feasible solution even though the matrix is sparse due to an exponential increase in arithmetic operations especially for transient three dimensional problems.

LAPLACIAN TERM APPROXIMATIONS. Consider the heat conduction equation in a three dimensional body of length (a), width (b) and height (c) with the following boundary conditions:

$$\frac{\partial T}{\partial t} = \alpha \left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right) \qquad (1)$$

$$T = T(x, y, z, t) = T^{\lambda}_{i,j,k} \qquad (2)$$

$$T(0,y,z,t)=0;\ T(x,0,z,t)=0;\ T(x,y,0,t)=0$$

$$\frac{\partial T}{\partial x}(a,y,z,t)=0;\ \frac{\partial T}{\partial y}(x,b,z,t)=0;\ \frac{\partial T}{\partial z}(x,y,c,t)=0 \qquad (3)$$

For any chosen finite increment Δx, Δy, Δz and Δt in both finite difference and finite element system with a fixed value of $\alpha \Delta t(1/\Delta x^2+1/\Delta y^2+1/\Delta z^2)$, the efforts required to calculate the solution up to a given time is proportional to the number of spatial nodal points raised to the power of three. The number of spatial points changes drastically for multi dimensional problems. Therefore, a variety of finite difference schemes, as well as the stability, accuracy, and oscillation characteristics of three dimensional problems are essential for economical and practical reasons. The following analysis is prepared to fulfill such an objective among other considerations. Rewriting the first time derivative in a finite difference form, the governing equation becomes

$$\frac{T^{+}_{ooo}-T^{o}_{ooo}}{\alpha\,\Delta t}=\phi\,\nabla^2 T^{+}_{ooo}+(1-\phi)\,\nabla^2 T^{o}_{ooo} \qquad (4)$$

Where ϕ is a weighted paramenter with respect to time and varies between 0 and 1 and the Laplacian term, $\nabla^2 T$, is written as

$$\nabla^2 T=\frac{\partial^2 T}{\partial x^2}+\frac{\partial^2 T}{\partial y^2}+\frac{\partial^2 T}{\partial z^2} \qquad (5)$$

Let us now derive second derivative (∂^2) following Strickwerda [6,7]. By the use of Taylor's series (spatial step size = h), one can derive the following:

$$h\delta_{+}=e^{h\partial}-1$$

$$h\delta_{-}=1-e^{-h\partial}$$

$$\delta^2=\delta_{+}\delta_{-}=h^{-2}(e^{h\partial}-1)(1-e^{-h\partial})=\left(\frac{\sinh\frac{1}{2}h\partial}{h/2}\right)^2 \qquad (6)$$

Where δ_{+} and δ_{-} are forward and backward differences respectively. Eq (6) can be written as

$$h\delta=2\sinh\tfrac{1}{2}h\partial \qquad (7)$$

or

$$\partial=\frac{\sinh^{-1}\frac{1}{2}h\delta}{h/2}$$

Second derivatives can be formulated by expansion into series and eventually squaring of eq (7):

$$\begin{aligned}\partial^2 &= \left(1-\tfrac{1}{12}h^2\delta^2+\tfrac{1}{90}h^4\delta^4\right)\delta^2+O(h^6)\\ &= \left(1+\tfrac{1}{12}h^2\delta^2\right)^{-1}\left(1-\tfrac{1}{240}h^4\delta^4\right)^{-1}\delta^2+O(h^6)\end{aligned} \tag{8}$$

Substitution of eq (8) into eq (5) yields the following result:

$$\nabla^2 T=\sum_{i=1}^{3}\left(1+\tfrac{1}{12}h^2\delta_i^2\right)^{-1}\left(1-\tfrac{1}{240}h^4\delta_i^4\right)^{-1}\delta_i^2 T+O(h^6) \tag{9}$$

One may obtain the following equation by clearing out denominators:

$$\nabla^2 T=\sum_{\substack{i=1\\ j\neq i\neq K}}^{3}\left(1+\tfrac{1}{12}h^2\delta_i^2\right)\left(1+\tfrac{1}{12}h^2\delta_K^2\right)\left(1-\tfrac{1}{240}h^4\delta_i^4\right)\left(1-\tfrac{1}{240}h^4\delta_K^4\right)\delta_i^2 T+O(h^6) \tag{10}$$

Further simplification yields the following result:

$$\nabla^2 T=\sum_{\substack{i=1\\ j\neq i\neq K}}^{3}\left(1+\tfrac{1}{12}h^2\delta_j^2+\tfrac{1}{12}h^2\delta_K^2+\tfrac{1}{144}h^4\delta_j^2\delta_K^2\right)\left(1-\tfrac{h^4}{240}\left(\delta_j^4+\delta_K^4\right)\right)\delta_i^2 T+O(h^6) \tag{11}$$

Three finite difference schemes may be formed from eq (11). The simplist and also the least accurate is O (h):

$$\nabla^2 T=\sum_{i=1}^{3}\delta_i^2 T+O(h^2) \tag{12}$$

Substitution of central second difference operator in all three directions yields the following result:

$$\nabla^2 T=\left(T_{-00}+T_{0-0}+T_{00-}-6T_{000}+T_{00+}+T_{0+0}+T_{+00}\right)/h^2 \tag{13}$$

Here, the commas between subscripts are omitted for conciseness. For example, T_{-00} is equivalent to $T_{i-1,j,K}$ in a formal notation. Similarly, T_{0-0} is same as $T_{i,j-1,K}$. Next, another finite difference scheme can be formulated by retaining terms up to $O(h^4)$ from eq (11).

$$\begin{aligned}\nabla^2 T &= \sum_{\substack{i=1\\ j\neq K\neq i}}^{3}\left(1+\frac{h^2}{12}\delta_j^2+\frac{h^2}{12}\delta_K^2\right)\delta_i^2 T+O(h^4)\\ &= \sum_{\substack{i=1\\ j>i}}^{3}\left(1+\frac{h^2}{6}\delta_j^2\right)\delta_i^2 T+O(h^4)\end{aligned} \tag{14}$$

One can write eq (14) in a finite difference form if we can define terms, such as $\delta_x^2\delta_y^2 T$. This term can be written as

$$\delta_x^2\delta_y^2 T = \delta_x^2\left(T_{0+0} - 2T_{000} + T_{0-0}\right)/h^2$$

$$= \left[\left(T_{++0} - 2T_{0+0} + T_{-+0}\right) - 2\left(T_{+00} - 2T_{000} + T_{-00}\right) + \left(T_{+-0} - 2T_{0-0} + T_{--0}\right)\right]/h^4 \qquad (15)$$

Substitution of eq (15) and similar results of other terms into eq (14) yields the following 19-point finite difference formula for Laplacian, $\nabla^2 T$:

$$\nabla^2 T = \left(T_{--0} + T_{-0-} + 2T_{-00} + T_{-0+} + T_{-+0} + T_{0--} + 2T_{0-0} + T_{0-+} + 2T_{00-} - 24T_{000} + 2T_{00+} + T_{0+-} + 2T_{0+0} + T_{0++} + T_{+0-} + 2T_{+00} + T_{+0+} + T_{++0}\right)/6h^2 \qquad (16)$$

The accuracy is $O(h^4)$. Another most accurate formula for Laplacian term can be formulated by retaining terms up to $O(h^6)$ in eq (11):

$$\nabla^2 T = \sum_{i=1}^{3}\left(1 + \frac{h^2}{6}\delta_j^2\right)\delta_i^2 T + \frac{h^4}{144}\sum_{\substack{i=1\\ j\neq k\neq i}}^{3}\delta_j^2\delta_k^2\delta_i^2 T - \frac{h^4}{240}\sum_{\substack{i=1\\ j\neq k\neq i}}^{3}\left(\delta_j^4 + \delta_k^4\right)\delta_i^2 T + O(h^6) \qquad (17)$$

$$= \sum_{\substack{i=1\\ j>i}}^{3}\left(1 + \frac{h^2}{6}\delta_j^2\right)\delta_i^2 T + \frac{h^4}{48}\prod_{i=1}^{3}\delta_i^2 T + \frac{h^4}{80}\prod_{i=1}^{3}\delta_i^2 T + O(h^6)$$

$$= \sum_{\substack{i=1\\ j>i}}^{3}\left(1 + \frac{h^2}{6}\delta_j^2\right)\delta_i^2 T + \frac{h^4}{30}\prod_{i=1}^{3}\delta_i^2 T + O(h^6) \qquad (18)$$

Where $\prod$ is the product symbol. The last term alone contains all 27 nodal points if expanded in a manner similar to eq (15). The final result of the last term is as follows:

$$\frac{h^4}{30}\prod_{i=1}^{3}\delta_i^2 T = (T_{---} - 2T_{--0} + T_{--+} - 2T_{-0-} + 4T_{-00} - 2T_{-0+} + T_{-+-}$$

$$-2T_{-+0} + T_{-++} - 2T_{0--} + 4T_{0-0} - 2T_{0-+} + 4T_{00-} - 8T_{000} + 4T_{00+}$$

$$-2T_{0+-} + 4T_{0+0} - 2T_{0++} + T_{+--} - 2T_{+-0} + T_{+-+} - 2T_{+0-} + 4T_{+00}$$

$$-2T_{+0+} + T_{++-} - 2T_{++0} + T_{+++}) / 30h^2 \qquad (19)$$

Substitution of eq (19) and eq (16) into eq (18) yields the following most accurate ($O(h^6)$) finite difference approximation for the Laplacian term

$$\nabla^2 T = (T_{---} + 3T_{--0} + T_{--+} + 3T_{-0-} + 14T_{-00} + 3T_{-0+} + T_{-+-}$$

$$+3T_{-+0} + T_{-++} + 3T_{0--} + 14T_{0-0} + 3T_{0-+} + 14T_{00-} - 128T_{000}$$

$$+14T_{00+} + 3T_{0+-} + 14T_{0+0} + 3T_{0++} + T_{+--} + 3T_{+-0} + T_{+-+}$$

$$+3T_{+0-} + 14T_{+00} + 3T_{+0+} + T_{++-} + 3T_{++0} + T_{+++}) / 30h^2 \qquad (20)$$

As before, the subscripts i,j and k, and commas are omitted. Same notation, as explained above, is used. For example,

$T_{+++} = T_{i+1,j+1,k+1}$.

So far, three finite difference schemes are derived for the Laplacian term, i.e., eq (13), eq (16), and eq (20) with orders of accuracy $O(h^2)$, $O(h^4)$, and $O(h^6)$ respectively. However, it doesn't seem to be enough variety to compare especially finite difference and finite element schemes. Towards this goal, another finite difference scheme for three dimensional laplacian term is derived. This one also contain all 27 nodal points available in a typical three dimensional finite element by combination of nine rays (lines) passing through central node (i,j,k) and central second derivatives along those rays. The resulting finite difference approximation for the Laplacian term is given below. However, the order of accuracy is $O(h^2)$ far lower than its 27-node counterpart.

$$\nabla^2 T = (T_{---} + T_{--0} + T_{--+} + T_{-0-} + T_{-00} + T_{-0+} + T_{-+-} + T_{-+0} + T_{-++}$$

$$+T_{0--} + T_{0-0} + T_{0-+} + T_{00-} - 26T_{000} + T_{00+} + T_{0+-} + T_{0+0} + T_{0++} \qquad (21)$$

$$+T_{+--} + T_{+-0} + T_{+-+} + T_{+0-} + T_{+00} + T_{+0+} + T_{++-} + T_{++0} + T_{+++})/9h^2$$

FINITE DIFFERENCE SHEMES. Until now, a variety of finite difference shemes are derived for three dimensional Laplacian term in order to obtain not only more accurate numerical solutions but also to unify and compare finite element and finite difference techniques. In general, it is understood that the higher order scheme yields more accurate solution than a lower order scheme. The accuracy of the numerical results can also be improved by reducing the grid spacing, h. Usually, grid spacing reduction improves the numerical results up to a certain extent. At this point, the numerical resutls are as accurate as can be with the chosen order of finite difference scheme. Further reduction in grid spacing will lead to increasing round off errors due to enormous increase in number of computations and thus, overall, less accurate results. However, the accuracy of the results can be improved by the use of higher order finite difference scheme.

The accuracy of a numerical solution may also be improved by proper selection of the weighted parameter, ϕ, introduced in eq (4). This parameter also plays a major role in stability and oscillation characteristics of a numerical scheme. The parameter $\phi (0 \leq \phi \leq 1)$ allows a weighted average of sum of three second order spatial derivatives at two discrete times (old and new). An explicit scheme is the result when ϕ is set to zero; otherwise, an implicit scheme will be the result for remaining range of parameter, ϕ.

The generic Laplacian is represented by equations (13), (16), (20) and (21). Appropriate time superscripts (o = old time, + = new time) have to be introduced into Laplacian finite difference approximations before substitution of equations (13), (16), (20) and (21) into eq (4) in order to obtain various finite difference schemes. The format of finite difference equation for an unsteady three dimensional problem is shown in eq (22).

$$
\begin{aligned}
&AT^{+}_{---}+BT^{+}_{--0}+AT^{+}_{--+}+BT^{+}_{-0-}+CT^{+}_{-00}+BT^{+}_{-0+}+AT^{+}_{-+-}+BT^{+}_{-+0}+AT^{+}_{-++}\\
&+BT^{+}_{0--}+CT^{+}_{0-0}+BT^{+}_{0-+}+CT^{+}_{00-}+DT^{+}_{000}+CT^{+}_{00+}+BT^{+}_{0+-}+CT^{+}_{0+0}+BT^{+}_{0++}\\
&+AT^{+}_{+--}+BT^{+}_{+-0}+AT^{+}_{+-+}+BT^{+}_{+0-}+CT^{+}_{+00}+BT^{+}_{+0+}+AT^{+}_{++-}+BT^{+}_{++0}+AT^{+}_{+++}\\
&=ET^{o}_{---}+FT^{o}_{--0}+ET^{o}_{--+}+FT^{o}_{-0-}+GT^{o}_{-00}+FT^{o}_{-0+}+ET^{o}_{-+-}+FT^{o}_{-+0}+ET^{o}_{-++}+\\
&FT^{o}_{0--}+GT^{o}_{0-0}+FT^{o}_{0-+}+GT^{o}_{00-}+HT^{o}_{000}+GT^{o}_{00+}+FT^{o}_{0+-}+GT^{o}_{0+0}+FT^{o}_{0++}+ET^{o}_{+--}\\
&+FT^{o}_{+-0}+ET^{o}_{+-+}+FT^{o}_{+0-}+GT^{o}_{+00}+FT^{o}_{+0+}+ET^{o}_{++-}+FT^{o}_{++0}+ET^{o}_{+++}
\end{aligned}
\qquad (22)
$$

However, the coefficients are different. These are given in Table 1 for all Laplacian term approximations L13, L16, L20 and L21 discussed above. L13, L16, L20 and L21 are named after equations (13, (16), (20) and (21) respectively. θ is the welknown dimensionless Fourier number ($\alpha \Delta t/h^2$). The quantity on the right hand side of eq (22) is known due to known nodal temperatures at the old time. One can generate a system of equations, one at each interior node. Even if one divides the body into 11 equal parts in each direction, 1000 equations with 1000 unknows will be generated. However, there are atmost 27 unknowns in each equation. Therefore, a sparse matrix is generated. Special numbering of nodes yields a minimum bandwidth for nonzero terms. Sparse matrix algorithms that take advantage of minimum bandwidth, storage and efficint computations are available in the literature for its solution. The coefficient A is associated with 8 corner nodal temperatures whereas the coefficient B is associated with 12 edge nodal temperatures. Similarly, the coefficient C is connected with 6 face center nodal temperatures. The coefficient D exists only with one central nodal temperature. The coefficients E, F, G, and H are associated with the same nodal temperatures as the coefficients A, B, C, and D respectively.

TABLE 1 COEFFICIENTS OF VARIOUS FINITE DIFFERENCE EQUATIONS

FD Scheme by Laplacian	COEFFICIENTS							
	A	B	C	D	E	F	G	H
L13 (7 nodes)	θ	0	$\phi\theta$	$1+6\phi\theta$	0	0	$-\theta(1-\phi)$	$1-6\theta(1-\phi)$
L16 19 nodes	0	$\frac{\phi\theta}{6}$	$\frac{2}{6}\phi\theta$	$1+4\phi\theta$	0	$-\frac{\theta}{6}(1-\phi)$	$-\frac{2\theta}{6}(1-\phi)$	$1-4\theta(1-\phi)$
L20-$O(h^6)$ 27 nodes	$\frac{\phi\theta}{30}$	$\frac{3}{30}\phi\theta$	$\frac{14}{30}\phi\theta$	$1+\frac{128}{30}\phi\theta$	$-\frac{\theta}{30}(1-\phi)$	$-\frac{3\theta}{30}(1-\phi)$	$-\frac{14}{30}\theta(1-\phi)$	$1-\frac{125}{30}\theta(1-\phi)$
L21-$O(h^2)$ 27 nodes	$\frac{\phi\theta}{9}$	$\frac{\phi\theta}{9}$	$\frac{\phi\theta}{9}$	$1+\frac{26}{9}\phi\theta$	$-\frac{\theta}{9}(1-\phi)$	$-\frac{\theta}{9}(1-\phi)$	$-\frac{\theta}{9}(1-\phi)$	$1-\frac{26}{9}\theta(1-\phi)$

OTHER FINITE DIFFERENCE SCHEMES. It is obvious, by now, that one is confornted with a large system of equations for multi dimensional problems and its numerical solution is expensive either by the use of direct or indirect (iterative) methods of system of equations generated by implicit techniques. The utilization of explicit techniques is limited due to small time step requirements in order to enforce stability.

Several methods that employ the useful characteristics of both implicit and explicit methods are also developed. These are becoming the most popular techniques for solving parabolic partial differential equations, such as transient heat conduction in a multi dimensional environment. In particular, the alternating direction implicit (ADI) method is ideal for solving a two dimensional problem. This method uses the implicit formulation in one direction and considers the other direction explicitly. The two directions are interchanged from one time step to the next time step. This results in a simple tridiagonal system of equations even for a two-dimensional problem as in a one-dimensional problem. The ADI method may belong to the class of splitting methods.

The Crank-Nicolson (CN) implicit scheme, equals one-half, is mentioned extensively in the literature for the solution of transient heat conduction problems. The Douglas scheme is not mentioned that much. However, the following expamle (Table 2) shows that Douglas scheme is better then Crank-Nicolson scheme.

Table 2. Comparison of Douglas & CN ($t=0.1, \theta=1$)

X=	0.1	0.2	0.3	0.4	0.5
Exact	0.0934	0.1776	0.2444	0.2873	0.3021
Douglas	0.0941	0.1789	0.2463	0.2895	0.3044
CN	0.0948	0.1803	0.2482	0.2918	0.3069

One can improve the stability of classical explicit finite difference technique ($\theta \leq 1/6$) by the following equation ($\theta \leq 1/2$) for a three dimensional problem:

$$T^{+} = \prod_{i=1}^{3} (1 + \theta h^2 \delta_i^2) T^{0} \qquad (23)$$

Here, computations are required at all 27 grid points at the old time level. However, it is more economical if it is used in the following split form:

$$\begin{aligned} T^{*} &= (1+\theta h^2 \delta_3^2) T^{0} \\ T^{**} &= (1+\theta h^2 \delta_2^2) T^{*} \\ T^{+} &= (1+\theta h^2 \delta_1^2) T^{**} \end{aligned} \qquad (24)$$

Here, the problem is how to satisfy the requirement of

intermediate boundary conditions within the time step. Therefore, explicit difference methods are rarely used to solve initial boundary value problems in three dimensional problems. More often, ADI and locally-one-dimensional (LOD) schemes are used instead of explicit methods.

The Douglas-Rachford [8] ADI scheme for a three dimensional case can be written as

$$\begin{aligned}(1-\theta h^2\delta_1^2)T^* &= (1+\theta h^2(\delta_2^2+\delta_3^2))T^\circ \\ (1-\theta h^2\delta_2^2)T^{**} &= T^*-\theta h^2\delta_2^2 T^\circ \\ (1-\theta h^2\delta_3^2)T^+ &= T^{**}-\theta h^2\delta_3^2 T^\circ\end{aligned} \qquad (25)$$

The following ADI scheme is more accurate depending on parameters, γ and β :

$$\prod_{i=1}^{3}(1-\gamma h^2\delta_i^2)T^+ = \prod_{i=1}^{3}(1+\beta h^2\delta_i^2)T^\circ \qquad (26)$$

Of course, this can be split into three equations involving the solution of only tridiagonal system of equations along x, y, and z at first, second and third steps, respectively. It is better to experiment various schemes not only mentioned above but also available elsewhere and decide the most appropriate one based on accuracy, stability and computer time due to need of repetetive and intense computations for a transient three dimensional problem. Keep in mind, practicality and boundary conditions in a final showdown.

CONCLUSIONS: The simulation of hypervelocity projectile, in-flight, involves not only computational fluid dynamic study around the projectile but also heat transfer in the projectile. It is highly desirable to couple the two problems whenever feasible. Derived various finite difference approximations, ranging in accuracy $O(h^2)$ to $O(h^6)$, for Laplacian term in three dimensions. Constructed numerous finite difference formulas, both explicit and implicit, for the solution of transient three dimensional heat conduction problems. One of the finite difference formula is found to be equivalent to classical finite element scheme. However, it is not proved here due to space and time limitations. One of the numerical example indicates that the Douglas scheme is superior to Crank -Nicolson scheme. Also, other economical schemes such as split methods, alternating direction implicit, locally one dimensional and explicit methods are briefly tonched due to practical considerations. However, limited experimentation is desirable based on a given problem and boundary conditions.

ACKNOWLEDGEMENT. Grateful acknowledgement is given to Prof. John C. Strikwerda, University of Wisconsin, Madison, for clever manipulations of sixth order accurate finite difference approximation of Poisson equation.

REFERENCES

1. Carslaw, H.S., and J.C. Jaeger, "Conduction of Heat in Solids", Oxford University Press (1959).
2. Wilson, E.L., and R.E. Nickell, "Application of the Finite Element Method to Heat Conduction Analysis", Nuclear Engineering and Design, Vol. 4 pp. 276-286 (1966).
3. Gurtin, M.E., "Variational Principles for Linear Initial Value Problems," Quarterly Applied Mathematics, Vol. 2, pp. 252-256 (1964).
4. Yalamanchili, R.V.S., and S.C. Chu, "Stability and Oscillation Characteristics of Finite Element, Finite Difference, and Weighted Residual Methods for Transient Two Dimensional Heat Conduction in Solids," Journal of Heat Transfer, Trans. of ASME, Vol. 95, Series C, #2, (1973).
5. Yalamanchili, R., "Accuracy, Stability, and Oscillation Characteristics of Transient Two Dimensional Heat Conduction,: ASME Paper # 75-WA/HT-85 (1975).
6. Strikwerda, J.C., "Finite Difference Schemes and Partial Differential Equations," Wadsworth, Inc. (1989).
7. Strikwerda, J.C., Private Communications, Univ. of Wisconsin, Madison (Oct. 1990).
8. Douglas, A., and H.H. Rackford, "On the Numerical Solution of Heat Conduction Problems in Two and Three Space Variables," Trans. of American Mathematical Society, Vol 82, pp. 421-39.

High performance Simplification-Based Automated Deduction *

Maria Paola Bonacina **Jieh Hsiang**
Department of Computer Science
SUNY at Stony Brook
Stony Brook, NY 11794-4400
{bonacina,hsiang}@sbcs.sunysb.edu

1 Introduction

Equational logic is one of the most important domains of research in computer science. Specifications of types of data structures and assertions about the behaviour of programs are naturally written in equational form. Programs made of equations are called *equational programs* and appear in functional programming, logic programming and in most combinations of high level programming paradigms [19, 23]. First order logic can be expressed equationally [20]. This formulation makes it possible to express logic programming equationally and to employ the computational model of equational languages in logic programming [7]. Set theory can also be expressed equationally [33], enabling one to reason about query languages and optimization in data bases [11].

Such a wide range of applications, not to mention the traditional applications to algebra, makes automated deduction in equational logic an important subject of research. However, the seemingly insurmountable search space caused by the symmetry and replacement properties of the equality predicate had been a serious obstacle which baffled researchers in automated deduction for several decades. It is not until very recently that methods capable of effectively reason with equations have been designed and successfully applied to an interesting range of challenging problems. These methods are based on the *term rewriting* approach to equational reasoning, which was started in [24].

The key idea in term rewriting based theorem proving is to regard a derivation as a process of *proof reduction.* Equations are oriented into rules according to a *well-founded ordering*, and equational replacement is performed only in one direction. When an expression (term, equation, clause) is *simplified* by a rule, the old expression is discarded and replaced by the new one, which is smaller in the ordering. The generation of new lemmas, the *superposition process*, is also done according to the ordering. By keeping every piece of data fully simplified at all time, the search space is *drastically* reduced.

Section 2 presents in greater detail the simplification-based theorem proving approach, according to the theoretical framework which we have proposed in [8]. Section 3 describes our theorem prover *SBR3*, which implements the simplification-based methodology. Section 4 relates some original proofs obtained automatically with *SBR3*. The last section is devoted to some discussion on our current work on *distributed theorem proving.*

*Research supported in part by grants CCR-8805734 and CCR-8901322, funded by the National Science Foundation. The first author is also supported by a scholarship of Universitá degli Studi di Milano, Italy.

2 Simplification-based automated deduction

A *theorem proving problem* consists in finding a proof of a given sentence φ in a given set of axioms S. The set S is a *presentation* of the theory $Th(S)$ of all the theorems of S, $Th(S) = \{\psi \mid S \models \psi\}$. For instance, in equational logic, S is a set of equations E, the axioms for an equational theory. The sentence φ to be proved is the *target* or *goal*. In equational theorem proving, the target is an equation $\forall \bar{x}\, s \simeq t$, where all variables are universally quantified. We write $(S; \varphi)$ to denote the problem of proving φ from S.

The first component of a *theorem proving strategy* $\mathcal{C}$ is a set I of *inference rules*. An application of an inference rule to $(S; \varphi)$ transforms it into another problem: $(S; \varphi) \vdash_I (S'; \varphi')$. Clearly, the two problems must be equivalent. This is ensured by requiring that for all inference steps $(S; \varphi) \vdash_I (S'; \varphi')$, the theory of S' is not larger than the theory of S, i.e. $Th(S') \subseteq Th(S)$, and $\varphi \in Th(S)$ if and only if $\varphi' \in Th(S')$. We have termed these two properties *monotonicity* and *relevance* respectively.

An inference mechanism I defines for every given input $(S_0; \varphi_0)$ the *space* of all the problems or *states* $(S; \varphi)$, which can be derived from $(S_0; \varphi_0)$ by I in zero or more steps. This space can be represented as a tree, where the nodes are labeled by pairs $(S; \varphi)$, the root is labeled by $(S_0; \varphi_0)$ and there is an arc from node $(S; \varphi)$ to node $(S'; \varphi')$ if and only if $(S; \varphi) \vdash_I (S'; \varphi')$. We call this tree the *I-tree rooted at* $(S_0; \varphi_0)$, because it is determined by the inference mechanism I and the input problem $(S_0; \varphi_0)$. Accordingly, a sequence of inference steps $(S; \varphi) \vdash_I (S'; \varphi')$ is an *I-path*. In general, the I-tree is a directed graph, rather than a tree, since a node $(S; \varphi)$ may be reachable starting from the root by more than one I-path. However, it is always possible to transform it into a tree by allowing different nodes to have the same label.

If φ_0 is indeed a theorem of S_0, i.e. $\varphi_0 \in Th(S_0)$, the inference mechanism I should be able to prove it. This is the intuitive meaning of the *refutational completeness* of an inference system. Refutational completeness can be described on the I-tree as follows: I is refutationally complete if and only if, whenever $\varphi_0 \in Th(S_0)$, the I-tree rooted at $(S_0; \varphi_0)$ contains at least a node labeled by a *successful state* $(S; true)$, i.e. a state where the target is proved.

If our theorem proving strategy $\mathcal{C}$ has a refutationally complete inference mechanism, we know that for every true input target, we can find a proof. However, ensuring that the inference rules are sufficiently powerful to prove all theorems is just the beginning. We now face the problem of *searching* the I-tree to reach a solution. Thus, the second component of a strategy $\mathcal{C}$ is a *search plan* Σ: $\mathcal{C} =< I; \Sigma >$. Given the input state $(S_0; \varphi_0)$, Σ selects an inference rule f in I and a tuple of premises $\bar{x}$ in $S_0 \cup \{\varphi_0\}$. The first step consists then in applying f to $\bar{x}$, generating a new state $(S_1; \varphi_1)$. Choosing an inference step corresponds to choosing one of the arcs leaving node $(S_0; \varphi_0)$ in the I-tree. The process is repeated, generating a *derivation*

$$(S_0; \varphi_0) \vdash_{\mathcal{C}} (S_1; \varphi_1) \vdash_{\mathcal{C}} \ldots (S_i; \varphi_i) \vdash_{\mathcal{C}} \ldots,$$

where at each step an inference is performed according to the search plan. The derivation computed by $\mathcal{C}$ on input $(S_0; \varphi_0)$ is the unique I-path selected by Σ in the I-tree rooted at $(S_0; \varphi_0)$. A derivation is successful if it reaches a successful node $(S; true)$.

The refutational completeness of I guarantees that successful derivation exist. We need another property to ensure that the specific derivation computed by $\mathcal{C}$ is successful. This property is the *fairness* of the search plan: Σ is fair if and only if, whenever the I-tree rooted at $(S_0; \varphi_0)$ contains successful nodes, the derivation controlled by Σ finds one. The refutational completeness of the inference rules and the fairness of the search plan together imply the *completeness* of the strategy $\mathcal{C}$: whenever $\varphi_0 \in Th(S_0)$, the computation by $\mathcal{C}$ halts

successfully. In other words, C is a *semidecision procedure* for theorem proving.

Meeting the completeness requirement alone is not difficult. Many refutationally complete inference systems are known and a search plan which tries exhaustively all steps is trivially fair. The more challenging question of automated deduction is to obtain a strategy which is both complete and *efficient*: not only should the strategy succeed, but it should also do it by consuming "reasonable" amounts of resources, i.e. time and memory. The notion of efficiency is clearly not an absolute one. Rather, it can be used for comparisons. Informally, given two complete strategies C_1 and C_2, a problem $(S_0; \varphi_0)$ and a fixed amount of memory (elapse of time), C_1 is more efficient in time (in memory) than C_2 on problem $(S_0; \varphi_0)$, if it reaches a solution in shorter time (using a smaller amount of memory).

The issue of efficiency can, and in fact should, be considered at both the inference level and the search level. At the inference level, the goal is to devise inference mechanisms which generate "small" search spaces, while preserving refutational completeness. It is desirable that the search space is small, since searching a small space is intuitively easier than searching a large one, but not at the expense of losing all the solution nodes! Similarly, at the search level, the goal is to design search plans which find "fast" solutions, while preserving fairness.

We attack these problems as follows. We have seen that a theorem proving derivation transforms a theorem proving problem into equivalent problems. Intuitively, it is desirable that a problem is reduced to one which is in some sense "smaller". In fact, at the end of a successful derivation we have a solved problem $(S; true)$, where the dummy target "true" simply indicates that the original target has been proved. Thus, we need to identify what is being reduced during a theorem proving derivation. We observe that if a target φ_0 is indeed a theorem of the input set S_0, then there exist some proofs of φ_0 in S_0. On the other hand, the proof of the dummy target "true" is *empty*. At any stage $(S_i; \varphi_i)$ in between there is a (non-unique) *minimal* proof of φ_i in S_i, which represents the least amount of work which still needs to be done in order to prove φ_i from S_i. If the derivation gets closer to a solution, a minimal proof of the target gets reduced, i.e. the amount of work which is left becomes smaller. When the problem is solved, no more work needs to be done. Therefore, we regard theorem proving as *reduction of a minimal proof of the target* to the empty proof.

In order to compare proofs and to have a notion of minimal proofs, we need an *ordering* of proofs. Furthermore, this ordering needs to be *well founded*, having as bottom element the empty proof. A notion of well founded orderings on proofs, called *proof orderings*, has been introduced in [5, 6] and used to prove that Knuth-Bendix type completion procedures generate confluent systems of rewrite rules [16]. We use the same notion for a different purpose. Given a proof ordering $>_p$, at each stage $(S_i; \varphi_i)$ of a derivation, we consider the set $\Pi(S_i, \varphi_i)$ of the *minimal* proofs of φ_i in S_i, according to the ordering $>_p$. A successful derivation progressively reduces a proof in $\Pi(S_i; \varphi_i)$ to the empty proof.

This view has several advantages, both theoretical and practical. On the theoretical side, it has allowed us to give a coherent mathematical foundation to theorem proving. All concepts in theorem proving are defined and related to each other by using proof orderings and proof reduction with respect to such orderings. For instance, the above informal notions of refutational completeness and fairness can be formalized in terms of proof reduction [8, 9].

On the practical side, we require that the inference rules are *proof-reducing*. As we derive $(S_{i+1}; \varphi_{i+1})$ from $(S_i; \varphi_i)$, the set $\Pi(S_i, \varphi_i)$ is replaced by $\Pi(S_{i+1}, \varphi_{i+1})$. Clearly, we need to forbid all inference steps which would replace a proof P in $\Pi(S_i, \varphi_i)$ by a proof Q in $\Pi(S_{i+1}, \varphi_{i+1})$ such that $Q >_p P$. Such steps certainly do not help. On the other hand, we cannot impose that at every step a minimal proof of the target

be reduced. This is impossible, since theorem proving is a process of search and therefore many steps generally do not contribute to the final result. We require that for every step $(S_i;\varphi_i)\vdash_{\mathcal{C}}(S_{i+1};\varphi_{i+1})$, every proof P in $\Pi(S_i,\varphi_i)$ is either preserved, i.e. P is also in $\Pi(S_{i+1},\varphi_{i+1})$, or reduced, i.e. P is replaced by a proof Q in $\Pi(S_{i+1},\varphi_{i+1})$ such that $Q <_p P$. This condition is still not sufficiently general, since inference steps may not affect immediately any minimal proof of the target and still be necessary to prove it eventually. Therefore, we need to extend our attention to a larger set of theorems, which we call the *domain* $\mathcal{T}$ of the derivation. A step $(S_i;\varphi_i)\vdash_{\mathcal{C}}(S_{i+1};\varphi_{i+1})$, such that $\Pi(S_i,\varphi_i)=\Pi(S_{i+1},\varphi_{i+1})$, is also proof-reducing, provided that for all ψ in $\mathcal{T}$, every minimal proof is either preserved or reduced and for at least a ψ in $\mathcal{T}$ a minimal proof is reduced. Intuitively, we would like the domain $\mathcal{T}$ to be as small and as "related" to the target as possible. In practice, for the known simplification-based strategies, the domain is the set of all ground equations.

2.1 The simplification-based inference engine UKB

The most significant characteristic of inference rules in simplification-based strategies is that they are proof-reducing [8]. As an example, we present in the following the ones which are used in our prover *SBR3*, an automated deduction system for equational theories. Collectively, they form the **unfailing Knuth-Bendix completion procedure**, or **UKB** for short. UKB is a semi-decision procedure for the validity problem of equational theory.

The most important one is *Simplification* [28] itself. If we consider a derivation in equational logic, a presentation is a set of equations E and a target is an equational theorem $\forall\bar{x}\, s\simeq t$. We write the target as $\hat{s}\simeq\hat{t}$ to denote that it contains only universally quantified variables and therefore can be regarded as a ground equality. The definition of simplification involves two orderings. The first one is a *well founded* ordering on terms $\succ$ which is used to ensure that simplification replaces an equation by a smaller equation [15]. The second one is the *encompassment* ordering $\trianglerighteq$ which is defined as follows: $t\trianglerighteq s$ if $t|u = s\sigma$ for some position u and substitution σ, i.e. an instance $s\sigma$ of s occurs as a subterm in t. We write $t\triangleright s$ if $t\trianglerighteq s$ and either u is not the root position or σ is not just a renaming of variables [16].

Simplification applies to the presentation:

$$\frac{(E\cup\{p\simeq q, l\simeq r\};\hat{s}\simeq\hat{t})}{(E\cup\{p[r\sigma]_u\simeq q, l\simeq r\};\hat{s}\simeq\hat{t})}\quad \begin{array}{l} p|u = l\sigma \qquad p\succ p[r\sigma]_u \\ p\triangleright l\vee q\succ p[r\sigma]_u\end{array}$$

and to the target:

$$\frac{(E\cup\{l\simeq r\};\hat{s}\simeq\hat{t})}{(E\cup\{l\simeq r\};\hat{s}[r\sigma]_u\simeq\hat{t})}\quad \begin{array}{l}\hat{s}|u = l\sigma \\ \hat{s}\succ\hat{s}[r\sigma]_u.\end{array}$$

Intuitively, a simplification step replaces an equation by a smaller equation and therefore it reduces all the proofs where the replaced equation occurred.

The second basic inference rule, a deductive inference rule called **Superposition** [21], is also proof-reducing:

$$\frac{(E\cup\{p\simeq q, l\simeq r\};\hat{s}\simeq\hat{t})}{(E\cup\{p\simeq q, l\simeq r, p[r]_u\sigma\simeq q\sigma\};\hat{s}\simeq\hat{t})}\quad \begin{array}{l} p|u\notin X \qquad (p|u)\sigma = l\sigma \\ p\sigma\not\preceq q\sigma, p[r]_u\sigma\end{array}$$

where X is the set of variables and σ is the most general unifier of $(p|u)$ and l. The key point is that the step is performed only if $p\sigma\not\preceq q\sigma$ and $p\sigma\not\preceq p[r]_u\sigma$. This conditions guarantee that the rule is proof-reducing.

An operator f is said to satisfy the *right cancellation law* if for every x,y,z, $f(x,z)=f(y,z)$ implies $x=y$. The *left cancellation law* is defined symmetrically. Cancellation laws can be incorporated as inference rules,

which may reduce considerably the size of the equations. We present two such inference rules here. A complete list can be found in [22].

Cancellation 2:

$$\frac{(E \cup \{f(d_1, d_2) \simeq y\}; \hat{s} \simeq \hat{t})}{(E \cup \{f(d_1, d_2) \simeq y, d_1\sigma \simeq x\}; \hat{s} \simeq \hat{t})} \quad \begin{array}{l} y \in V(d_1) \quad \sigma = \{y \mapsto f(x, d_2)\} \\ y \notin V(d_2) \quad x \textit{ is a new variable} \end{array}$$

Cancellation 4:

$$\frac{(E \cup \{f(p, u) \simeq f(q, u)\}; \hat{s} \simeq \hat{t})}{(E \cup \{p \simeq q\}; \hat{s} \simeq \hat{t})}$$

where the function f is right cancellable. In *Cancellation 2*, if the substitution $\sigma = \{y \mapsto f(x, d_2)\}$ is applied to the given equation, it becomes $f(d_1\sigma, d_2) \simeq f(x, d_2)$, since y does not occur in d_2. The cancellation law reduces this equation to $d_1\sigma \simeq x$. *Cancellation 4* is not necessary for the purpose of completeness, but it helps in improving efficiency.

Simplification-based strategies also feature rules such as **Functional subsumption**,

$$\frac{(E \cup \{p \simeq q, l \simeq r\}; \hat{s} \simeq \hat{t})}{(E \cup \{l \simeq r\}; \hat{s} \simeq \hat{t})} \; (p \simeq q) \trianglerighteq (l \simeq r)$$

which delete equations subsumed by other equations, and **Deletion**

$$\frac{(E \cup \{s \simeq s\}; \hat{s} \simeq \hat{t}\})}{(E; \hat{s} \simeq \hat{t})}$$

which delete trivial equations. These rules do not reduce any minimal proofs, but they delete equations, which are *redundant*, in the sense that they do not contribute to any minimal proofs and therefore are not needed in the derivation. Deletion also applies to the target

$$\frac{(E; \hat{s} \simeq \hat{s}\})}{(E; true)}$$

in order to detect that the target is proved.

Another inference on the target is superposition of an un-orientable equation onto a target equality $\hat{s} \simeq \hat{t}$ to generate a new target equality. A newly generated target equality is first simplified as much as possible and then it is kept only if it is smaller than $\hat{s} \simeq \hat{t}$. This rule is called **Ordered saturation** [1]:

$$\frac{(E \cup \{l \simeq r\}; N \cup \{\hat{s} \simeq \hat{t}\})}{(E \cup \{l \simeq r\}; N \cup \{\hat{s} \simeq \hat{t}, \hat{s}' \simeq \hat{t}'\})} \quad \begin{array}{l} \hat{s}|u = l\sigma \qquad \hat{s}[r\sigma]_u \rightarrow_E^* \hat{s}' \qquad \hat{t} \rightarrow_E^* \hat{t}' \\ \{\hat{s}', \hat{t}'\} \prec_{mul} \{\hat{s}, \hat{t}\} \end{array}$$

Ordered saturation applies if $\hat{s} \prec \hat{s}[r\sigma]_u$, since if $\hat{s} \succ \hat{s}[r\sigma]_u$ holds, simplification would apply. The target equality $\hat{s}' \simeq \hat{t}'$ might have a shorter proof than the other target equalities. Ordered saturation allows us to generate more than one target in order to broaden our chance of reaching the proof as soon as possible.

Rules such as simplification, subsumption and deletion are called *contraction* inference rules, because they delete equations or replace them by smaller equations. Rules like superposition and ordered staturation instead are *expansion* inference rule, because they generates new equations and add them to E or to the target. Roughly speaking, a step which deletes a sentence also deletes the portion of the search space which depends on that sentence, i.e. all the inferences which could be applied to that sentence. On the contrary, an expansion step expands the data base and therefore the search space. It follows that in order to keep the size of the search space manageable, it is desirable to apply as much as possible the contraction rules and to restrict as much

as possible the application of the expansion rules. This is in fact the philosophy of the simplification-based strategies. First of all, these strategies adopt *simplification-first* search plans [21], i.e. search plans which give priority to contraction inference rules. Under such search plans, expansion rules are applied only if no contraction rule applies. Consequently, the current set of equations and therefore the current search space is always kept as reduced as possible. Secondly, simplification-based strategies impose strong ordering based restrictions on the expansion rules, such as those embedded in the definitions of superposition and ordered saturation. Such restrictions make the inference rules proof-reducing and limit their applicability, thereby reducing their capability of expanding the search space. These choices have turned out to be very successful in practice, up to the point of bringing within reach unsolved challenge problems, as described in the following section.

3 Putting theory into practice

We have developed a family of theorem provers for equational theories whose design strictly adheres to the aforementioned methodology. The latest versions is $SBR3$, written in CLU and runs on Sun3. A new version in C++, $SBR4$, with the same functionalities is being developed. $SBR4$ runs on any machines that supports C++, and is much faster than $SBR3$. On the problems which we have tested on both $SBR3$ and $SBR4$ (the latter on a Sparcstation), the latter is usually at least ten times faster.

$SBR3$ takes as inputs an equational theory E and an equation $s \simeq t$ and tries to prove that $s \simeq t$ is a theorem of E. It proves a theorem the refutational way. That is, it replaces all variables in $s \simeq t$ by new Skolem constants and tries to find a contradiction to $E \cup \{\hat{s} \neq \hat{t}\})$ where $\hat{s}$ and $\hat{t}$ are the skolemization of s and t. Then the prover will try to deduce an instance of $x \neq x$ which yields the contradiction.

In addition to the theory and the equation, the user should also provide an ordering for comparing the terms. Usually the ordering should be a *complete simplification ordering* (a simplification ordering which is total on ground terms). In $SBR3$ the user has the choice of assigning a precedence among the operators in the theory and choose an ordering from a list implemented in the system. However, $SBR3$ will not check the totality for the user. The lack of totality on ground terms may actually be turned into a powerful search strategy similar to *Ordered Saturation* described in the previous section.

The backbone of $SBR3$ is a variation of unfailing Knuth-Bendix completion, mentioned in the previous section, which also incorporates the commutative and associative (AC) axioms of an operator into the unification algorithm. We term this procedure AC-UKB. Although the AC axioms can be handled simply as equations, it is advantageous to treat them implicitly in the unification process to the number of unnecessary *Superposition* inferences.

What differ $SBR3$ from the other provers, in addition to the simplification-based inference system, are its simple yet extremely powerful search plans. Search plans are usually treated in theorem proving in an *ad hoc* and incomplete way – anything that produces proofs is allowed. Fairness (thus completeness of the proof strategy) is usually compromised by the concern for greater efficiency. Using the notion of proof reduction, we have demonstrated that it is possible to achieve both completeness (fairness of the search plan) as well as efficiency. In $SBR3$, only fair search plans are implemented. Our experiments showed that they, if done properly, can indeed be both complete and very efficient.

The most important design choice common to all the search plans in $SBR3$ is that they are *simplification-*

first plans. That is, no superposition step is ever performed if there are still simplification steps and functional subsumption steps to be done. This search plan, coupled with cancellation, controls the growth of the number and size of equations sufficiently enough to obtain proofs for simple to moderately difficult problems. For more difficult problems, however, the search space quickly grows to an unmanageable size.

The first question we tackle is one of finding a shorter path to a solution. UKB, being complete, guarantees the existence of a proof through simplification and superposition should there be one. It does not, however, guarantee to provide a *short* proof. Suppose the prover can look at several different inequalities and tries to find a contradiction simultaneously[1], then conceivably one can find a proof faster. On the other hand, one should also keep in mind not to inundate the search space with irrelevant inequalities.

SBR3 provides a facility for maintaining a reasonable number of inequalities, to check for shorter proofs, by modifying the *ordered saturation* rule. When an un-orientable equation is generated, we superimpose it into an existing inequality (say A) to create a new inequality if possible. Then the new inequality is simplified using the rest of the equations and rules into B. The inequality B is kept, without deleting A, if $A \not\leq B$ according to the ordering. We term this method the *inequality ordered-saturation strategy*. This strategy is indispensable for proving some of the more difficult problems which we experimented [1].

Another challenge is to eliminate redundant critical pairs. This problem is especially serious in AC-rewriting due to the potentially astronomical number of AC-unifiers. In the term rewriting literature there are a handful of critical pair criteria, whose purpose is to eliminate unnecessary critical pairs. However, all of them are designed not to destroy the confluence property of any given two terms. In refutational theorem proving, on the other hand, we are only interested in the confluence of the two terms of the targeted theorem. Therefore a critical pair can be deleted or suspended as long as it does not destroy the confluence of the intended terms.

Taking advantage of this property, we employed a notion of *measure* in *SBR3*. A measure is defined syntactically on the structure of terms: for example, the number of occurrences of a specific operator may be a measure. The measure estimates the likelihood of whether a critical pair may contribute to an eventual proof of the intended theorem. Critical pairs are ordered according to the measure which decides the next equation to be chosen to perform superposition. Certain measures even allow us to delete critical pairs if they are deemed irrelevant for producing a proof. This search strategy is called *filtration-sorted strategy* and its details can be found in [2]. Three different types of measure are implemented in *SBR3*.

4 Experimental results: automatic proofs by SBR3

We have conducted extensive experiments on *SBR3*. We tested the prover on all the examples in equational theorem proving which we could find, as well as some new ones. The experiments we performed showed a dramatically small search space, just as expected. As a simple example, for the well-known *Salt and Mustard* puzzle of Lewis Carroll, first suggested by the Argonne Theorem Proving Group as a challenge problem for theorem provers, the Argonne prover Otter [25] generated more than 32,000 clauses before finding the solution while ours succeeded after generating less than 2000 rewrite rules.

The performance of *SBR3* on serious mathematical problems is even more impressive. The celebrated Jacobson's Theorem of ring theory for $n = 3$ [31], the independence of ternary algebra axioms [27], etc., have all been proved in a few minutes. In the following we describe some of the problems for which *SBR3* provided

[1] The basic UKB only looks at one.

the *first* computer proofs.

Classical Regular Languages

In [14], there is an equational formulation of classical regular languages by Yanov (page 108 of [14]) which completely axiomatize regular languages containing the empty string. The axioms are:

$x + x == x$

$z.(x + y) == (z.x) + (z.y)$

$(x + y).z == (x.z) + (y.z)$

$(x^*)^* == x^*$

$x^*.x^* == x^*$

$x + (x + y)^* == (x + y)^*$

$(x + y)^* == (x^* + y)^*$

$(x + y)^* == (x.y)^*$

$(x.y)^*.x == (x.y)^*$

$x.(x.y)^* == (x.y)^*$

$y + (x.y) == x.y$

$x + (x.y) == x.y$

$x + y == y + x$

$(x.y).z == x.(y.z)$

$(x + y) + z == x + (y + z)$

where "." is concatenation. $SBR3$ proved that

$$(\sum_{i=1}^{n} A_i)^* = (\sum_{i=1}^{n} A_i.\overline{A_i}^*.A_i)^*(1 + \sum_{i=1}^{n} A_i.\overline{A_i}^*)$$

where $\overline{A_i} = A_1 + \cdots + A_{i-1} + A_{i+1} + \cdots + A_n$, for $n = 3$ and $n = 4$, and the languages contain the empty strings[2]. In [14], Conway used an entire chapter to introduce a new technique to prove these two problems and remarked (page 119) that *"... even for $n = 3$ it is difficult to produce a proof without using the general ideas of this chapter, and for $n = 4$ I doubt if a completely written out proof could be fitted into 10 pages"*. The *direct* proof, produced by $SBR3$, needs no more than five new critical pairs, in addition to the simplication steps! The cpu time needed for $n = 3$ is about 4 minutes and 42 minutes for $n = 4$.

Dependency of Lukasiewicz's fifth axiom

Lukasiewicz's many-valued logic is defined using the following four axioms:

$true \Rightarrow x == x$

$(x \Rightarrow y) \Rightarrow ((y \Rightarrow z) \Rightarrow (x \Rightarrow z)) == true$

[2]The equations are not true in classical regular algebra when $n \geq 5$.

$(x \Rightarrow y) \Rightarrow y == (y \Rightarrow x) \Rightarrow x$

$(not(x) \Rightarrow not(y)) \Rightarrow (y \Rightarrow x) == true.$

The problem is whether the fifth axiom $x \Rightarrow y \vee y \Rightarrow x == true$ is necessary [17]. The conjecture of its dependency was given by Lukasiewicz in the 20's, as reported in [32], and proved many years later [13, 26].

The proof by $SBR3$ is done by first deriving a few lemmas from the axioms, one of which leads to the definition of an additional operator *or*. Then $SBR3$ proves that *or* is AC. Finally, the conjecture is proved in about 2 minutes. For the final session, the inputs are

$true \Rightarrow x == x$

$x \Rightarrow x == true$

$x \Rightarrow true == true$

$(x \Rightarrow y) \Rightarrow ((y \Rightarrow z) \Rightarrow (x \Rightarrow z)) == true$

$not(not(x)) == x$

$(x \Rightarrow y) \Rightarrow y == (y \Rightarrow x) \Rightarrow x$

$or(not(x), y) == x \Rightarrow y$

$x \vee y == (x \Rightarrow y) \Rightarrow y$

Declared AC-operator: *or*.

Theorem proved: $x \Rightarrow y \vee y \Rightarrow x == true$.

A detailed description of the experiments in Lukasiewicz logic can be found in [3, 4, 10].

Moufang identities in alternative rings

Alternative rings are rings with the associativity of $*$ replaced by two *alternative* axioms. The Moufang identities are a set of equational theorems of alternative rings. The Moufang identities as a challenge to theorem provers was first suggested in [30], although no automated proof was given. They were later proved automatically using a special-purpose theorem prover designed for ring theory [35]. $SBR3$ is the first syntactic theorem prover which proved them automatically.

Alternative rings are defined by

$0 + x == x$

$0 * x == 0$

$x * 0 == 0$

$g(x) + x == 0$

$g(x + y) == g(x) + g(y)$

$g(g(x)) == x$

$x * (y + z) == (x * y) + (x * z)$

$(x + y) * z == (x * z) + (y * z)$

$(x * y) * y == x * (y * y)$

$$(x * x) * y == x * (x * y)$$
$$g(x) * y == g(x * y)$$
$$x * g(y) == g(x * y)$$
$$a(x, y, z) == ((x * y) * z) + g(x * (y * z)$$

where a is an auxiliary operator.

SBR3 proved the following properties (the middle alternative law and two skew-symmetries of a) within 20 seconds:

$$(x * y) * x == x * (y * x)$$
$$a(y, x, z) == g(a(x, y, z))$$
$$a(z, y, x) == g(a(x, y, z))$$

The Moufang identities are defined as:

$$(((x * y) * x) * z) = (x * (y * (x * z))) \text{ (left Moufang)}$$
$$(((z * x) * y) * x) = (z * (x * (y * x))) \text{ (right Moufang)}$$
$$((x * y) * (z * x)) = ((x * (y * z)) * x) \text{ (middle Moufang)}$$

and they are proved in 49, 55, and 41 minutes respectively.

By adding the left and right Moufang into the input set, we are able to give a direct proof of

$$a(x * x, y, z) == ((a(x, y, z) * x) + (x * a(x, y, z)))$$

in 13 minutes. A full account of our experiments in alternative rings is given in [1].

Another series of problems which we are working on now is to verify the theorems of the book *A Formalization of Set Theory without Variables* by Tarski and Givant. As pointed out in [11], this will have direct implication on the design and optimization of query languages in relational data bases and program synthesis.

Our experiments are encouraging. They show us that high performance automated deduction is feasible even with our current knowledge and technology. We believe that the philosophy of simplification underlying our prover is the most significant reason for the dramatic reduction of search space, which made all our automatic proofs possible.

5 Distributed theorem proving

We are currently working on the design of a simplification-based strategy for *parallel automated deduction* in a *distributed multi-processing* environment. We feel that simplification-based theorem proving is an ideal candidate for application of parallel computation, because the rewriting approach couples a strong and elegant theoretical foundation with an extremely encouraging experimental record. A deep understanding of the problem at hand is necessary to design an architecture that exploits successfully the increased computing power of a parallel

environment. It would also open a new perspective of application for parallel computation which has not been investigated before.

Relatively little work has been done in this area so far. Parallelizing a simplification-based strategy is significantly different from parallelizing a conventional, space consuming theorem proving strategy. The latter uses mostly expansion inferences and it is relatively easy to perform expansion steps in parallel, because expansion steps are more or less independent from each other. More precisely, any two inference steps which do not have premises in common are trivially independent and can proceed concurrently, at least in principle. For expansion inferences, two steps which share one or more premises are also independent, because expansion steps do not modify their premises. Expansion steps simply need to be granted *read-access* to their premises. Since concurrent read can be safely admitted, the parallelization of expansion inferences does not raise basic conceptual problems. In a simplification-based strategy, however, inference rules are intertwined. The reason is that contraction inferences do modify their premises. A contraction step needs not just read-access, but also *write-access* to its premises. Therefore, two contraction steps which share premises may cause a *write-write* conflict if they attempt to modify concurrently the same data. Also, contraction steps may have *read-write* conflicts with concurrent expansion steps.

Even this very basic analysis of the problem shows that the presence of contraction rules makes the design of a parallel strategy harder. However, we think that the gain is well worth the additional effort. Firstly, there is ample empirical evidence that sequential strategies with contraction rules are much more powerful than those without contraction. This behaviour is also justified theoretically by our proof reduction view. Based on this, it is reasonable to foresee that the same pattern of behaviour will appear when comparing parallel strategies. In fact, we expect an even much better improvement. By grossly simplifying the problem, let C_0 be a sequential strategy without contraction rules and let t be the time spent by C_0 to prove a given input $(S;\varphi)$. Let C_1 be the sequential strategy obtained by adding contraction to C_0 and let t/s, for some $s > 1$, be the time required by C_1 on $(S;\varphi)$. Furthermore, let C_2 and C_3 be respectively a parallel version of C_0 and a parallel version of C_1. We expect that if C_2 takes time t/n, $n > 1$, to prove φ, C_3 will take time t/p, where $p > n \cdot s$. In other words, we expect the speedup of a parallel simplification-based strategy to be much higher than the mere combination of the speedup induced by simplification and the speedup induced by parallelism. This may not be true for all inputs, but we expect it to hold for most targets. The intuitive reason for our expectation is the following. Roughly speaking, if we execute in parallel an expansion-only strategy, we will be able to perform expansion steps by batches rather than one by one. The equations will be generated faster and the derivation will succeed at an earlier stage than the sequential one. However, the solution obtained is in some sense the same, as the same equations are generated. On the other hand, if we execute in parallel a simplification-based strategy, powerful simplifiers may be generated much sooner than in the sequential derivation. In a simplification-first strategy, the early application of such simplifiers may trigger the early generation of other simplifiers and an eventual radical modifications of the data base, leading the prover to find a different and much faster successful path than the one found by the sequential execution.

Problems related to those of parallel deduction have been addressed by the study of parallel and distributed implementations of the *Buchberger algorithm* [34, 29, 18]. The Buchberger algorithm works on polynomials, equated to 0 and treated as oriented equations. It takes as input a set of polynomials and gives as output a *basis* for the ideal generated by the input polynomials. The basis has the property that it reduces to 0 all and only the polynomials belonging to the ideal [12]. The Buchberger algorithm is related to the simplification-based strategies because it features an expansion inference rule which is similar to superposition and a contraction

rule which is similar to simplification. There are also substantial differences, because the Buchberger algorithm has a much less general purpose than a theorem proving strategy. The Buchberger algorithm is an algorithm, whereas the theorem proving strategies are semidecision procedures. Its inferences do not use unification, since there are no variables, as the "variables" in the polynomials are constants logically. It follows that expansion steps are much less expensive than in theorem proving. Also, the equations are all trivially oriented into rewrite rules, because they are obtained by equating polynomials to 0. Nonetheless, parallel implementations of the Buchberger algorithm need to deal with the problem of the coexistence of expansion and contraction inferences. The three approaches presented in [34, 29, 18] address the problem within three different models of parallel computation: a shared memory multi-processor in [34], a data-flow machine in [29] and a distributed memory multi-processor in [18]. All three algorithms have interesting features. However, none of them implements a simplification-first methodology. In fact, the data base of polynomials is not maintained fully simplified by any of these three implementations. In particular, very little *backward contraction*, i.e. simplification of formerly existing equations by newly generated ones, is performed. As a consequence, expansion rules are applied to equations which are not fully reduced, unnecessary equations are generated and the search space swells. It seems that this phenomenon has prevented these three implementations from achieving better speedups. The trouble is that requiring equations to be fully simplified, before they are allowed to expand, introduces some sequentiality. An expansion process cannot be granted read-access to an equation until all simplification processes have had write-access to it. We have then two at least partially conflicting desiderata: on one hand, we would like to simplify as much as possible before expanding, while in the meantime we would like to perform as many steps in parallel as possible. The problem is to find a satisfactory trade-off between these two.

We have kept this issue in mind since the early stages of our project. So far, we have settled on a few basic choices. The first one is *coarse grain* versus *fine grain* parallelism or, equivalently, *coarse granularity* of protection versus *fine granularity* of protection. For the purpose of this discussion, we regard as fine granularity the *term level* and as coarse granularity the *equation (or clause) level.* Thus, fine granularity means that every term is a *grain* of memory with its own access rights. Fine granularity allows parallel processes to access different subterms of the same term. Parallel matching, parallel rewriting and parallel unification are examples of fine grain parallelism. On the other hand, coarse granularity means that if a process is granted access to an equation, no other process can access any part of it. Fine grain parallelism is well suited for equational programs, where just one term needs to be reduced by a static set of equations. In theorem proving we have a dynamic set of equations where every single term is subject to simplification. It seems to us that under these conditions the overhead of handling fine granularity would be unreasonably high. Therefore, we choose to concentrate ourselves on coarse grain parallelism, although some fine grain parallelism might be considered at a later stage.

The second basic choice is *shared memory* versus *distributed memory.* This choice is related to the previous one. Fine grain parallelism leads in general to adopt a shared memory, since it does not seem realistic to scatter the terms of an equation over a distributed memory. Coarse grain parallelism can be implemented in principle in both a shared memory and a distributed memory. However, we are oriented toward distributed memory, for the following reasons. Theorem proving is basically search for solutions in a generally huge search space. We expect parallelism to help in two ways: by keeping the search space small by eager, parallel simplification and by searching it in parallel along different paths. In order to realize this intuitive idea of *parallel search*, we need the parallel processes to be rather independent. Thus, the processors should be rather *loosely coupled*, with no shared memory. We envision a situation where each processor has in its own memory a set of equations S^i and

the union of all the S^i's form the current data base S. The S^i's are initially disjoint, but in general they do not remain disjoint during the derivation. Also, each processor is originally given a copy of the input target φ_0. Since different processors perform different steps, φ_0 may be reduced to different, yet equivalent targets, one per processor. Each processor performs its own inference steps searching for a proof. However, the processors do communicate by broadcasting their equations to all the other processors. When receiving equations from the outside, a processor uses them to perform inferences with its own equations. The simplification-first methodology is strictly enforced at the local level. Each processor maintains its own data base fully reduced, including the equations received as messages. No expansion step is performed if the equations involved are not fully reduced, at least locally. Clearly, they are not guaranteed to be reduced with respect to the global data base. However, our strategy is *fair* in the sense that it guarantees that any two equations generated at remote sites will be able to interact through messages, if they are not simplified locally beforehand. The cost of handling such messages is the price to pay for the high degree of independence of the processors. In addition, this scheme induces a certain amount of redundancy, as the data bases at different sites are not guaranteed to be disjoint and therefore it may happen that a same step is executed by more than one processor.

This is just a very brief sketch of a few basic ideas in our work. We are currently studying the details, trying to minimize redundancy and the cost of message passing. Based on the investigations conducted so far and on the observation that the implementations in [34, 29, 18] obtained significant speedups even in the absence of full simplification, we expect that this on going research will ultimately increase the speed of a theorem prover like *SbReve* by at least a hundred times.

References

[1] S.Anantharaman and J.Hsiang, Automated Proofs of the Moufang Identities in Alternative Rings, *Journal of Automated Reasoning*, Vol. 6, No. 1, 76–109, 1990.

[2] S.Anantharaman and A.Andrianarivelo, Heuristical Critical Pair Criteria in Automated Theorem Proving, in A.Miola (ed.), *Proceedings of the International Symposium on the Design and Implementation of Symbolic Computation Systems*, Capri, Italy, April 1990, Springer Verlag, Lecture Notes in Computer Science 429, 184–193, 1990.

[3] S.Anantharaman and M.P.Bonacina, Automated Proofs in Lukasiewicz Logic, Technical Report, Department of Computer Science, SUNY at Stony Brook, November 1989.

[4] S.Anantharaman and M.P.Bonacina, An Application of the Theorem Prover SBR3 to Many-valued Logic, in M.Okada and S.Kaplan (eds.), *Proceedings of the Second International Workshop on Conditional and Typed Term Rewriting Systems*, Montréal, Canada, June 1990, Springer Verlag, Lecture Notes in Computer Science, to appear.

[5] L.Bachmair, N.Dershowitz and J.Hsiang, Orderings for Equational Proofs, in *Proceedings of the First Annual IEEE Symposium on Logic in Computer Science*, 346–357, Cambridge, Massachussets, June 1986.

[6] L.Bachmair and N.Dershowitz, Equational inference, canonical proofs and proof orderings, *Journal of the ACM*, to appear.

[7] M.P.Bonacina and J.Hsiang, On Rewrite Programs: Semantics and Relationship with Prolog, *Journal of Logic Programming*, to appear.

[8] M.P.Bonacina and J.Hsiang, Completion procedures as Semidecision procedures, in M.Okada and S.Kaplan (eds.), *Proceedings of the Second International Workshop on Conditional and Typed Term Rewriting Systems*, Montréal, Canada, June 1990, Springer Verlag, Lecture Notes in Computer Science, to appear.

[9] M.P.Bonacina and J.Hsiang, On fairness of completion-based theorem proving strategies, in R.V.Book (ed.), *Proceedings of the Fourth International Conference on Rewriting Techniques and Applications*, Como, Italy, April 1991, Springer Verlag, Lecture Notes in Computer Science 488, 348–360, 1991.

[10] M.P.Bonacina, Problems in Lukasiewicz logic, in *Newsletter of the Association for Automated Reasoning*, No. 18, June 1991.

[11] P.Broome, Applications of Algebraic Logic to Recursive Query Optimization, in *Proceedings of the Eighth Army Conference on Applied Mathematics and Computing*, 1990, to appear.

[12] B.Buchberger, An Algorithm for Finding a Basis for the Residue Class Ring of a Zero-dimensional Polynomial Ideal, (in German), PhD thesis, Department of Mathematics, University of Innsbruck, Austria, 1965.

[13] C.C.Chang, in *Transactions American Mathematical Society*, No. 87, 55–56, 1958.

[14] J.H.Conway, *Regular Algebra and Finite Machines*, Chapman and Hall, 1971.

[15] N.Dershowitz, Termination of Rewriting, *Journal of Symbolic Computation*, Vol. 3, No. 1 & 2, 69–116, February/April 1987.

[16] N.Dershowitz and J.-P.Jouannaud, Rewrite Systems, Chapter 15, Volume B, *Handbook of Theoretical Computer Science*, North-Holland, 1989.

[17] J.M.Font, A.J.Rodriguez and A.Torrens, Wajsberg algebras, *Stochastica*, Vol. 8, No. 1, 5–31, 1984.

[18] D.J.Hawley, A Buchberger Algorithm for Distributed Memory Multi-Processors, in *Proceedings of the International Conference of the Austrian Center for Parallel Computation*, Linz, Austria, October 1991, Springer Verlag, Lecture Notes in Computer Science, to appear.

[19] C.M.Hoffmann and M.J.O'Donnell, Programming with Equations, *ACM Transactions on Programming Languages and Systems*, Vol. 4. No. 1, 83–112, January 1982.

[20] J.Hsiang, Refutational Theorem Proving Using Term Rewriting Systems, *Artificial Intelligence*, Vol. 25, 255–300, 1985.

[21] J.Hsiang and M.Rusinowitch, On word problems in equational theories, in Th.Ottman (ed.), *Proceedings of the Fourteenth International Conference on Automata, Languages and Programming*, Karlsruhe, Germany, July 1987, Springer Verlag, Lecture Notes in Computer Science 267, 54–71, 1987.

[22] J.Hsiang, M.Rusinowitch and K.Sakai, Complete set of inference rules for the cancellation laws, in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milano, Italy, August 1987, 990–992.

[23] C.Kirchner, H.Kirchner and J.Meseguer, Operational semantics of OBJ3, in *Proceedings of the 9th International Conference on Automata, Languages and Programming*, LNCS 241, Springer Verlag, 1988.

[24] D.E.Knuth and P.B.Bendix, Simple Word Problems in Universal Algebras, in J.Leech (ed.), *Proceedings of the Conference on Computational Problems in Abstract Algebras*, Oxford, England, 1967, Pergamon Press, Oxford, 263–298, 1970.

[25] W.W.McCune, OTTER 2.0 Users Guide, Technical Report ANL-90/9, Argonne National Laboratory, Argonne, Illinois 1990.

[26] C.A.Meredith, in *Transactions American Mathematical Society*, No. 87, 54, 1958.

[27] A.J.Nevins, A human-oriented logic for automatic theorem proving, *Journal of the ACM*, Vol. 4, 606–621, 1974.

[28] M.Rusinowitch, Theorem-proving with Resolution and Superposition, *Journal of Symbolic Computation*, Vol. 11, No. 1 & 2, 21–50, January/February 1991.

[29] K.Siegl, Gröbner Bases Computation in STRAND: A Case Study for Concurrent Symbolic Computation in Logic Programming Languages, Master thesis and technical Report No. 90-54.0, RISC-LINZ, November 1990.

[30] R.L.Stevens, Challenge Problems from Nonassociative Rings for Theorem Provers, in E.Lusk and R.Overbeek (eds.), *Proceedings of the Ninth Conference on Automated Deduction*, Argonne, Illinois, May 1988, Springer Verlag, Lecture Notes in Computer Science 310, 730-734, 1988.

[31] M.E.Stickel, A case study of theorem proving by the Knuth-Bendix method: Discovering that $x^3 = x$ implies ring commutativity, in *Proceedings of the Seventh Conference on Automated Deduction*, Springer Verlag, Lecture Notes in Computer Science 170, 248–258, 1984.

[32] A.Tarski and J.Lukasiewicz, Investigations into the sentential calculus, Chapter IV in A.Tarski, *Logic, Semantics and Metamathematics*, 38–56, Clarendon Press, Oxford, 1956.

[33] A.Tarski and S.Givant, *A Formalization of Set Theory Without Variables*, American Mathematical Society, Colloquium Publications, Vol. 41, 1987.

[34] J.-P.Vidal, The Computation of Gröbner Bases on A Shared Memory Multiprocessor, in A.Miola (ed.), *Proceedings of the International Symposium on the Design and Implementation of Symbolic Computation Systems*, Capri, Italy, April 1990, Springer Verlag, Lecture Notes in Computer Science 429, 81–90, 1990.

[35] T.C.Wang, Case Studies of Z-module Reasoning: Proving Benchmark Theorems from Ring Theory, *Journal of Automated Reasoning*, Vol. 3, No. 4, 1987.

Constructive Relational Programming: A Declarative Approach to Program Correctness and High Level Optimization[†]

Paul Broome
U.S. Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD 21005-5066
(broome@brl.mil)

and

James Lipton
Dept. of Mathematics
University of Pennsylvania
Philadelphia, PA
(lipton@saul.cis.upenn.edu)

Abstract. Program efficiency and program correctness are often conflicting aims. The efficient program may be unreadable and the well structured, obviously correct program may have unnecessary steps.

We offer an approach for attaining both correctness and efficiency. Our solution includes a binary rewriting language based on Tarski and Givant's system of relation combinators. In this language smaller, correct programs can be straightforwardly combined to give larger programs. Programs can often be proved semantically equivalent using the equations of relation algebras, to give a reliable optimization method.

We illustrate the expressiveness of this system by applying it to a simplified version of the stable marriages problem. We also illustrate a natural application of non-monotonic logic in which a program query accepts a database as a parameter, constructed from a complex expression.

This is a new style of program construction based on a traditional, mathematical notation.

1. Introduction. Advances in the theory of programming languages and program correctness in recent years have been impressive. They have given a major boost to program reliability by providing clear, high-level tools for program development that stress modularity and increasingly transparent connections between programs and their specifications. The work has already resulted in dramatic declines in software development time –the costliest factor in computing– and in the undertaking of projects orders of magnitude larger than those one could have conceived of a few decades ago.

Several key paradigms have emerged in this work:

> Strongly typed programming languages provide an expressive type discipline to promote modularity, clarity in the definition of data, and a certain degree of compile-time error checking. Type-safe programs may, however, be incorrect: most type disciplines are not expressive enough to be a specification language.

On the other hand, declarative programs, in their purest form, come very

[†] (Partly) supported by the U.S. Army Research Office through the Mathematical Sciences Institute of Cornell University. We gratefully acknowledge the assistance of Raymond Ng, Barbara Broome, and Brint Cooper.

close to the goal of programming *directly with executable specifications.* These so-called logic programming languages (e.g., Prolog) are ideally suited to symbolic computation, and have been very successful in expert systems, databases, and other sophisticated applications. Sometimes, however, precisely because of the distancing from implementation encouraged by such languages, there is considerable loss of efficiency.

Functional Programming languages provide some benefits of both imperative and declarative languages. They are more algorithmic in spirit, equally suited to symbolic computation, and definition-based, in a clear and modular way. Some even provide a type-free language for the programmer, with automatic type inference at compile time (ML).

Relational programming, because of its compositional nature, extends functional programming in a natural way. But it shares the reversibility, non-deterministic robustness, and declarative nature of logic programming: relations are assertions.

Correctness and Efficiency. In high level language programming, clarity and ease of software development are a principal benefit, sometimes at the expense of efficiency. This has tended to place a burden on compilation as an optimizing process. Work in optimization has tended to be too low-level, almost independent of the transparency considerations above. It remains a significant problem to integrate this work with any of the programming concepts aimed at enforcing correctness.

The logic of binary relations is an attractive formalism for addressing both correctness and efficiency issues. The *Relation Algebras* of Tarski and Givant permit a variable-free, combinatory formalization of set theory that can specify input/output relations in a computationally useful, declarative way. Work discussed below and in related papers suggests that operations on proper binary relations between terms permits both a declarative database style of programming and a concise compilation technique for logic programs.

The formalism rests upon a well-developed algebraic theory. It is therefore well suited to the development of verifiably correct code and correctness proofs. Yet it lends itself to efficient program transformation techniques that constitute one of the more promising vehicles for high level optimization. Its elementary axiomatization and variable-free nature make it an interesting candidate for symbolic computation and metalogical programming.

High level optimization is addressed in this framework by introducing an inherently relational combinator to describe single linear recursions. New equations about this combinator supplement those of relation algebras and Q-relation algebras. These equations provide rules for collapsing loops and propagating constraints. Thus efficiency, which has often been an issue seemingly independent of correctness considerations, turns out to be closely related to, and justified by, correctness arguments.

Several new directions of research have issued from the development of this relational programming paradigm. The computational analysis of relations provides a framework for typing logic and constraint logic programs in a way that suggests conventional Curry-Howard style typing of functional programming. This points to a logically sound way of combining declarative and functional code, bringing together the benefits of both paradigms.

Our objectives are to illustrate the expressiveness of a small collection of operations on proper binary relations. This collection comprises a declarative discipline that subsumes functional programming and database operations. We also illustrate their relevance to non-monotonic reasoning. These operations are based on the calculus offered in [Tarski] and are particularly important for correctness and efficiency concerns. The set theoretic equivalents are representatives of an algebraic structure called a relation algebra that can be used for high level optimization and program synthesis by heuristic rewriting[Broome].

2. Equational systems. The set theoretic representation of basic operations of a relational language are the following.

sum(union)	$\mathtt{F}+\mathtt{G} \equiv \{x,y \mid x\ \mathtt{F}\ y \vee x\ \mathtt{G}\ y\}$,
product(intersection)	$\mathtt{F}*\mathtt{G} \equiv \{x,y \mid x\ \mathtt{F}\ y \wedge x\ \mathtt{G}\ y\}$,
relative product(composition)	$\mathtt{F};\mathtt{G} \equiv \{x,y \mid \exists z(x\ \mathtt{F}\ z \wedge z\ \mathtt{G}\ y)\}$,
converse	$\mathtt{converse(F)} \equiv \{x,y \mid y\ \mathtt{F}\ x\}$,
complement	$\mathtt{complement(F)} \equiv \{x,y \mid \neg x\ \mathtt{F}\ y\}$,
identity	$\mathtt{id} \equiv \{x,y \mid x = y\}$,
relation equality	$\mathtt{R} = \mathtt{S} \iff (\forall xy, x\ \mathtt{R}\ y \leftrightarrow x\ \mathtt{S}\ y)$,
left projection	$\mathtt{hd} \equiv \{x,y \mid \exists z, x = [y \mid z]\}$,
right projection	$\mathtt{tl} \equiv \{x,y \mid \exists z, x = [z \mid y]\}$,
atomic binary predicates	$\mathtt{F_1}, \mathtt{F_2}, \cdots, \mathtt{F_n}$.

Except for recursion, the language consists of program forming operations suggested by these definitions. Domain and range objects are terms. Program inputs and outputs are described in the theory of ordered pairs over the free algebra of terms in finitely many constants and function symbols. Relation equality formalizes a notion of program equivalence.

3. Syntax. A program is a sequence of definitions and queries. The fundamental concepts that distinguish this system are relational expressions and set formulas. Solutions to queries are sets finitely described by constraints. That is, interpretation is a search for a representation of a relation as a certain canonical formula defining a set of pairs of terms.

We describe the language in Backus-Naur form.

< *program*> : : = < *definition*>...< *definition*> | **solve** < *relation*> .
< *definition*> : : = **define** < *relation name*> => < *relation*> .
< *relation name*> : : = < *constant*> | < *constant*>(< *variable*>,...,< *variable*>)
< *constant*> : : = < *a string beginning with a lower case letter*>
< *variable*> : : = < *a string beginning with an upper case letter*>

<*relation*> : : = **id** | **hd** | **tl**
: : = < *relation*> + < *relation*>
: : = < *relation*> * < *relation*>
: : = < *relation*> ; < *relation*>
: : = **converse**(< *relation*>)
: : = **complement**(< *relation*>)
: : = **plus**(< *relation*>)
: : = **pi**(< *relation*> ,< *relation*> ,< *relation*>)
: : = < *constant*> | < *constant*>(< *relation*> ,..., < *relation*>)
: : = { < *term*>;< *term*> ,..., < *term*>;< *term*>}

The projections `hd` and `tl` are special cases of a more general n-ary projection denoted `p^i/j` . This is to be understood as a binary relation between a term and a subterm. The term has function symbol `p` with `i` arguments; the subterm is the `j`th argument of the term.

Definitions permit variables but queries do not. The relation expressions in a **define** <*relation name*> and <*relation*> permit variables for relations whereas the expression after **solve** must not have relation variables. Some examples of definitions are the diversity relation, the universal relation, and a sample, one element, ground relation defined as follows.

```
define di => complement(id).
define 1 => id + di.
define bits => {0;1}.
```

Solutions are described with two free variables `X` and `Y` and possibly other universally or existentially quantified variables. Values for variables are terms. A term is a constant, a variable, or a function symbol with arguments that are terms. A function symbol is also a constant. A variable begins with an upper case letter. A constant with a lower case letter.

All variables other than `X` and `Y` are either existentially quantified on the outside or locally universally quantified. Complements of projections introduce universally quantified variables. For example, the existentially quantified variable `Z` in the definition of `hd` becomes universally quantified when complemented. Universally quantified variables are denoted by their enclosure in parentheses. In particular, `complement(hd)` is the set of `X;Y` such that for every `X2`, `X` is different from the pair `[Y|X2]`. This is expressed as

```
{X;Y:  (X2)(X # [Y|X2])}.
```

Two illustrative examples are the solutions to both `bits` and `complement(bits)`. The following interaction shows these solutions.

```
solve bits.

bits = {X;Y: X=0 ^ Y=1}.

solve complement(bits).

complement(bits) = {X;Y: X=0^Y#1 v X#0}.
```

Set formulas are defined as follows.

< *set formula*> ::= {**X;Y** :< *logical expression*> }
< *logical expression*> ::= *true* | **X = Y** | **(X # Y)** | < *disjunction*>
< *disjunction*> ::= < *disjunct*> **v ... v** < *disjunct*>
< *disjunct*> ::= < *conjunct*> **^ ... ^** < *conjunct*>
< *conjunct*> ::= < *variable*> **=** < *term*> | < *quantifiers*>< *variable*> **#** < *term*>
< *quantifiers*> ::= {(< *variable*>)}

We illustrate with definitions that can be viewed as constructions of complex data types. These are called half product and the traditional cartesian product functor which we call a cross product.

```
define [R|S] => (R;converse(hd))*(S;converse(tl)).
define A#B => (hd;A;converse(hd))*(tl;B;converse(tl)).
```

Both `R` and `S` share domain objects in the half product. The range object is a pair; its first component is constructed with `R` and the second with `S`. For example, we have the following interaction:

```
solve  [{a;b}|{a;c}].
```

```
[{a;b}|{a;c}]={X;Y: X=a^Y=[b|c]}.

solve {a;b}#{a;c}.

{a;b}#{a;c}={X;Y: X=[a|a]^Y=[b|c]}.
```

4. Recursion. The primitive recursion operator, `pi`, is a schema for the divide and conquer paradigm that builds relations between terms. This recursor is defined as the following.

```
define pi(D,S,C) => S + D;(id#pi(D,S,C));C.
```

Either the terms are related with `S` or else they are more highly structured. In the second case the domain object is divisible by a relation `D` into a pair. In addition, the range object can be constructed from a pair by a relation `C`. The first components of these pairs must be identical; the second components are related by `pi(D,S,C)`. There are several advantages to this recurser; a major one is that it satisfies equations for cooperative loop merging and the propagation of constraints [Broome]. Transitive closure and maplist can be defined with `pi`.

```
define plus(R) => pi(R;converse(tl),R,tl).
define maplist(R) => pi(id,null,R#id).
define null => {[ ];[ ]}.
```

As an example, consider the relation `maplist(id)`. It is the (infinite) set of pairs of identical lists, but it has a finite description and contains the relation `null`. The relation `maplist(id)` would be described as

```
maplist(id) = {X;Y: X pi(id,null,id#id) Y}.
```

The query `solve maplist(id)*null` involves delaying the solution of `maplist(id)`, solving `null`, then solving the intersection to give

```
maplist(id)*null = {X;Y: X=[ ] ^ Y=[ ]}.
```

5. Interpretation. Interpretation is a search for a representation as a set formula. Any recursion free relational expression in the free term algebra is logically equivalent to a two quantifier, normal set formula as defined in the BNF syntax[CLPCR]. The argument makes use of quantifier elimination and other properties of term algebras. The form includes a limited use of quantifiers. Every variable not explicitly universally quantified in an disequation is understood to be existentially quantified. Given an arbitrary relational expression, each relation is successively expanded into equations and disequations and simplified into disjunctive normal form. In particular, the sum of two relations is effectively the disjunction of their two set expressions. Two relations are composed by unifying the range objects in the first relation with domain objects in the second and normalizing. A solution for `converse(R)` is simply an exchange of the variables `X` and `Y` in the solution of `R`.

An implementation technique for complementing a set of equations and disequations in disjunctive normal form has been described in [Chan] and further expanded in [Mayer, Plaza]. The algorithm normalizes the expression again into a disjunction of conjuncts. The simplification phase also simplifies to *true* any obviously valid or irrelevant equations or disequations in a conjunction and removes any conjunction that contains an unsatisfiable equation or disequation. For any finitely described set, R, two iterations of this algorithm gives an equivalent

set expression. That is, for finitely described sets, the set theoretic equation $\overline{\overline{R}} = R$ is satisfied operationally; `complement(complement(R))` is equivalent to `R`. However, this equation will not necessarily be satisfied for expressions R that contain certain occurrences of recursion.

We can illustrate interpretation with a predicate that insists that its argument (a sequence) has no duplicates. This predicate makes use of a `select` relation that chooses one element of an arbitrary length sequence. The relation `different` is a specialized identity relation on sequences. Only sequences without duplicate elements are acceptable.

```
define select => pi(id,hd,tl).
define different => pi(id, id#null, [hd*(tl;complement(select))|tl]).
```

6. Matching Problems. The stable marriages problem[SMP, MNR] is the generic version of a wide range of matching problems that generally includes assigning resources to tasks with the additional constraint that the resource be appropriate for the task. Sample applications may include assignments such as targets to weapons, residents to hospitals, and employees to jobs.

The stable marriages problem is the following: Given two sequences `B` and `G` of boys and girls and a binary relation `K` (knows) between names of boys and names of girls. A complete matching marries each boy to one and only one girl. The boy and girl must be members of `K` (i.e., they know each other). Let `m(B,K,G)` be the set of all complete matchings. That is, each member of `m(B,K,G)` is a one-to-one function defined from `B`→`G`. The problem is male-biased in that every boy, but not every girl, must be matched.

A solution is a set of one-to-one functions. There are a number of ways of representing a one-to-one function. The usual is a binary relation with no duplicate domain or range objects. This system cannot represent sets of sets at top level. It can, although, encode a function by use of indices and sequences. Thus a suitable representation of a one-to-one function is as a single pair of sequences, of the same size, with no duplicates in each sequence. Our set of all solutions, `m(B,K,G)`, uses the second representation, as a set of pairs of sequences.

An example problem uses the specific binary relations `b,k,g` representing an indexed set of boys, a knows relation, and an indexed set of girls. The sets `b` and `g` are automatically indexed with the integers. Thus {`b1,b2,b3`} is indexed with the elements $1, 2, 3$ to form the binary relation {`1;b1, 2;b2, 3;b3`}.

```
define b => {b1,b2,b3}.
define k => {b1;g1,b1;g2,b2;g1,b2;g2,b2;g3,b3;g3,b3;g4}.
define g => {g1,g2,g3,g4}.
```

In this example the answer is represented as six 1-1 functions.

```
solve m(b,k,g).
m(b,k,g)={X;Y:
  X=[b3,b2,b1] ^ Y=[g3,g1,g2]
V X=[b3,b2,b1] ^ Y=[g3,g2,g1]
V X=[b3,b2,b1] ^ Y=[g4,g1,g2]
V X=[b3,b2,b1] ^ Y=[g4,g2,g1]
V X=[b3,b2,b1] ^ Y=[g4,g3,g1]
V X=[b3,b2,b1] ^ Y=[g4,g3,g2]}.
```

The overall strategy for solving the stable marriage problem is to first form a matching of all of `B` to some of `G` insisting that there are no duplicates among the `G`'s, then to constrain this set to corresponding `B,G` pairs that are members of `K`. This is described as the intersection of two sets. The first is a set of unique matchings of boys to girls; the second insists that

`maplist(K)` hold for each pair in each matching.

```
define m(B,K,G) => unique(B,G)*maplist(K).
```

The construction of the unique matching, ignoring the `K` relation, involves forming arbitrary pairs, then insisting there are no duplicates. That is, while the matching is unique it may still include couples that are unknown to each other. Assume that `pairs(B,G)` can be defined to construct arbitrary sequences of girls of the same size as a given sequence of boys. If we can be provided with a unique sequence of boys then we only need to check that no girl is listed twice. Define `unique(B,G)` with the following.

```
define unique(B,G) => pairs(B,G);different.
```

7. Data Representations. We will briefly address issues in data representation. A common difficulty with most data representations is that numeric and symbolic data are typically kept separate; databases are often queried with a complex interface from a procedural programming language. This is a major source of complexity in program construction and the cause of error-prone programming.

The most powerful aspect of this notation as a database query language is that it is also a programming language. Both symbolic data and computations on them agreeably merge. In addition, changes of representation are easily performed.

In the stable marriages problem, a sequence of boys `B` is a one-to-one function. As described previously, it may be represented as a set of ordered pairs where the domain objects are nonnegative integers and the range objects are boys. A sequence may also be constructed with projections so that the one-to-one function could be represented as a single pair of sequences; the domain object is a sequence of nonnegatives and the range object is a sequence of names. Assume that `Bs` is the set of boys in the second representation.

If we were provided the second representation then `pairs(Bs,G)` could easily be defined as `converse(Bs);maplist(1;G)`. The purpose of the universal relation in `maplist(1;G)` is to free up the domain variable so that an arbitrary `G`, not necessarily the one with the same index, can be assigned to each `B`. This relation could play the role of `pairs(B,G)`.

On the other hand, if the set `B` is provided in the first representation then we would need to construct a sequence of integers so that the set `B` could be packaged into a single sequence. This could be done with a special identity relation `ints(B)` that would count the size of `B` and return a pair of sequences of nonnegatives from the size of `B` down to 1. Given this definition we could define `pairs(B,G)` as the following.

```
define pairs(B,G) => maplist(converse(B));ints(B);maplist(1;G).
```

Again we can gather `B`'s into a sequence by finding the `B` with the largest index, constructing a sequence of integers from 1 to the largest, and mapping `B` onto this sequence. The relation `ints(B)` is defined using `largest(B)` and `iota` where `largest(B)` provides the size of `B` and `iota` constructs the sequence of nonnegatives. For example, one element of `iota` would be the pair `3;[3,2,1]`. The relation `ints(B)` is a pair of sequences of integers the same size as the number of `B`'s. That is

```
define ints(B) => converse(iota);largest(B);iota.
define iota => pi([id|pred],(zero;1;null),id).
define largest(B) => ((B;1)*(succ;complement(B;1)))*id.
```

Indices are nonnegatives represented as sequences of 1 bits. The successor `succ` is an extension of a sequence to a sequence one bit longer. The converse of `succ` is `pred` defined as

the following:

```
define pred => tl*(hd;converse(bits);1).
```

8. Nonmonotonic Reasoning. A common problem with updates of real databases is closely related to belief revision in nonmonotonic logics. This arises when a formally valid conclusion must be revised or rejected. Often this must be dealt with in the program with side effects by updating the database and asking the query again. If the user is interested in testing possible scenarios one would like to integrate this scheme directly into the language.

The relation combinator formalism gives a very straightforward way of dealing with updated or changed conclusions resulting from revisions of the data. Our solution extends the notion exhibited in the functional query language FQL[BFN]. In FQL, an update to a database is the application a function to a database and a transaction to return a new database.

In our solution, the database is a relation and the set of transactions is a relation. The merger of the two is the relation sum which can be given a name. However, an experimental update would likely remain unnamed and passed as an argument to an expression to query the database. That is, we can incorporate the new information in an adjusted parameter without changing the program or any data files. Not only can a relation sum be substituted, but any expression or program that constructs that database can be substituted.

For example, consider deleting a fact. That is, what would be the stable marriage assignment if `b3` and `g4` no longer knew each other? We can perform this experiment by intersecting `k` with the set that does not contain the element `{b3;g4}`, that is replacing `k` in the query with `complement({b3;g4})*k`. This leaves `k` unchanged but gives two 1-1 functions as a solution.

```
solve m(b, complement({b3;g4})*k, g).

m(b,k*complement({b3;g4}),g)={X;Y:
   X=[b3,b2,b1] ^ Y=[g3,g1,g2]
 V X=[b3,b2,b1] ^ Y=[g3,g2,g1]}.
```

To find the three assignments that result from additionally including a new fact `{b3;g1}` we would solve the query `m(b, complement({b3;g4})*k + {b3;g1}, g)`.

Nonmonotonic reasoning is a powerful concept that permits the quick reformulation of concise and expressive queries. It may eventually be appropriate for situations in which interaction time is the bottleneck. This system of combinators permits a straightforward and expressive style of program construction.

9. Conclusion. We have described progress toward the definition of a small collection of program instructions based in set theory. Its simplicity is attractive for many reasons. Programs in this style are very expressive, widely applicable and important to program correctness, efficiency, and reusability concerns. Secondly, this style clarifies and enhances nonmonotonic reasoning techniques that goes beyond existing systems. It enables a flexible style of what-if reasoning that permits the declarative construction of a database for the purpose of an experimental query. Finally, the equations of a relation algebra form the foundation for program equivalence proofs to justifiably bridge the gap between correct and efficient programs.

REFERENCES

[CLPCR] P. Broome, J. Lipton, *Combinatory Logic Programming: Computing with Relations*, forthcoming.

[Chan] D. Chan, *Constructive Negation Based On the Completed Database*, Logic Programming: Proceedings of the Fifth International Conference and Symposium Eds. R. A. Kowalski and K. A. Bowen, MIT Press, Cambridge, MA, pp 111–125, 1988

[Mayer] M.J. Mayer, *Complete Axiomatizations of the Algebras of Finite, Rational and Infinite Trees,* Proc. 3rd Symp. on Logic in Computer Science, Edinburgh, pp 348–357, 1988.

[Plaza] J.A. Plaza, Fully Declarative Programming with Logic, Mathematical Foundations, dissertation, SUNY, 1991

[Tarski] A. Tarski, S. Givant, *A formalization of set theory without variables,* Colloquium publications, V. 41, American Mathematical Society, Providence, RI, 1987

[BFN] P. Buneman, R.E. Frankel, R. Nikhil, An Implementation Technique for Database Query Languages, ACM TODS Vol 7 No. 2, pp 164–186, 1982

[Broome] P. Broome, *Applications of Algebraic Logic to Recursive Query Optimization,* 8th Army Conference on Applied Mathematics and Computing, 1991

[SMP] D. Gusfield, R. Irving, *The Stable Marriage Problem: Structure and Algorithms,* Foundations of Computing Series, MIT Press, 1989

[MNR] W. Marek, A. Nerode, J. Remmel, *A Theory of Nonmonotonic Rule Systems I,* Annals of Mathematics and Artificial Intelligence, V. 1, p. 241-273, 1990

Real-Time Reasoning in Deadline Situations*

Madhura Nirkhe, Sarit Kraus, Donald Perlis
Department of Computer Science
and
Institute for Advanced Computer Studies
University of Maryland,College Park, MD 20742

Abstract

In deadline situations the salient resource is time: all preparations must be carried out in advance of the deadline. If action is called for, an appropriate plan must be formulated and enacted before the deadline. This puts interesting constraints on the reasoning that goes into forming the plan and its simultaneous or subsequent execution.

Step-logics were introduced as a mechanism for reasoning situated in time. We first describe them in brief. We then show their application to creating a step-logic planner that lets a time-situated reasoner keep track of an approaching deadline as she/he makes (and enacts) her/his plan, thereby treating *all* facets of planning (including plan-formation and its simultaneous or subsequent execution) as deadline-coupled. We use a key example of a tight deadline situation to illustrate the problem and our approach.

*This research was supported in part by U.S. Army Research Office grant DAAL03-88-K0087, and in part by NSF grant IRI-8907122.

Motivation

Hard Deadlines:

Example An automated helicopter pilot with a mission to rescue an injured soldier in time before the advancing enemy patrol reaches the soldier in distress.

An infinite cost of overshooting the deadline

BASIC TRADEOFF: Meta-planning (thinking about the planning process) improves performance, but every second spent on planning is one less second for acting.

The problem

Time taken to plan brings the deadline CLOSER

The agent must account for the passage of time during the *same* reasoning

Step-logics; account for *all* the time taken

Applied here to the planning domain

Deliberation Time

Action occurs in the mere form of thinking or reasoning
Traditionally actions are viewed as separate from the planning

Is planning a different beast?
Just as deliberation over the features of actions will lead to better plans, taking account of the features of planning will lead to more intelligent decisions about the plans.

Routine tasks: Little or no deliberation
Reasoning *about* time-bounded tasks:
Deliberation is required, but is outside the action (real) time-frame

Dudley's planning problem

Novel situation, cannot a priori assign utilities; must think about them in real time

Must meta-plan on-going deliberations vis-a-vis the passage of time

Need: Not an ultimate plan but a plan which evolves in a changing world

Total effort, partial plan formulation, making decisions about available and conceivable alternatives, plan sequencing, plan failure and revision, MUST ALL STAY WITHIN THE DEADLINE, AND MUST ADJUST TO MEET THE DEADLINE.

Step-logics for planning in real-time

- Inferences are characterized by the time elapsed during the inference
- $Now(i)$ and the other time parameters appear in the on-going process of reasoning
- Observations become instant beliefs
- Contradictions are not necessarily bad, they are permitted, and resolved in subsequent steps
- Projections are made about the future in the context of each plan to conjecture the state of the world upon hypothetical execution of actions in the plan
- Inherently non-monotonic formalism, must retract older(incorrect) beliefs in the face of new evidence

Sample Inference Rules

Agent looks at the clock

$$\frac{i : \ldots}{i+1 : \ldots, \mathbf{Facts}(i+1, \{\ldots, Now(i+1)\})}$$

Modus Ponens(MP)

$$\frac{i : \ldots, \mathbf{Facts}(i, \{\ldots, \alpha, \ldots, (\alpha \rightarrow \beta)\})}{i+1 : \ldots, \mathbf{Facts}(i, \{\ldots, \beta\})}$$

Example:

$$\frac{i : \ldots, \mathbf{Facts}(i, \{\ldots, \phi, \alpha, (\alpha \rightarrow \beta), (\beta \rightarrow \delta), (\phi \rightarrow \psi), \ldots\})}{i+1 : \ldots, \mathbf{Facts}(i, \{\ldots, \beta, \psi\})}$$

Related to planning:

Forms the first partial plan:

$$\frac{i : \mathbf{Goal}(G)}{i+1 : \mathbf{Ppl}(p, i+1, \{G\}), \mathbf{Feasible}(p, i)}$$

Sample Axioms

- $Run(T_1 : T_2, Y, L_1 : L_2) \rightarrow At(T_2, Y, L_2)$,
  $T_2 = T_1 + (L_2 - L_1)/v_Y$
- $condition(Run(T_1 : T_2, Y, L_1 : L_2), At(T_1, Y, L_1))$
- $result(Run(T_1 : T_2, Y, L_1 : L_2), At(T_2, Y, L_2))$

Do not require Dudley to figure out *how* to run, this is a routine task, and as such requires only one time step to break down into atomic paces.

AN OVERVIEW OF THE MODULAR UNIX®-BASED VULNERABILITY ESTIMATION SUITE

Jill H. Smith
Wendy A. Winner
Phillip J. Hanes
US Army Ballistic Research Laboratory
ATTN: SLCBR-VL-V
Aberdeen Proving Ground, MD 21005-5066

The Modular UNIX®-based Vulnerability Estimation Suite (MUVES) is the new computing environment for the conduct of vulnerability/lethality studies within the Vulnerability/Lethality Division of the Ballistic Research Laboratory. MUVES employs the latest software technologies both in design and implementation to leverage scarce vulnerability/lethality analyst resources, improve the ability to incorporate methodology advances, provide an audit trail of the analyses, and facilitate configuration management and archiving of analyses. MUVES is a suite of packages that are ANSI C compliant and run on System V® compatible UNIX® platforms. MUVES provides a user-friendly, menu-driven interface for the conduct of vulnerability/lethality analyses. Currently, the compartment-level vulnerability/lethality model, VAMP (Vulnerability Analysis Methodology Program) is implemented under this environment and the stochastic point-burst model, SQuASH (Stochastic Quantitative Analysis of System Hierarchies) is to begin implementation this FY.

®UNIX and System V are trademarks of AT&T.

I. Introduction

The Modular UNIX-Based Vulnerability Estimation Suite (MUVES) is the new software environment under which all vulnerability/lethality analyses conducted by the Vulnerability/Lethality Division of the Ballistic Research Laboratory (BRL) will be performed [1, 2, 3]. MUVES is a very general environment that is designed to evaluate the interaction of a threat with a target where the target information is provided via ray-tracing. Target descriptions built using the BRL Multi-device Graphics EDitor (MGED) [4] are ray-traced via an interface to the BRL Computer-Aided Design (BRL-CAD) package [5].

Although currently only the compartment-level vulnerability/lethality model has been implemented under MUVES, all models in the vulnerability/lethality hierarchy of models will be converted to run under the MUVES environment. MUVES is written in the C programming language and employs state-of-the-art computer programming techniques, such as structured programming, for ease of maintenance and extension. MUVES incorporates a user-friendly menu-driven user interface to facilitate the conduct of vulnerability/lethality analyses and a set of post-processors for the textual and graphical display of results.

II. Background and Goals

The Vulnerability/Lethality Division has a hierarchy of vulnerability/lethality models including the low-resolution compartment-level model VAMP[6], the component-level point-burst model VAST[7] and the component-level stochastic point-burst model SQuASH[8]. All of these models are coded in Fortran and exist in multiple copies within the Division. Each vulnerability analyst modifies the code to perform the specific analyses requested and iterates on this procedure for the target/threat combinations included in a study. The maintaining of multiple copies has lead to configuration control and audit trail problems and the burden of maintaining what should be the same code many

times over. Also, extensions and improvements to the code have not been uniform within the Division. For many years it was felt that a single code could not support the various analyses because of the need to modify or tailor the code to specific analyses. The goal of the MUVES project was to consolidate the code where possible without losing the flexibility to accommodate the different study requirements. Other goals of the MUVES project were to keep the audit trail of the inputs to the associated outputs and facilitate the archiving of the inputs and outputs.

III. Vulnerability Computations

A MUVES analysis has two basic inputs, the threat and the target. The threat information is stored in data files containing the physical characteristics of a particular threat. This may include such things as the velocity, caliber, and mass of a kinetic-energy penetrator, or perhaps the power and wavelength of a laser beam, or any other data required to describe the damage-producing capabilities of that threat.

The target information is stored in several different files, each defining a specific aspect of the target. The target characterization may be thought of in three general categories. The geometry comprises the shape and spatial location of each component, plus names of the components. The relevant physical characteristics of each component (*e.g.*, material, density, reflectivity) are recorded for use in interactions with the threat. Finally, there is the system structure which defines each target system in terms of its constituent components and defines the measures of effectiveness in terms of the system and component functionalities.

A vulnerability analysis consists of determining the effects of a threat against a target. Due to current geometry interrogation techniques, the motion of the threat must be piece-wise linear.

MUVES uses a ray-tracing approach to simulate a threat's trajectory to and (possibly) through a target. The ray-tracing package constructs a path consisting of the geometric information about each component in a trajectory; the threat information is then attached to the first component of the path. Figure 1 shows a simplified representation of a threat path and a schematic of the computations performed on the information along that path.

Each component in a target is assigned to a category. This category is used to select an Interaction Module (IM) appropriate for computing the effects of a specific threat impacting that component. Within this module, several things may occur: threat parameters may be altered, damage may be produced for that component, and new threats may be generated. The threat may then be propagated to the next component, possibly with updated parameters. The interaction will then be computed for the next component. In the interaction module, all damage is recorded as physical parameters (*e.g.*, number of impacting fragments, hole diameter, deposited energy); interpretation of this damage is deferred until later in the process. If new threats are generated, new threat paths must be determined (via ray-tracing), new interactions will occur, and more damage may be produced. This cycle is continued until all threats have exited the target or have been stopped by various components. As shown in Figure 1, all damage is stored until the interactions are completed.

When all damage-producing interactions have ceased, the evaluation phase begins. The recorded damage is sorted for each component, so that all damage to a single component may be evaluated together. An Evaluation Module (EM) is called for each damaged component; the selection of an EM is also determined by the category of the component. These modules compute an engineering estimate of the level of damage to a component based on the physical damage from the interaction module(s). Typically, this estimate is expressed as a value between 0.0 and 1.0 for each component. The exact meaning of this value may differ depending on the method of analysis in use.

These component damage values are combined using the system structure of the target to determine the damage level of each system within the

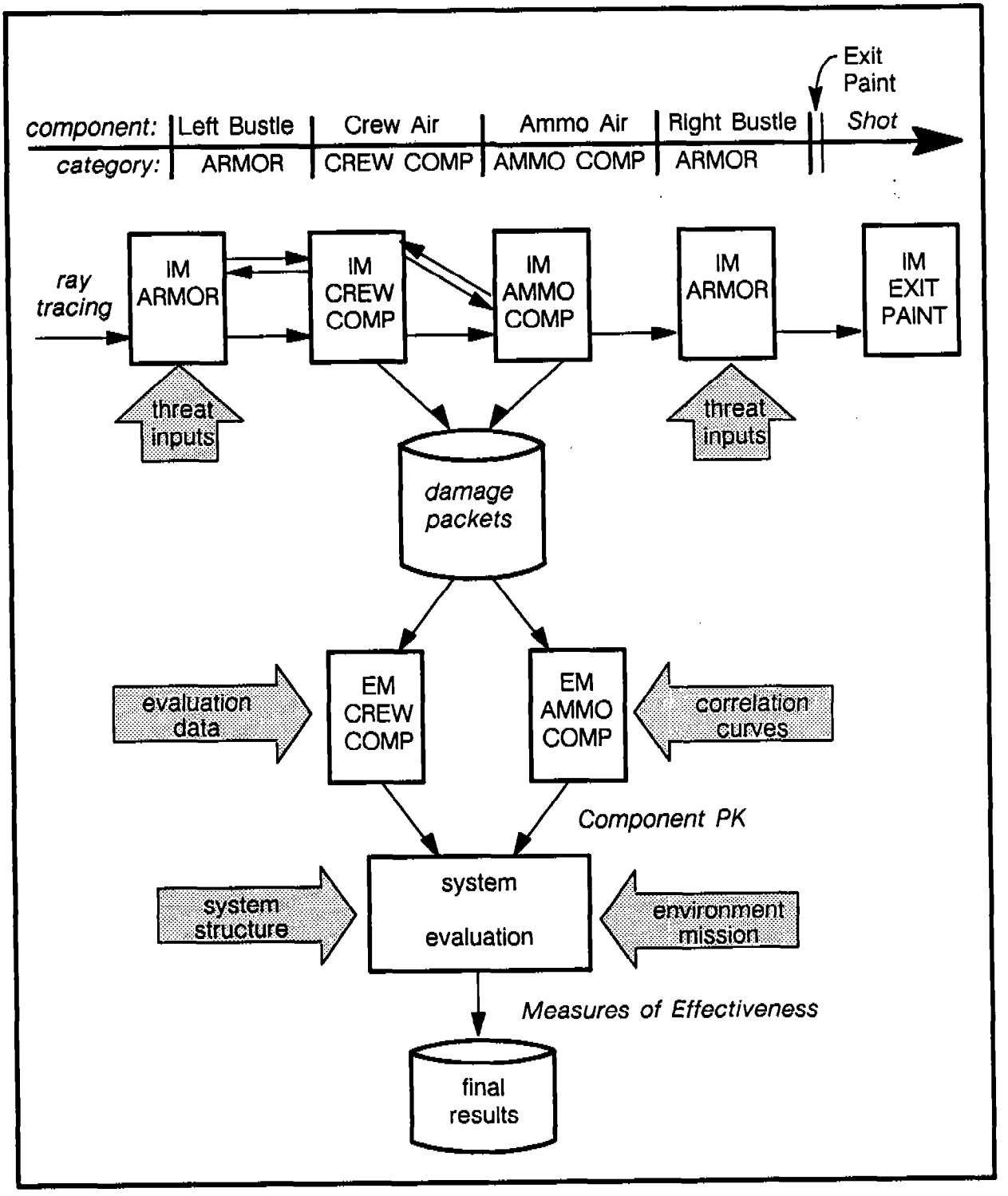


Figure 1. MUVES Vulnerability Computation

target. The target's ability to perform one or more missions may then be assessed using the measures of effectiveness for that target based on the functionalities of its systems.

This process is repeated for each shot in the array requested by the analyst.

IV. Software Packages

MUVES is designed to improve the long-term flexibility of vulnerability methodology development. Software design and structured programming techniques were employed to maintain a high standard of quality for all MUVES code. Basic software functions were defined and their interfaces designed to minimize code redundancy. Software modules with well-defined interfaces were written to perform singular tasks. Modules were then combined into software packages; each package contains software related to a set of similar tasks. For instance, there is a pseudo-random sequence package which provides several random number generators. Random number generator functions in this package return pseudo-random numbers from a variety of distributions for use in various situations. Although future applications may require the use of random number generators which are not presently in this package, the modular nature of the code facilitates enhancing the functionality.

There are three primary classes of software packages: (1) general-purpose, (2) MUVES-specific, and (3) model-specific. Figure 2 lists the software packages by category.

General-purpose software packages handle tasks which are common to software environments other than MUVES. For example, there is a doubly-linked list package (Dq) which handles the creation, insertion, traversal, and deletion of nodes in a queue where each node has forward and backward pointers to other nodes. There is also a package which performs piece-wise linear interpolation of tabular data. Another package provides the interface to one or more ray-tracing slave processes running on the same host or network accessible hosts (Rt). Yet another package provides an interface to the terminal handler for controlling input/output processing (Tc). This is only a sample of the general-purpose software within MUVES. Of the approximately 130,000 lines of MUVES code written, about 33% is general-purpose in nature.

MUVES-specific software packages are common to the vulnerability/lethality assessment process and may be used for any vulnerability/lethality model. These packages form a standard library which may be applied to a general class of threat-target interaction models. Examples include an interactive user interface (Ui), a threat-component interaction package (Im), and a final analysis results I/O interface (Fr). Software packages (such as the user interface) might require some additional code for a new model but would utilize these basic modules. For instance, for a each new model, menu entries would have to be added to the user interface; however, the manipulation and behavior of the menus would remain the same. Approximately 44% of the code is MUVES-specific.

Model-specific packages include all software packages which are unique to a particular methodology. These packages are required to complete the implementation of the model and to postprocess final results. It may also be advantageous to provide some software to set up model inputs. As previously stated, the only model currently available under the MUVES environment is the compartment-level model. For this model, the *compart* package contains the crucial Interaction and Evaluation Modules which assess and evaluate damage, respectively. Four postprocessors are provided to examine results in tabular and graphical formats; additional postprocessors will likely be added as analysts identify various needs. It is important to notice that this model only represents 23% of the total MUVES software; the remaining 77% has general applicability to other vulnerability/lethality models.

As additional vulnerability/lethality models are implemented under this environment,

GENERAL PURPOSE

NAME	DESCRIPTION	LINES
Db	Database server	466
Dq	Doubly-linked queues	1253
Dx	Inter-process data exchange	5077
Er	Error handling	1674
Hm	Hierarchical menus	4696
In	Interpolation	3013
Io	Input Operations	1776
Lk	Resource Locking	1776
Mm	Dynamic memory manager	3256
Nm	Name pools	1173
Rn	Pseudo-random sequences	1303
Rr	Reusable rays	2859
Rt	Target geometry ray-tracing	5751
Sa	Shot array generator	1198
Sc	Terminal screen manager	1229
Tc	Terminal I/O control	673
Vm	Vector math	1323
Uc	Units conversion	346
tools	Software development tools	4633
	SUBTOTAL	43475
	TOTAL	130602
	% OF TOTAL	33.29

MUVES-SPECIFIC

NAME	DESCRIPTION	LINES
Ap	Analysis parameters	4795
At	Post-shot utility assessment	542
Cd	Component damage records	3912
Dd	Data dependencies	6285
Em	Component damage evaluation	462
Fr	Final analysis results I/O	3736
Im	Threat-component interaction	1484
Ir	Intermediate analysis results I/O	4432
Se	Contextual system evaluation	5809
Ti	Threat-target interaction	2627
Vu	Weighted-view utility assessment	822
Ui	Interactive user interface	17359
muverat	Analysis control program	541
data	Data files for installation testing	4595
	SUBTOTAL	57401
	TOTAL	130602
	% OF TOTAL	43.95

COMPARTMENT APPROXIMATION METHOD

NAME	DESCRIPTION	LINES
compart	Compartment Model	23222
cellxcell	Cell-by-cell file (Final Results)	1031
colorsil	Color silhouette (Final Results)	1264
input_tools	Tools for setting up input files	1366
ir2ascii	Intermediate Results Converter	664
siv	Summary, IUA, and View averages	2179
	SUBTOTAL	29726
	TOTAL	130602
	% OF TOTAL	22.76

Figure 2. MUVES Software Packages

the relative percentages of MUVES-specific and model-specific code is expected to rise in comparison to general-purpose code. The stochastic point-burst model is expected to require more model-specific packages than the compartment-level model.

V. User Interface

One of the largest packages in MUVES is the User Interface (Ui). This package provides a menu-driven environment in which an analyst specifies the parameters of a vulnerability analysis by selecting menu entries and entering information via the keyboard for both required and optional inputs. The analyst's task is eased because the available selections are clearly visible in the menus.

The User Interface automatically maintains a record of the selections made in the course of an analysis. This session information may be loaded at the start of a new analysis to repeat a previous analysis or to run an analysis which differs only slightly from it. The session information is part of the audit trail available for every analysis. The User Interface maintains a record of every input file used during an analysis to prevent accidentally over-writing of these files.

The User Interface also has access control lists so that an analyst may prevent unauthorized access to files used for sensitive projects.

VI. Advantages

MUVES has been written in the C programming language to be portable across a variety of hardware platforms. The code conforms to the American National Standards Institute (ANSI) C standard [Federal Information Processing Standard (FIPS) 160] and the IEEE Standard Portable Operating System for Computer Environments (POSIX). Compliance with these standards promotes longevity of the code.

All code changes are monitored and documented. An audit trail of these changes is saved using maintenance and enhancement tracking tools [Source Code Control System (SCCS)]. The Division's algorithms for assessing and evaluating vulnerability/lethality damage have been closely scrutinized. Recommended improvements to existing algorithms and suggestions for new algorithms are being incorporated only after consulting a BRL panel of vulnerability experts and terminal ballisticians. Algorithms are well-documented within the code as well as in the MUVES Analyst's Guide [2] and individual BRL reports.

Optimizing the capabilities of today's distributed computing assets (*e.g.*, desktop workstations, mini-supercomputers, *etc.*) has been achieved by providing the ability to divert computationally intensive, ray-tracing tasks to higher-performance, network-accessible, computing assets. At the analyst-level, ray-tracing information may also be captured to a file and re-used at a later date. Thus, the time required for target geometry interrogation can be significantly reduced. Preliminary use of reusable ray traces has been shown to reduce analysis run-times by a factor of five.

VII. References

[1] Phillip J. Hanes, Karen Ross Murray, Douglas A. Gwyn, and Helen R. Polak, "An Overview and Status Report of MUVES (Modular UNIX-based Vulnerability Estimation Suite)," *Ballistic Research Laboratory Memorandum Report No. 3679, July 1988.*

[2] Phillip J. Hanes, Scott L. Henry, Gary S. Moss, Karen R. Murray, and Wendy A. Winner, "Modular UNIX®-based Vulnerability Estimation Suite (MUVES) Analyst's Guide," *Ballistic Research Laboratory Memorandum Report*, in preparation.

[3] Douglas A. Gwyn, "Modular UNIX®-based Vulnerability Estimation Suite (MUVES) Administrator's Guide," *Ballistic Research Laboratory Memorandum Report*, in preparation.

[4] Keith A. Applin, Michael J. Muuss, and Robert J. Reschly, "Users Manual for the BRL-CAD Graphics Editor MGED," US Army Bal-

listic Research Laboratory, Draft copy, 6 October 1988.

[5] Michael J. Muuss, Phillip Dykstra, Keith Applin, Gary Moss, Paul Stay, and Charles Kennedy, "The Ballistic Research Laboratory CAD Package Release 3.0 - A Solid Modelling System and Ray-Tracing Benchmark," edited by Donald Merritt, SECAD/VLD Computing Consortium, US Army Ballistic Research Laboratory, 1 October 1988.

[6] C. L. Nail, T. E. Bearden, and E. Jackson, "Vulnerability Analysis Methodology Program (VAMP): A Combined Compartment-Kill Vulnerability Model," *Computer Sciences Corporation Technical Manual CSC TR-79-5585*, October 1979.

[7] C. L. Nail, "Vulnerability Analysis for Surface Targets (VAST) - An Internal Point-Burst Vulnerability Assessment Model - Revision I," *Computer Sciences Corporation Technical Manual CSC-TR-82-5740*, August 1982.

[8] Aivars Ozolins, "Stochastic High-Resolution Vulnerability Simulation for Live-Fire Programs," *The Proceedings of the Tenth Annual Symposium on Survivability and Vulnerability of the American Defense Preparedness Association*, May 1988.

A Logical Framework for Operations on Distributed Data *

P. Broome and B.D. Broome
US Army Ballistic Research Laboratory
Aberdeen Proving Ground, MD
21005-5066

Abstract

In this paper we consider logic programming as a means of both computing and formulating complex queries in the same system. These concepts are applied to a medium sized database. In particular, we establish a term representation of the data used in a prototype battlefield information system and conceptually extend this database with rules. We develop browsing operations for that system by logically combining constraints.

An extended language on binary predicates with richer operations is considered. In this language, programs and queries have mathematical properties that can be specified as equations between relations. These equations support program transformations that improve query efficiency. This work increases the likelihood of performing declarative operations on distributed data.

*The authors thank George Hartwig, Eric Heilman, Ken Smith, James Lipton and Morton Hirschberg. This report is a revision of BRL-MR-3882.

1 Introduction

Database management systems have become widely recognized as a means of sharing and maintaining data in a way that avoids redundancy and inconsistency. They allow the user to insert, delete and modify data and perform simple queries with a minimum of effort.

In recent years, however, the use of database systems has been extended to more and more complex applications. Databases address not just the predictable information required by a personnel department of a company, but also the less predictable information required by an object oriented simulation, an expert system, or a battlefield commander. Techniques developed with business applications in mind do not always provide the query flexibility required. Further, they do not extend themselves easily to take advantage of rapidly developing technologies like parallel computation and automatic program transformation.

Logical databases are very attractive for maintaining and manipulating knowledge and are predicted by some to be the data management system of the future[1]. Reasons for this prediction are that the approach is: well founded, as it is based on logic; cohesive, as it allows data structures, queries and computations in a single notation; declarative and therefore non-sequential, providing more potential for tapping the faster computing speeds of parallel processors. These features can greatly improve program maintenance, reliability, generality and efficiency.

In this project we select an existing distributed fact base and reformulate it as a logical database. Next, we construct some sample queries. Finally, we address possible query transformations and their impact on the efficiency of the associated queries. This approach allows evaluation of the logical database approach: the relative ease of development, query flexibility and efficiency. These issues are addressed in this paper. Further, the dynamic nature of the knowledge base selected allows us to examine compromises between absolute logical correctness and conclusions based on imperfect, incomplete, or changing data. Future work will examine this problem, as well as data visualization and query scheduling.

2 The Information Distribution System

Battlefield management has been identified as a major thrust for future Army technological development[2]. Here we find a prime example of the

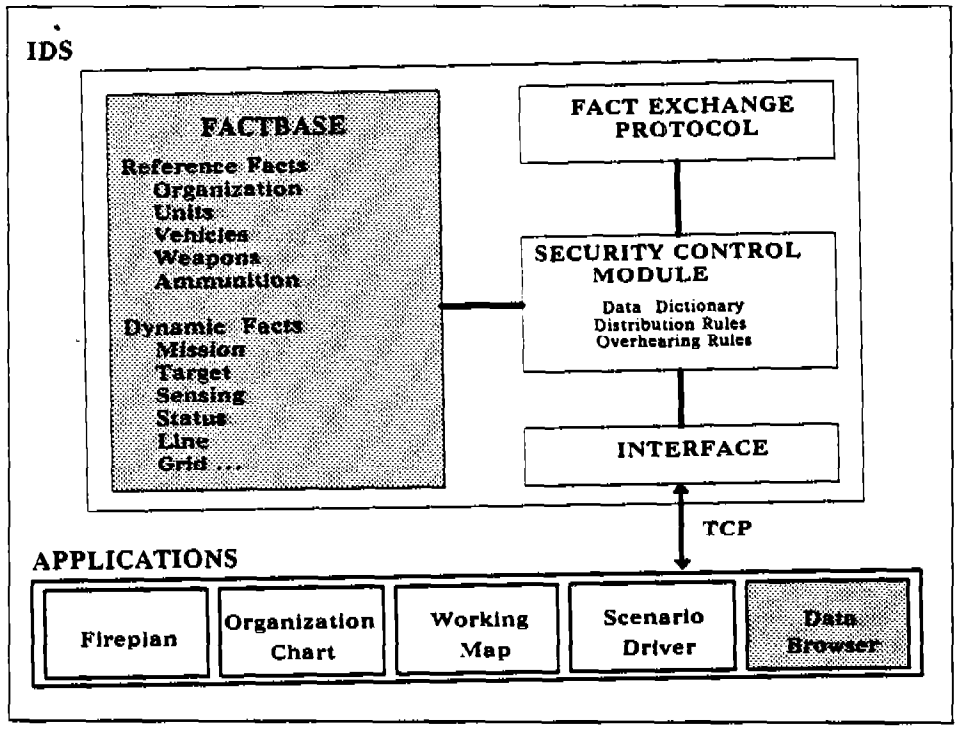


Figure 1: The Information Distribution System. (For this project, browsing operations are being developed to query the FACTBASE.)

need for both query flexibility and efficiency. In a highly dynamic, unpredictable and hostile combat environment, it is crucial that queries be easily formulated and quickly resolved.

The Information Distribution System (IDS) was developed as an experimental prototype to evaluate various data abstraction and distribution technologies for automatically distributing information to and among fighting level forces. It assumes low bandwidth communications in the tactical combat environment. Specifically, it addresses how to insure required battlefield information is available at the various locations where the battlefield management function is performed. As part of this prototype, a FACTBASE was developed, which accommodates the wide variety of information required at brigade and below. Various application programs then access the FACTBASE information through the IDS interface [3]. Figure 1 illustrates the IDS structure and its relationship to the various IDS applications.

The FACTBASE consists of various C programming structures and has a small query language with a C-like syntax. Some facts are relatively static over time, while others are more dynamic [4]. The information in the FACTBASE is complex, requiring all three possible database schemes: hierarchical, for the organizational structure; network, for the communications connectivity; and relational, for the logistics data found in TO&E or equipment manuals. This FACTBASE serves as the foundation for our logical database.

3 Logical Databases

Logic is a branch of mathematics which allows the explicit expression of goals, knowledge, and assumptions. It supplies a foundation for deducing conclusions from premises and for determining validity and consistency. Logic programming is a formal system for specifying objects and relations between objects. It departs radically from the mainstream of computer languages. It is not derived from a physical machine's instruction set, but is instead founded on an abstract model based on first order logic[5]. A logical, or deductive, database is a set of facts that are combined with a set of rules to allow new facts to be inferred and new relationships to be defined. A logical database is firmly and declaratively founded on a small, but powerful, set of primitives. This characteristic increases reliability, confidence, and efficiency.

Some of the dominant areas of interest in logic programming are program correctness, program optimization, parallelism and program synthesis. Major applications of logic programming have been made to intelligent databases, natural language processing, computer aided design, molecular biology, and high level compilation.

Logic programming attempts to apply the rigor of formal logic to complex, computer-based systems that lack such logical foundations. It is an ideal that has not been, and may never be, realized on an existing machine. One approximation is given by the programming language, Prolog. Prolog compilers have become very efficient primarily as a result of work by Warren and his colleagues[6]. This application is being developed in Prolog.

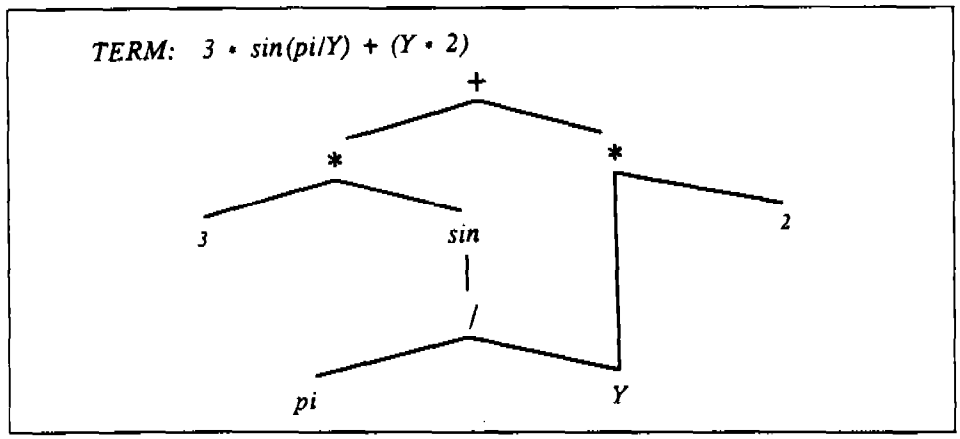


Figure 2: A graphical depiction of a term.

4 Developing a Logical FACTBASE

We began this project by constructing a parser and translator to transform the IDS FACTBASE into equivalent logical relations, which we refer to as the Logical FACTBASE. The result of the translation is a collection of approximately 30,000 Prolog clauses. This representation can include networks, hierarchies and relations. For the initial phase of the project, we have confined ourselves to the static portions of the database, intending to address the dynamic portions in the future. The static portions include the general unit or system properties while the dynamic portions include such changing values as unit location or assignment.

The founding data structure for the database is the term, made up of variables and constants. *Variables* are represented by character strings beginning with an upper case character. Special characters and strings beginning with lower case characters are *constants.* As Figure 2 illustrates, a term may be thought of as a tree-like structure with leaves that are variables or constants (like 3, pi, Y or 2 in Figure 2). The root and internal nodes of the graph are constants and are called *function symbols* ($+, *, sin$ and /). The root (+) is the *principal function symbol.* It is important to note that function symbols are passive, syntactic objects without any implied interpretation.

More precisely, a *term* is either a variable, a constant, or a function symbol with arguments that are terms. The most general term is simply

a variable. A term whose leaves are all constants is called a *ground term.* In the usual Prolog system, constants are stored only once and all other occurrences are simply pointers to the centrally-stored constant. Similarly, if a variable occurs twice in a term, both occurrences refer to the same variable (like Y in Figure 2). Thus, a term is not really a tree but a directed, acyclic graph, that is, a tree with shared branches. This sharing can mean significant savings in storage and is a side effect of the unification algorithm, discussed in the next section.

One special kind of term is the list. A *list* is made up of a nested sequence of pairs indicated with the period as principal function symbol. For example, a list of the first five integers is $.(1,.(2,.(3,.(4,.(5,[\ ])))))$, where we are representing the empty list with [ ]. More conveniently, we can represent this list as $[1,2,3,4,5]$.

Intuitively, a term may make up an entire fact or it may be the argument in a rule stating a fact. Terms also play the role of arrays, pointers, and record data structures.

A *rule* is the fundamental statement in a logic program or logical database. A rule has a head and body separated by ': −'; it ends with a period. The head contains at most one term, and the body contains zero or more terms separated by a comma. We can read a rule declaratively, that is as a statement of fact. For example,

$$P :- Q, R.$$

means that P is true if Q is true and R is true. A rule is also called a *clause.* A *unit clause* is a clause in which the body is empty. A *logic program* is a set of clauses.

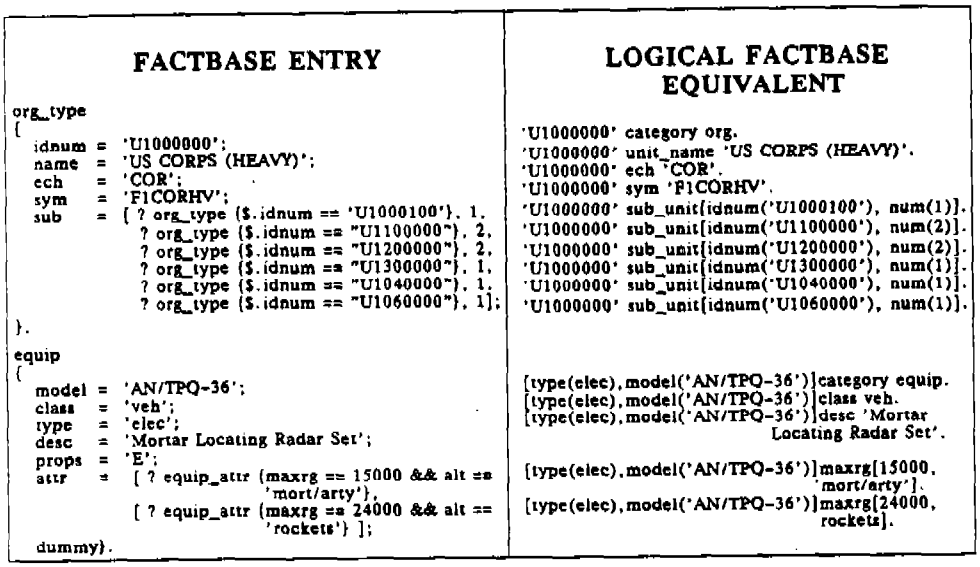
The IDS data was translated into unit clauses whose principal function symbols have two arguments. These define proper binary relations and are to be read as statements of fact. An example would be the clause

$$ech('U1000000','COR').$$

This is a unit clause whose head is a single term. The principle function symbol is *ech* and it has two arguments, *'U1000000'* and *'COR'*. The function symbol can also be placed between its arguments, in *infix* form, as

$$'U1000000'\ ech\ 'COR'.$$

Binary representation was chosen for several reasons. First, it is simple; database entries are easily written, easily searched, and can often be read



FACTBASE ENTRY

```
org_type
{
  idnum =  'U1000000';
  name  =  'US CORPS (HEAVY)';
  ech   =  'COR';
  sym   =  'F1CORHV';
  sub   =  [ ? org_type ($.idnum == 'U1000100'), 1,
             ? org_type ($.idnum == "U1100000"), 2,
             ? org_type ($.idnum == "U1200000"), 2,
             ? org_type ($.idnum == "U1300000"), 1,
             ? org_type ($.idnum == "U1040000"), 1,
             ? org_type ($.idnum == "U1060000"), 1];
}.
equip
{
  model =  'AN/TPQ-36';
  class =  'veh';
  type  =  'elec';
  desc  =  'Mortar Locating Radar Set';
  props =  'E';
  attr  =  [ ? equip_attr (maxrg == 15000 && alt ==
                           'mort/arty'),
             [ ? equip_attr (maxrg == 24000 && alt ==
                           'rockets') ];
  dummy).
```

LOGICAL FACTBASE EQUIVALENT

```
'U1000000' category org.
'U1000000' unit_name 'US CORPS (HEAVY)'.
'U1000000' ech 'COR'.
'U1000000' sym 'F1CORHV'.
'U1000000' sub_unit[idnum('U1000100'), num(1)].
'U1000000' sub_unit[idnum('U1100000'), num(2)].
'U1000000' sub_unit[idnum('U1200000'), num(2)].
'U1000000' sub_unit[idnum('U1300000'), num(1)].
'U1000000' sub_unit[idnum('U1040000'), num(1)].
'U1000000' sub_unit[idnum('U1060000'), num(1)].

[type(elec),model('AN/TPQ-36')]category equip.
[type(elec),model('AN/TPQ-36')]class veh.
[type(elec),model('AN/TPQ-36')]desc 'Mortar
                                Locating Radar Set'.

[type(elec),model('AN/TPQ-36')]maxrg[15000,
                                'mort/arty'].
[type(elec),model('AN/TPQ-36')]maxrg[24000,
                                rockets].
```

Figure 3: An example of an IDS fact and its translation to proper relation form.

as if they were sentences. Second, with this approach, there is no loss of computational power. Rules on binary relations can compute anything that rules on n-ary relations can compute[7]. Finally, the method we use later for transforming queries requires that the relations have two arguments[8].

Figure 3 illustrates the translation of two FACTBASE entries from their original C structure into their logical representation. The C structures typically consist of a fact type, followed by a series of subfield identifiers which are associated by = with a subfield value. In the example, *org_type* and *equip* are both fact types. Looking more closely at *org_type*, *idnum* is a subfield identifier, and its value is *U1000000*, a unique unit identification code developed for IDS applications. A unit clause is asserted for each of these triples, with the subfield identifier becoming the binary relation. The fact type and subfield value are the relation's arguments. A subfield value of *'E'* indicates an empty field and is not translated. In the example, one organizational fact is translated to 10 unit clauses. Their principal function symbols are *category, unit_name, ech, sym and sub_unit.* Each relation has 2 arguments. The sub_unit function, for example, has 2 arguments: parent unit id; and a list of 2 terms, the subunit and its number of occurrences.

After the translation was accomplished, a small parser was written in

Prolog, in which the operator precedence, position, class and associativity were established. The binary relations resulting from the translation were all defined in infix form.

Finally, the database was extended with new relations. These relations were not part of the organizational or logistical structure, but were created to help form new queries. For example, as illustrated in Figure 3, we know the maximum range of our weapons. We can extend the data by defining what we mean for a given distance, R, to be within firing range of a particular weapon of type T and model M:

$$\begin{array}{ll} [T, M] & can_fire_at_targets_at_range\ R :- \\ & [T, M]\ maxrg\ [Range, _Alt], \\ & Range\ >\ R. \end{array}$$

This new relation could be useful in searching for the right weapon to use against a given target. The new relations extend the translated database entries to a conceptually larger database. They are, in fact, rules that assist in formulating queries. This brings us to our next topic.

5 Querying the Logical FACTBASE

The next step in the application was to construct some queries. The fundamental tools for querying are unification and backward inferencing. We therefore begin this section with a brief explanation of these basic procedures.

The *unification algorithm* is a solution procedure that derives values for variables from an equation between two terms. Given two terms S and T the unification algorithm determines values for variables as follows:

- if S and T are both constants then unification succeeds if they are identical and fails if they are different.
- if S is a variable, then the value for S is $S = T$. (Symmetrically, if T is a variable, then the value for T is $T = S$.)
- if S and T are more general terms with the same function symbols, then the solution is determined by corresponding unification of their arguments.
- if S and T are more general terms with different function symbols then unification fails.

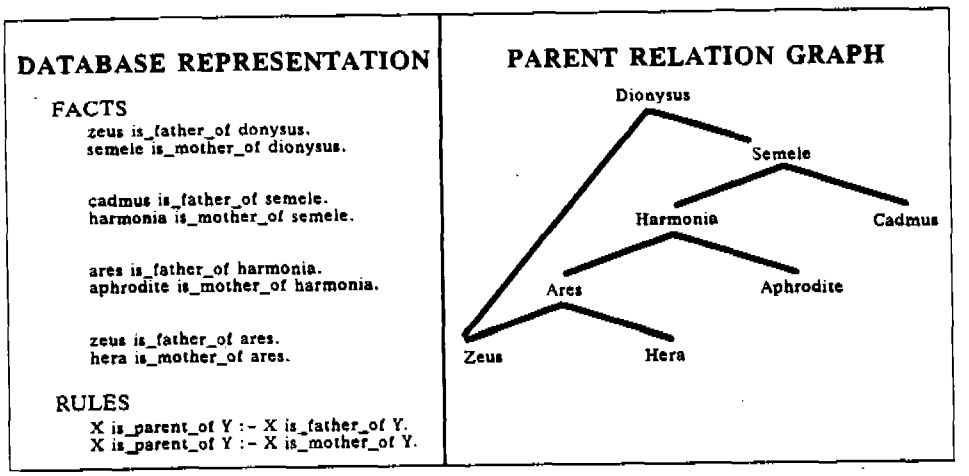


Figure 4: Representing the parent relationship in a logical database.

Unification, then, can be applied to extract components of clauses. Figure 4 illustrates a familiar example of a family database [9]. In this example, consider unifying the two terms X *is_father_of ares* and *zeus is_father_of* Y. From

$$X \; is_father_of \; ares = zeus \; is_father_of \; Y$$

we would conclude that a value for X is $X = zeus$ and a value for Y is $Y = ares$.

The second fundamental tool is *backward inferencing*, which is essentially the application of one rule to a goal, reducing it to a conjunction of subgoals. Inferencing allows us to arrive at conclusions from facts and rules. For example, in Figure 4, *zeus is_parent_of* Y can be reduced to *zeus is_father_of* Y using the very first rule allowing us to eventually infer that $Y = dyonysus$. If we look for more solutions, we find that $Y = ares$ also satisfies the query.

A *goal*, or in our case a database query, is a clause with an empty head. This goal is a conjunction of subgoals which is solved by solving all subgoals. Each subgoal is solved by unifying it with the head of a clause in the database. This creates values for variables. A single backward inference reduces this subgoal to another conjunction of subgoals until reaching the subgoal *true*, which is trivially solvable. In Prolog, subgoals are solved in sequential, left to right order and clauses are chosen in top to bottom order

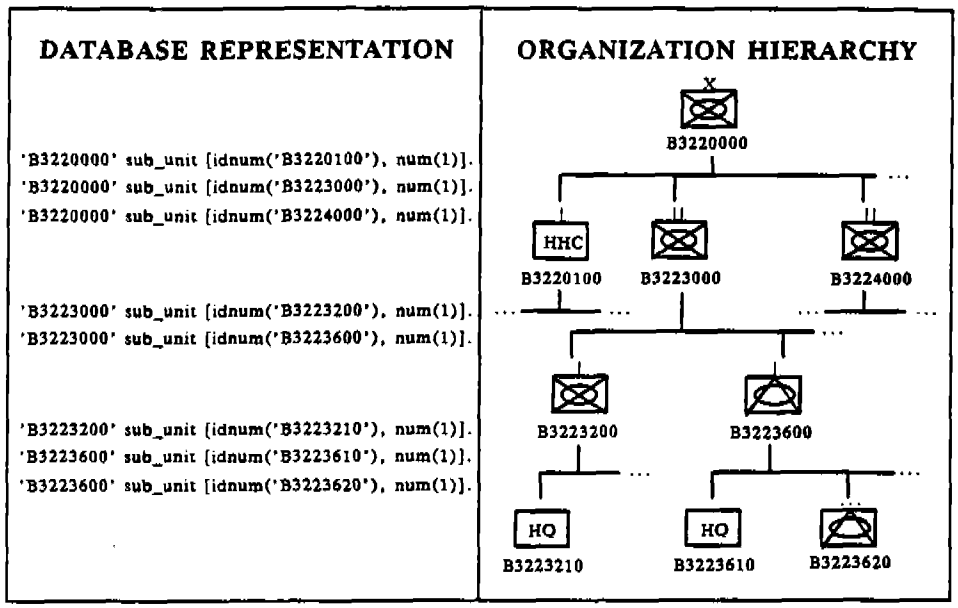


Figure 5: Representing the subunit relationship in the Logical FACTBASE.

with backtracking to find additional solutions. Again, looking at Figure 4, we can determine who are the parents of Semele by solving the goal

$$:-X\ is_parent_of\ semele.$$

This unifies with the head of the first rule, yielding $X\ is_father_of\ semele$. The solution for X in that subgoal is $X = cadmus$. Alternatively, the goal resolves to $X\ is_mother_of\ semele$, in which we find a alternate solution $X = harmonia$.

In Figure 5 we extend this technique to the FACTBASE data, using the subunit relation somewhat like the parent relation. *A subunit B* means that A and B are members of the *subunit* relation, with A being the parent unit.

Once the database has been established, a number of queries can be solved without any programming, by the selective placement of constants and variables in goals. Prolog attempts to unify the goal with unit clauses in the database. For example, using the data in Figure 5, we may identify all the subunits of B3220000, with the simple query, *'B3220000' sub_unit X*. Further, all relations defined with unit clauses can be queried in either

direction. This is a powerful aspect of the unification algorithm, for it allows us to answer questions about the converse of a relation in the database as well as about the relation itself. For example, the subunit relation has been defined, so we have immediate access to its converse, the parent relation. That is, we can identify the parent unit for B3223600 through the query, *X sub_unit [idnum('B3223600'), _num]*. Similar queries can be made for all relations established in the database. Queries solved with a single unification are satisfied almost immediately.

As indicated previously, more complex queries may require the definition of new relations. Suppose we wish to know whether B3223610 is under the control of B3223000. In this case, we would like to know if B3223610 is a subunit of B3223000, or if it is a subunit of a subunit of B3223000, etc. We define the *controls* relation recursively as follows:

$$\begin{aligned} A\ controls\ B :- &\ A\ subunit\ B. \\ A\ controls\ B :- &\ A\ subunit\ C, \\ &\ C\ controls\ B. \end{aligned}$$

Now, we may query with the goal *'B3223000' controls B3223610'*. Prolog verifies that there is a path through the organization graph in Figure 5 from B3223000 to B3223610 through B3223600, returning the answer *true*.

6 Query Transformations

Finally, we address possible query transformations and their resulting impact on the efficiency of the associated queries. Sometimes the most obvious expression of a query is not the most efficient for implementation, as illustrated in the example below. One of the benefits we hope to derive from this logical approach to computation is to be able to state queries in a straightforward manner, and then reliably transform these queries to optimize their execution.

The solution procedure for a query starts by unifying the goal with the head of a clause to determine values for variables. This environment is used to solve each subgoal of the body in turn. If any subgoal is unsolvable then alternate clauses are applied by backtracking to create possible alternate paths. A solution can be found more efficiently if the search can be correctly constrained in the appropriate direction. But note that an overconstrained system may be unsolvable.

Consider the problem of searching for a path through a graph described by a relation R. This is essentially asking if the two endpoints $\langle x, y \rangle$ of the

graph are members of the transitive closure R^+ of R. A pair is a member of the transitive closure of R if either the pair is in R or there is an intermediate point z such that $\langle x, z\rangle$ is in R and $\langle z, y\rangle$ is in R^+. In symbols this is written as

$$R^+ = \{\langle x, y\rangle | \langle x, y\rangle \in R \text{ or } \exists z, \langle x, z\rangle \in R, \text{ and } \langle z, y\rangle \in R^+\}.$$

Operationally, R^+ is the exhaustively repeated application of R.

The *controls* relation, that is the transitive closure of the *subunit* relation, provides a perfect example of how we can improve the efficiency of the solution procedure by transforming the query. In this example, we say that *A controls B* if there is a path from A to B in the graph formed by the *subunit* relation. The *controls* definition naturally schedules subgoals from the top of the command hierarchy downward. As illustrated in the following example, this schedule is inappropriate and inefficient for the database as structured. A bottom up search would have been better.

Consider the command hierarchy depicted in Figure 6. In this graph, the lines indicate the *subunit* relation, with higher nodes indicating parent units and lower nodes their subunits. This simplified example allows us to limit the *controls* relation to two levels. That is, a unit *controls* its subunits and its subunits' subunits. To determine if *B*3224600 is under the control of *B*3220000 we find an intermediate unit *V* such that *B*3220000 *subunit V* and *V subunit B*3224600. Efficiency greatly depends on which subgoal is selected first. If we start with the goal *B*3220000 *subunit V* then we have multiple solutions, requiring us to travel through the tree, first through node B3220100 and its subunits, then through node B3223000 and its subunits, and finally to our solution point under B3224000. On the other hand, if we start with the goal *V subunit B*3224600, it has a unique solution, quickly generating our solution path.

The reason that the second subgoal should be chosen first is that the converse of the *subunit* relation, denoted ($subunit^\smile$), is a function. Each unit has exactly one parent unit. Thus it would be much more efficient to carry out the search in this order, as each choice would be unique. We, therefore, transform the query to find a path in the tree with

$$controls^\smile = (subunit^\smile)^+,$$

denoting transitive closure with $^+$. The *subunit* relation does indeed define a tree, so *A controls B* is reversible. Since the converse of *subunit* is a function, the paths through the tree can be most efficiently found by

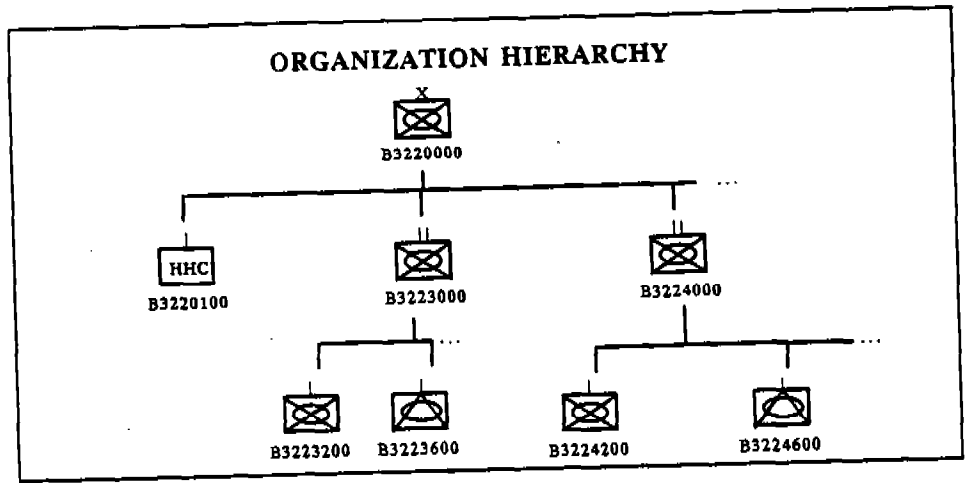


Figure 6: A sample command hierarchy graph.

searching up the tree instead of top down. The bottom up search requires no backtracking. Here we note that it is the nature of the *subunit* relation that suggests this transformation. For the large organizational structure in the IDS, the bottom up solution of a sample query was immediately solved whereas the corresponding top down query took more than an hour.

The reversibility of the unification algorithm is what allows us to represent converse relations. Some knowledge about reversibility can save a great deal of computation time. Searches both up and down the hierarchy in the originally defined IDS database would have required that we add the converse relations to the data, essentially doubling the storage requirements for the *subunit* relation. This trades storage for time, and sacrifices modularity and maintainability. With our new approach, the tradeoff is unnecessary.

On the other hand, while queries are completely reversible when solved with unit clauses, termination is unpredictable in general. In Prolog, some queries that can be easily solved in the forward direction may not terminate in the reverse direction. In addition, some operations in Prolog only have meaning when all arguments are ground terms. Attractive solutions to these problems are emerging from research in constraint logic programming and higher order extensions to logic programming[8,10]. These approaches solve bigger classes of problems by giving declarative extensions to some operations in logic programming such as negation, inequality, and ordering.

7 Future Work

Future work will emphasize three main areas: first, the notoriously difficult problem of synchronizing data updates with data queries, including determining constraints that can maintain integrity; second, methods of pictorially representing the relations in the Logical FACTBASE and the associated queries; and, finally, further query optimizations.

In Section 2 we indicated that the static portions of the FACTBASE were translated first. The dynamic data would be translated in future. This is because logic programming with a set of clauses does not accommodate axioms that may be modified in the middle of a deduction [11]. An attractive compromise, however, can be derived from a thorough treatment of binary relations[8,12]. Accepting the fact that change is an integral part of our distributed database, we concentrate on cleanly separating the abstract portions of our relations, the rules, from the facts. That is, we separate the program from the data. Once this is accomplished, the algebra of equations between relations is an appropriate formalism and an ideal foundation for query optimizations that hold independently of the data. The FACTBASE information will be set aside as an area designated to be modified. Queries operate on a snapshot of the database without attempting to maintain a notion of logical truth. Equations between combinations of relations hold independently of the data. We extend this concept and further partition the data into distinct relations to represent partitions of the database such as *subunit* and *owns_equipment*. Then we can pass these relations along as arguments to the previous operations. This adds another level of generality to the query language so that generic operations can be defined and applied to portions of the database or to other predefined operations on the database.

Secondly, we will experiment with ways of pictorially representing the relations in the logical FACTBASE and the associated queries. There is a close relationship between proper binary relations and combinatorial graphs. This strongly suggests a visualization technique for logical databases that may allow the casual user to bypass much of the notation and abstract syntax.

Finally, we will explore schedules for constraints as binary relations. This includes further methods for reordering subgoals, merging recursions, and propagating constraints. There is also a close relationship between declarative languages and parallelism. The mathematical properties of program operations such as associativity and commutativity indicate that order of some computations can be ignored.

8 Conclusions

We selected an existing distributed fact base and reformulated the static portion as a logical database of binary relations. A parser of C structures was built and a translator constructed to separate the information into relations for querying and updating. We identified the operations required to develop our queries. Finally, some high level, decision critical queries were formulated to test flexibility. Simple query transformations were applied to improve efficiency.

At the end of this first phase, we find that the logical database has a relatively simple structure. Once its structure was established, a number of queries were immediately available through unification. These were satisfied almost instantaneously. More complex queries were built using rules as statements of a recursive programming language, with power, flexibility and limited reversibility. The approach to date puts us in a position to begin examining the effort required to develop queries and the computation time required to perform those queries on the data one might expect in a battlefield environment.

A single inference is comparable to one statement executed in a procedural language. The number of inferences involved is critical to efficiency and may be very large if the order of subgoal selections is not carefully controlled. Prolog queries are not always reversible, partly because subgoals are chosen in a predetermined order. This makes naive queries more difficult to formulate and implies that careful attention must be paid to the solution procedure when scheduling subgoals. A view of programs as proper binary relations, along with an associated set of equations between relations, is a step toward understanding and harnessing the limited reversibility of logic programs.

The primary claim of this work is that logical databases are a convenient vehicle for the management of battlefield information. The primary advantages are improved program maintenance, reliability, efficiency and generality. While no system can perfectly represent a distributed database, we have begun applying a logical model that is an attractive compromise, viewing both queries and data as proper binary relations. The query language we will use has a set of operations with an associated theory. This theory is independent of the data and should be unaffected by its volatility.

References

[1] J. W. Lloyd, *Foundations of Logic Programming*, Springer-Verlag, Berlin, 1987.

[2] Science Applications International Corporation , "The Army Technology Base Master Plan," SAIC-MCDC 89-03002, April 1989.

[3] Samuel C. Chamberlain, "The Information Distribution System: IDS - An Overview," Ballistic Research Laboratory, BRL-TR-3114, Aberdeen Proving Ground, MD, August 1990.

[4] George W. Hartwig, *The Information Distribution System: The Factbase*, Ballistic Research Laboratory, Aberdeen Proving Ground, MD, (To be published).

[5] L. Sterling & E. Shapiro, *The Art of Prolog: Advanced Programming Techniques*, MIT Press, Cambridge, 1986.

[6] D.H.D. Warren, "An Abstract PROLOG Instruction Set," SRI International Technical Note 306, 1983.

[7] J. Sebelik and P. Stepanek, "Horn Clause Programs for Recursive Functions," in *Logic Programming*, K. L. Clark and S.-A. Tarnland, ed., Academic Press, London, 1982.

[8] Broome, Paul, "Program Transformation with Abstract Relation Algebras," BRL-MR-3784, October 1989.

[9] R. Kowalski, *Logic for Problem Solving*, North Holland, New York, 1979.

[10] J. Cohen, "Constraint Logic Programming Languages," in *Communications of the ACM* #33, July, 1990.

[11] J. McCarthy & P.J. Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence," in *Machine Intelligence*, D. Michie and B. Meltzer, ed. #4, Edinburgh University Press, Edinburgh, 1969.

[12] A. Tarski and S. Givant, "A Formalization of Set Theory Without Variables," in *Colloquium Publications, American Mathematical Society* #41, Providence, RI, 1987.

"An Object-Oriented Approach to Large-Scale Battlefield Simulation"

Michael Brewer and Pat Burns
Department of Mechanical Engineering
Colorado State University
Fort Collins, CO 80523
mbrewer@carbon.lance.ColoState.EDU
pburns@yuma.ACNS.ColoState.EDU
9th Army Conference
on
Applied Mathematics and Computing
Wednesday, 19 June 1991

ABSTRACT

Large scale computerized battlefield simulations have been in existence for a long period of time. CEM VI (Concepts Evaluation Model VI), upon which we have directed our effort, was first developed in 1968. Since then, it has evolved through several different authors and types of Fortran implementations. The last critical update occurred in 1983 with the introduction of ATtrition using CALibrated parameters (ATCAL) algorithm.

CEM VI is a discrete event simulation. As such, it is subject to random and *a priori* unknown branching. Thus, data are not contiguous in memory, and the data structure evolves with the simulation. The algorithm, as formulated, was unable not amenable to vectorization on the new Cray architectures. A typical CEM VI simulation, executed in the scalar CPU, typically consumes several to 10 hours of Cray 2 CPU time. To ameliorate this situation, we developed a strategy whereby the kernel of CEM VI (ATCAL) could be vectorized.

After careful investigation it was determined that data motion was the key in realizing the potential for vectorizing the ATCAL algorithm. Three different strategies were investigated, with execution rates determined for each method. Taking advantage of the Cray gather/scatter hardware was determined the most feasible of the strategies investigated. After implementing the strategy in ATCAL, a speedup of 8.09 was obtained.

With the implementation of the vectorized ATCAL algorithm into the CEM VI code, we expect considerable (up to a factor of 2) improvements in overall CPU run times.With increased performance, this will enable the Army to run more cases, and the cases each can be of greater fidelity (higher resolution). This may be particularly germane now that the development of a stochastic version of CEM VI is underway.

Theme of the Work

"I shall be accused, I suppose, of saying that no event in war can ever occur which may not be foreseen and provided for. To prove the falsity of this accusation, it is sufficient for me to cite the surprises of Cremona, Bergop-zoom, and Hochkirch. I am still of the opinion, however, that such events even as these might always have been anticipated, entirely or in part, at least within the limits of probability or possibility."

Baron de Jomini, General and Aid-de-Camp of the Emperor of Russia, *The Art of War*, 1862 (trans. by Capt. G. H. Mendell and Lieut. W. P. Craighill).

1. Description of CEM

1.1 History

The Concepts Evaluation Model (CEM VI) originated in 1968 as the Theater Combat Force Requirements Model (TCM) developed by Research Analysis Corporation as part of the FORWORN research program. TCM was designed to provide theatre level combat capabilities and requirements that would be sensitive to the mixes of units for both sides. After becoming operational, TCM was modified to include force evaluation and to satisfy needs for the army project Conceptual Design for the Army in the Field(CONAF). TCM then became known as CONAF Evaluation Model I (CEM I). During the next six years the model was modified several times improving methodology and applications in alternative theatre combat forces. In 1974 the project was turned over to the Army and renamed Concepts Evaluation Model IV, retaining the acronym CEM IV. With the advent of a radically different theater defense concept for Europe, CEM IV was improved once more and renamed CEM V, which was studied by the US Army Concepts Analysis Agency (CAA) from 1979 to 1983. In 1983 CEM VI evolved from CEM V with the onset of a new method for calculating combat attrition; this was the introduction of ATCAL (An Attrition Model Using Calibrated Parameters).

1.2 Discrete Event Simulation

The structure of the solution for the CEM VI model evolves with the simulation in a complex fashion, dependent upon input at the beginning of the simulation. There are multiple branching levels, each containing multiple constraints. The simulation is deterministic in that, with the same input file used to start the simulation, the same

results will be obtained. This type of simulation does however have branching *a priori* unknown, in the sense that the evolutionary structure of the problem can be different from simulation to simulation based on the difference in input, i.e. the structure of the simulation is input driven. A typical simulation over multiple time steps can consume few to ten hours of Cray 2 time. The evolutionary structure of the of the problem is depicted in Figure 1.

1.3 Battlefield Schematic

The battlefield for the CEM VI simulation is broken down into two distinct forces, one side containing a blue brigade the other a red division, the two sides being split by the Forward Edge of Battle Area (FEBA). The engagements are performed over diverse terrain - the smallest level of which is a sub-sector. The terrain is broken down into even smaller units called mini-sectors, as can be seen in figure 2.

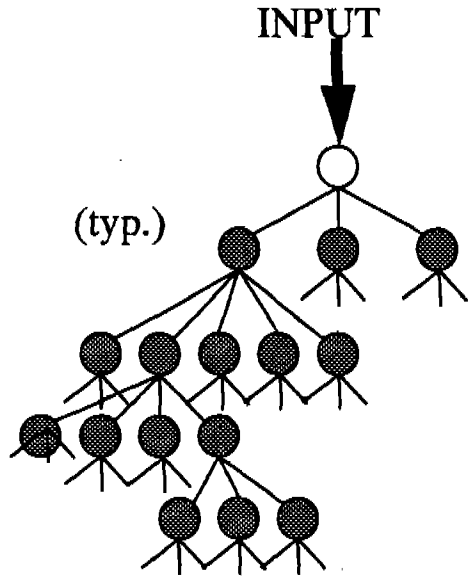


Figure 1 Evolutionary Structure of Problem

1.4 CEM VI - Scope and Fidelity

CEM VI as it relates to other battlefield simulations can best be shown in the following Figures 3 and 4. CEM VI encompasses large areas or theater level battles containing complete armies, whereas other simulations range from smaller areas where individual weapons are considered to groups and divisions. CEM VI is a low resolution simulation in which a kill matrix is used to encompass entire groups of weapons and targets, while high resolution simulations deal principally with the physics of individual weapon types.

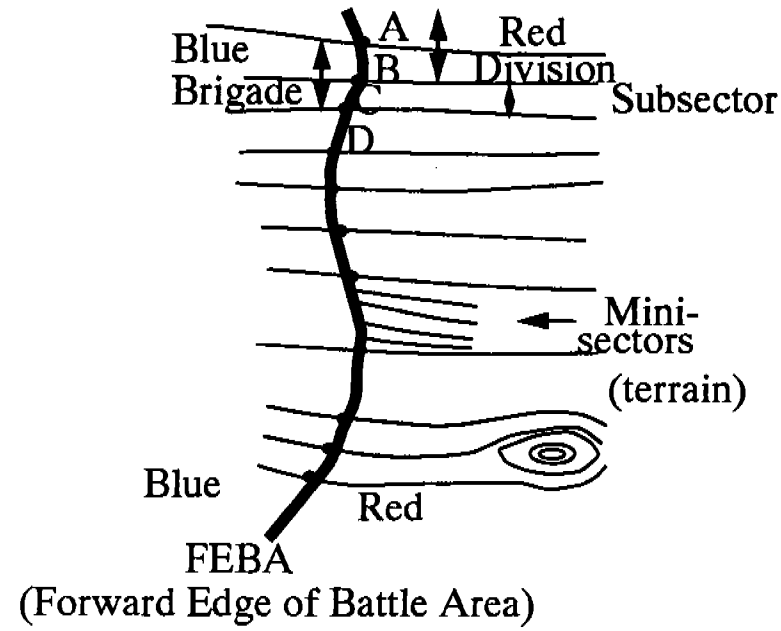


Figure 2 Terrain Distinction Controlling Engagements

2. Issues Involving CEM VI

There are two distinct issues involved with the CEM VI simulation: performance and composition of forces (as they evolve with the simulation). To wit, is the domain of the kill matrix used in the evolution of the problem that of the lower resolution simulation?

The main issues involved with the performance of the CEM VI simulation can be broken down into five distinct categories: data motion,vectorization,parallelization, input/output and debugging.

The data motion consists of the movement of killer victim scoreboards which are set up as arrays and used to calibrate attrition rates. These arrays are dependent upon the initial input into the simulation, and thus can cause different outcomes based on initial conditions. Because of their size, these arrays constitute large amounts of data motion and consume significant CPU time in the process.

With the advent of vector machines, vectorization of the code plays a key role in performance enhancement. For our purposes here, vectorization and parallelization are basically similar.

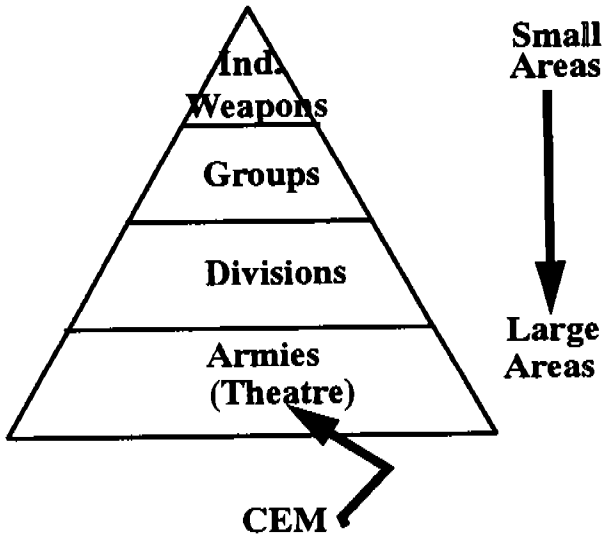


Figure 3.

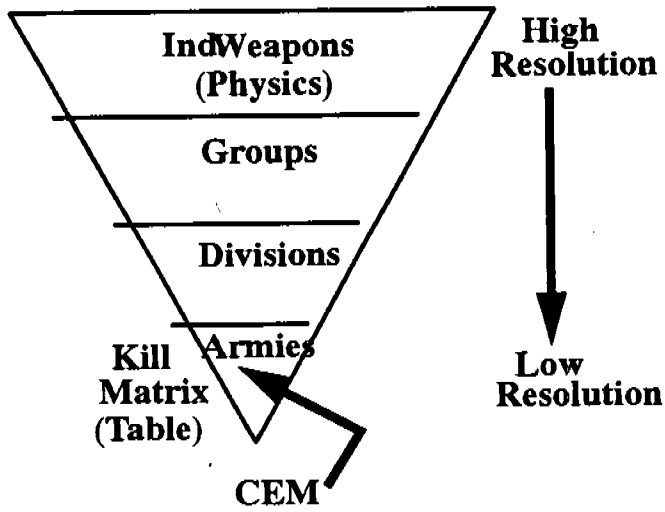


Figure 4.

CEM VI consumes significant i/o resources, required to set up correct killer victim scoreboards. Additionally, it is desirable to be able to interpret intermediate results. Therefore at this point, it is not advisable to consider changing this portion of CEM VI. Debugging plays an important role in performance monitoring as it does little good to increase the performance of the program if the results are incorrect.

3. Our Approach to Improvements in Performance

3.1 Vectorization of the Kernel

The kernel of CEM VI consists of the ATCAL algorithm which performs engagements over subsectors. The modification of ATCAL to enhance vectorization yields the best return in that it is the least invasive approach (fewest code modifications) and initially ATCAL, consumed 65% of the CPU time (from profiling) apart from input/output.

3.2 Development of Visualization Tools

With the development of visualization tools, a more accurate interpretation of the results can be made and performance can be monitored more easily.

3.3 Results of Improved Performance

With improved performance CPU time will be lowered, thus allowing for more runs. With more runs greater optimization of force mixes can be improved. Also, with less CPU time being consumed, it will allow an increase in the resolution of the simulation, resulting in greater physical fidelity.

4. Data Motion as the Key Problem

Through intensive study of the ATCAL algorithm, it was determined that data motion was the main cause for CPU time consumption in the kernel. With *a priori* unknown branching taking place involving ammunition constraints, weapon constraints, firepower constraints and target constraints, this presented a challenging problem in algorithm design.

4.1 In-Depth Study of Data Motion

Here, we study three approaches to perform the data motion to capture the dynamic structure of the solution. Amdahl's law illustrates the potential payoff, but yields no information as to the vector overhead involved, or as to the best strategy to employ. We examine in detail strategies which allow these factors to be determined.

4.2 Modified Amdahl's Law

The payoff for the added complexity of structuring the algorithm to perform data motion is well illustrated through a modification to Amdahl's Law [Amdahl, 1967] as follows.

Let f_p represent the fraction of time spent by the executing code in the vector hardware, D the ratio of time spent in the vector hardware performing purely data motion to that spent doing useful work, F the fractional inefficiency in vector calculations due to overhead (including start-up), V/S the ratio of scalar to vector execution rates, and R the vector (with data motion) to scalar speedup ratio. Then, accounting for all work to be done including scalar work, and vector work (real vector work and "useless" vector work, i.e., due to data motion), and as an approximation, neglecting overlap, then:

$$R = \frac{\Delta t_s}{\Delta t_p} = \frac{1}{(1 - f_p) + f_p \frac{V}{S}(1 + D + F)} \qquad (1)$$

The quantities D and F may be considered in combination, i.e. in the following D represents the sum of (D + F). The speedup ratio, R, calculated from equation (1) is plotted in Figure 5, with p varying along the abscissa, and D varying parametrically. The figure is constructed for V/S = 12, approximately representative of Cray hardware. Two facts are apparent from the figure: (1) good speedup may still be obtained for large amounts of data motion, and (2) the code must be highly vectorized to achieve close to maximum speedup with data motion. Both facts result from the much higher execution rate of vector hardware, when compared to scalar hardware.

4.3 Strategy I - Gather/Scatter

This involves utilizing the scatter/gather hardware in the Cray to access data elements non-contiguous in memory, according to a vector of indices. For example, suppose the vector of indices for those elements which pass a specific test are denoted INDEX(I). Specifically, the following Fortran pseudo-code effects the generation of such:

```
K = 0
DO I = 1, LENGTH
    IF (TRUTH(I)) THEN
       K = K + 1
       INDEX(K) = I
    END IF
 END DO
```

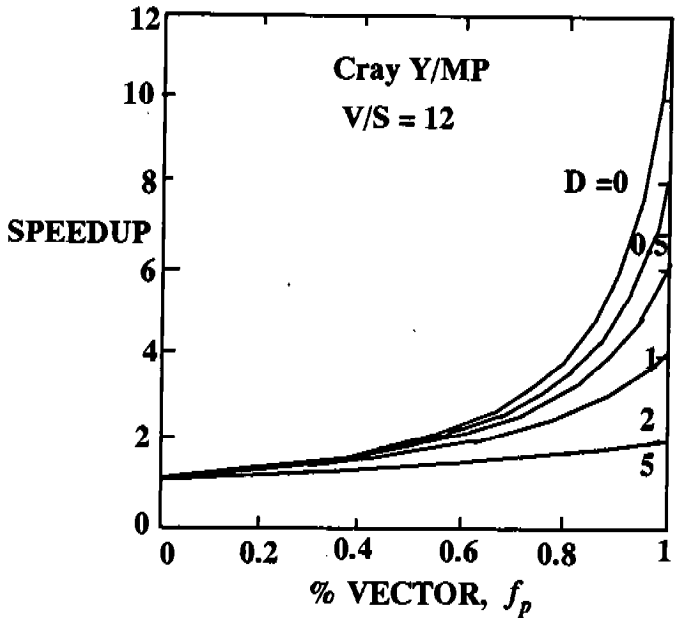


Figure 5 Amdahl's Law Modified for Data Motion

The above loop does not vectorize in Fortran (a deficiency of CFT77). To generate the above list in the vector hardware, it is necessary to use the Cray SCILIB routines such as WHENFLT, WHENFGT, etc. In Figure 6(a), we show rates at which vectors of such indices can be generated versus input vector length with the truth ratio (fraction of elements which pass the test) as a parameter. Here, we plot results in MOPS (Millions of OPerations per CPU Second), where we define one operation as the generation of one index.

Next, data elements must be gathered. The gather is done in a fashion which preserves the original order of elements. As such. we term this a vector "compress." Typically, the execution rate of a vector gather on a Cray is independent of the stride (increment in memory between elements). However, for a compress, memory bank conflicts arise due to the preservation of order. This makes the execution rate dependent upon truth ratio (or density) as shown in Figure 6(b).

Note that the peak rates are fairly high, but that long input vector lengths are required to achieve near peak performance. This makes the Cray function more like a long-vector architecture (such as the Cyber 205 used to be). Thus, it behooves us to structure the algorithm so as to employ long vectors (i.e., in the case of CEM VI, perform engagements over multiple, possibly many, sub-sectors).

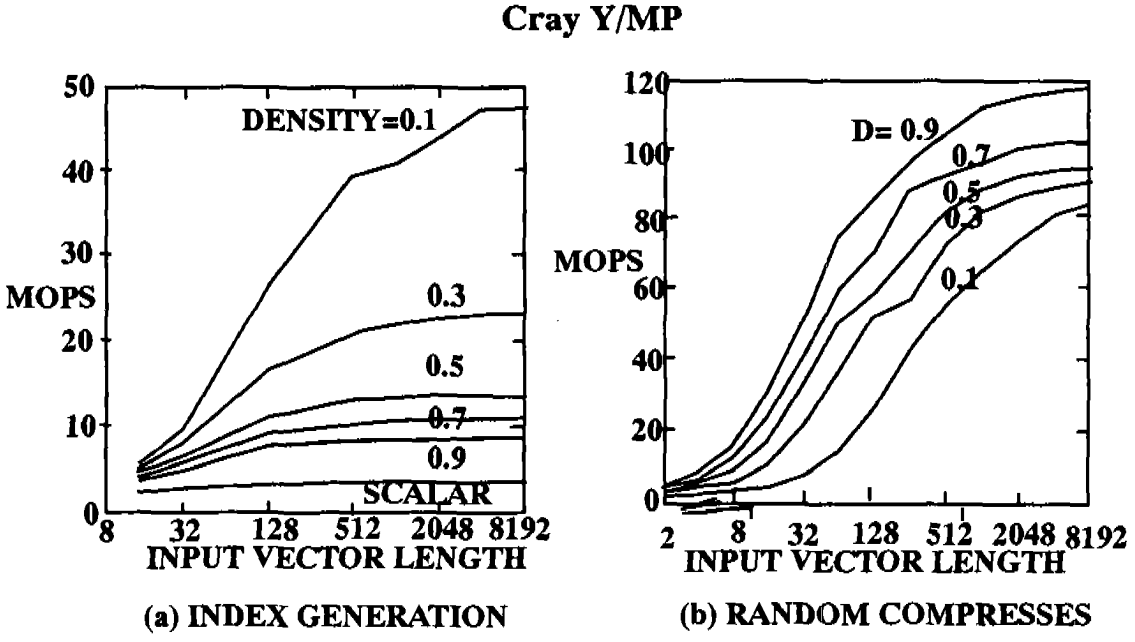


Figure 6 Execution Rates for Strategy I

4.4 Strategy II - IF THEN/ELSE Structures

Another strategy, allowing the overhead of index generation to be by-passed, is illustrated in the following loop:

```
DO I = 1, LENGTH
    IF (TRUTH1(I)) THEN
       execute statement 1
    ELSE IF (TRUTH2(I)) THEN
     execute statement 2
    ELSE
       execute statement N
    END IF
 END DO
```

Execution rates for this strategy are shown in Figure 7. Here, although the operation proceeds at vector speed, the execution rates are low. The highest execution rates are for the case of N=2 (IF THEN/ELSE) structures. Fortunately, most decisions in discrete event simulations are binary. Even where there are cases that do not, the logic may frequently be reduced to binary decision trees.

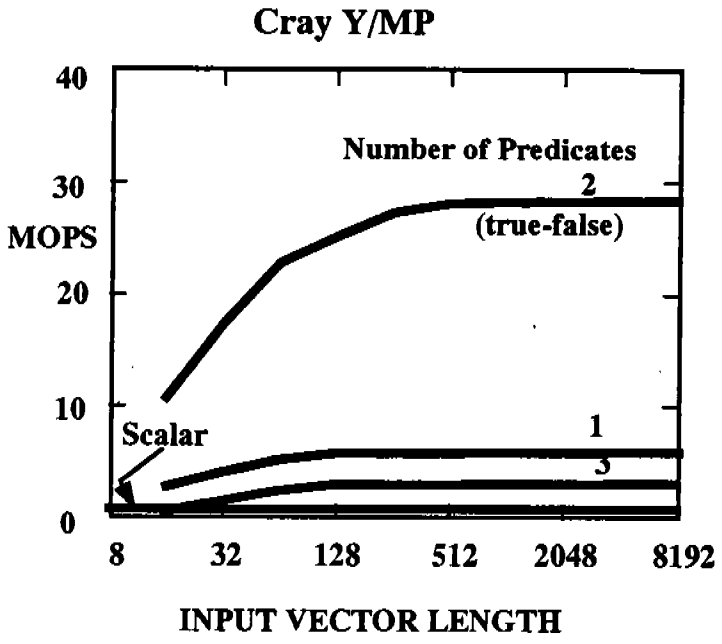


Figure 7 Execution Rates for Strategy II

4.5 Strategy III - Logical Truth Vectors

The third strategy is a spin-off of Strategy II, in that the predicates (conditionals) are stored as logical vectors rather than being evaluated (and lost) at the instant of run time. The advantage this approach offers over strategy II lies in the ability to perform successive levels of branching. The key to this strategy lies in the ability of the Cray to perform Boolean (logical) operations such as AND's and OR's. In Figure 8, we depict the execution rates for Cray architectures for logical operations. These operations must be done in series with those of Strategy II

5. Results and Discussion

In our case, since we have only a few levels of branching, and since most of the data motion in CEM VI occurs at only 1 level of indirection, we choose Strategy I as potentially the most effective strategy. We proceed to describe in greater detail the data motion in CEM VI for a single engagement.

In our case we loop over 51 vehicle types for the red side as targets with the blue side as shooters, then over 51 vehicle types for the blue side as the shooter and the red side as the target. For each of these target types, a bias array is used to determine whether the shooter vehicle employs direct fire or indirect. This constitutes a significant amount of data motion. In addition to using Strategy I, we have in-lined the subroutine which computes direct fire kills using the kill matrix.

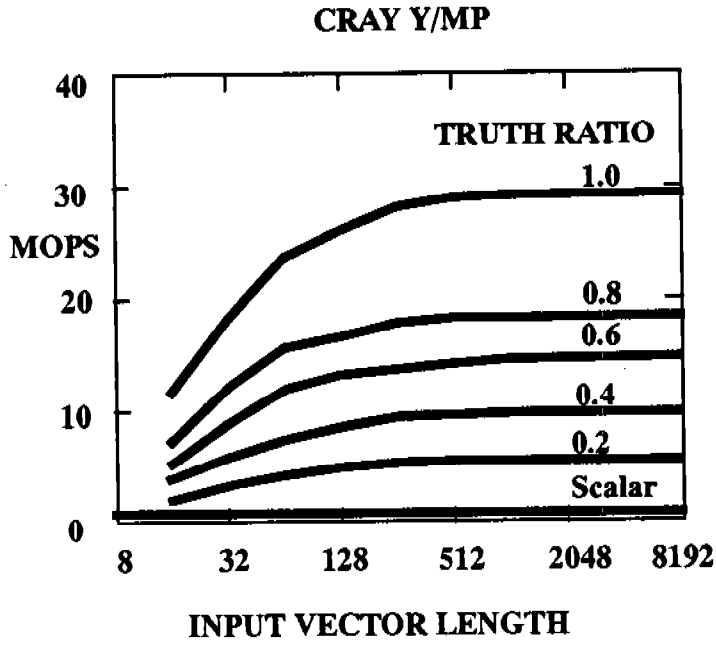


Figure 8 Execution Rates for Logical Operations

We instrumented the code to determine that 90% of the time is spent in performing direct fire. At a truth ratio of 90%, Figure 6(a) indicates a potential speedup of about 4 when generating indices, and Figure 6(b) indicates execution rates of about 50 MOPs. We have measured a total speedup of 8.09 after implementing these techniques. Figure 5 indicates that we are in the domain of from 0 to 50% penalty for vectorized data motion (D), and of from about 92% to 100% of the code being vectorized (fp).

6. Conclusions

During data motion, Cray architectures perform like long-vector machines such as a Cyber 205. Even with the penalty of data motion, high execution rates are possible as can be shown by the fact that a speedup of more than 8 was attained after vectorizing ATCAL. A video showing the improved performance and the state of the simulation was also developed in conjunction with the vectorization of ATCAL.

7. Recommendations

With the implementation of the vectorized ATCAL into CEM VI, actual speedups in the CEM VI code from the enhanced ATCAL should be determined. For future considerations in the improvement of CEM VI in both structure and performance, the exploration of alternative vectorization strategies should be studied. Some possible

directions might be vectorizing over some sub-sectors or possibly over all sub-sectors depending on memory constraints. A visualization package should be developed to show the additional capabilities of the enhanced code, and too improve the user interface between terminals and software. A visualization package could also assist in interpreting the process, especially if implemented with more error trapping. Other possible equation solution strategies should also be studied; now the code implements direct substitution for convergence of the main attrition loop. Other possible candidate solution techniques might be Newton iterations or Broyden updates.

8. REFERENCES

Amdahl, G. A.; 1967. "Validity of the Single-Processor Approach to Achieving Large-Scale Computing Capability," *Proceedings American Federation of Information Processing Societies, 30,* Washington, DC, pps. 483-485

9. ACKNOWLEDGEMENT

We gratefully acknowledge the support of the Army in this effort via ARO Grant No. - DAAL03-90-G-0200

EVOLVING PHASE BOUNDARIES IN DEFORMABLE CONTINUA

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

ABSTRACT. Recently, Gurtin and Struthers [2] developed a dynamical theory of phase transitions in crystal-crystal systems in which the interface is sharp, coherent, and endowed with energy, entropy, and superficial force. A fundamental conceptual ingredient of the theory is the use of three force systems: *deformational forces* that act in response to the motion of material points; *accretive forces* that act within the crystal lattice to drive the crystallization process; *attachment forces* associated with the attachment and release of atoms as they are exchanged between phases. Here I will discuss the main results of the theory, which are constitutive equations and balance laws for the interface.

CONSTITUTIVE THEORY. The surface energy and the accretive and deformational surface stresses are allowed to depend on the bulk deformation gradient $\mathbf{F}$, the normal $\mathbf{n}$ to the interface, the normal speed v of the interface, and a list z of subsidiary variables of lesser importance. It follows, as a consequence of thermodynamic admissibility, that: the surface energy and the accretive and deformational surface stresses are independent of v and z, and depend on $\mathbf{F}$ at most through the tangential deformation gradient $\boldsymbol{F}$; in fact, the energy

$$\psi = \widehat{\psi}(\boldsymbol{F}, \mathbf{n}) \tag{1}$$

completely determines the surface stresses through relations, the two most important of which are:

$$\boldsymbol{S} = \partial_{\boldsymbol{F}} \widehat{\psi}(\boldsymbol{F}, \mathbf{n}), \qquad \boldsymbol{c} = -D_{\mathbf{n}} \widehat{\psi}(\boldsymbol{F}, \mathbf{n}), \tag{2}$$

in which $\boldsymbol{S}$ is the deformational (Piola-Kirchhoff) surface stress, $\boldsymbol{c}$ is the normal accretive stress, $\partial_{\boldsymbol{F}}$ is the partial derivative with respect to $\boldsymbol{F}$, and $D_{\mathbf{n}}$ is the derivative with respect to $\mathbf{n}$ following the interface. A further consequence of thermodynamics is an explicit expression for the normal attachment force π:

$$\pi = k + \Psi + bv, \qquad b = \widehat{b}(\boldsymbol{F}, \mathbf{n}, v, \mathbf{z}) \geq 0, \tag{3}$$

where Ψ is the difference in bulk energies, while k is related to changes in momentum and kinetic energy across the interface. These results imply that the sole source of dissipation is the exchange of atoms between phases, with bv^2 the dissipation per unit interfacial area.

INTERFACE CONDITIONS. The system of constitutive equations and balance laws combine to give the interface conditions[1]

$$
\begin{aligned}
&div_{\mathcal{S}}\boldsymbol{S} + (\mathbf{S}_2 - \mathbf{S}_1)\mathbf{n} = \rho v(\mathbf{v}_1 - \mathbf{v}_2)\,, \\
&\Psi_1 - \Psi_2 = (\mathbf{S}_1\mathbf{n})\cdot(\mathbf{F}_1\mathbf{n}) - (\mathbf{S}_2\mathbf{n})\cdot(\mathbf{F}_2\mathbf{n}) - k - g - bv\,,
\end{aligned}
\tag{4}
$$

with

$$
\begin{aligned}
k &= \frac{1}{2}\rho v^2\{|\mathbf{F}_1\mathbf{n}|^2 - |\mathbf{F}_2\mathbf{n}|^2\} \\
g &= -\psi\kappa - div_{\mathcal{S}}\boldsymbol{c} + (\mathbf{F}^T\boldsymbol{S})\cdot\boldsymbol{L}\,.
\end{aligned}
\tag{5}
$$

The subscripts 1 and 2 denote the two phases: Ψ_1 and Ψ_2 are the bulk energies per unit reference volume; $\mathbf{S}_1$ and $\mathbf{S}_2$ are the bulk Piola-Kirchhoff stresses; $\mathbf{F}_1$ and $\mathbf{F}_2$ are the bulk deformation gradients; $\mathbf{v}_1$ and $\mathbf{v}_2$ are the material velocities; ρ is the reference density. The remaining quantities concern the interface: $\boldsymbol{L}$ is the curvature tensor with κ, its trace, the total curvature; $div_{\mathcal{S}}$ is the surface divergence.

SIMPLIFIED EQUATIONS.[2] Assume that both phases are isotropic with *linearized* stress-strain relations in each phase, and neglect all interfacial terms with the exception of the dissipative term bv in (4). Then for ***longitudinal motions*** with scalar displacement $u(x,t)$ and scalar tensile stress $\sigma(x,t)$ the basic equations are[3] the bulk equations

$$
(\textit{phase } 1) \qquad c_1^2 u_{xx} = u_{tt}, \quad \sigma = \beta_1 u_x, \quad \psi = \frac{1}{2}\beta_1 u_x^2
$$

$$
(\textit{phase } 2) \qquad c_2^2 u_{xx} = u_{tt}, \quad \sigma = \sigma_0 + \beta_2 u_x, \quad \psi = \psi_0 + \sigma_0 u_x + \frac{1}{2}\beta_2 u_x^2
$$

and the interface conditions

$$
\begin{aligned}
[\sigma] = -\rho v[u_t], \qquad [u_t] = -v[u_x]\,, \\
[\psi] = \langle\sigma\rangle[u_x] + bv\,,
\end{aligned}
$$

where $c_i^2 = \beta_i/\rho$ with β_i the elastic moduli; σ_0 and ψ_0 are constants; [ ] denotes the jump across the interface; $\langle\rangle$ designates the average interfacial value.

---

[1] For statical situations: $(4)_1$ was derived by Gurtin and Murdoch [6] as a consequence of balance of forces; $(4)_2$ and its counterpart for crystal-melt interactions were derived by Leo and Sekerka [5] (cf. Johnson and Alexander [3,4]) as Euler-Lagrange equations for stable equilibria. In the absence of surface stress and surface energy ($\boldsymbol{S} = 0, \boldsymbol{c} = 0, \psi = 0$): $(4)_1$ is a standard shock relation; $(4)_2$ (with $b \neq 0$) was established by Abeyaratne and Knowles [7] and Truskinovsky [11]. Counterparts of (4) for a rigid crystal in an inviscid melt were derived in [8]; an analog of $(4)_2$ for a rigid system was given in [1].

[2] Cf. [9]

[3] Cf. Abeyaratne and Knowles [10], whose treatment is slightly different.

For *antiplane shear* with scalar displacement $u(x,y,t)$ and shear-stress vector $\mathbf{T}(x,y,t)$ the basic equations are the bulk equations

$$(phase\ 1) \qquad s_1^2\Delta u = u_{tt},\ \mathbf{T} = \mu_1\nabla u, \qquad \psi = \frac{1}{2}\mu_1|\nabla u|^2$$

$$(phase\ 2) \qquad s_2^2\Delta u = u_{tt}, \quad \mathbf{T} = \mathbf{T}_0 + \mu_1\nabla u, \quad \psi = \psi_0 + \mathbf{T}_0\cdot\nabla u + \frac{1}{2}\mu_2|\nabla u|^2$$

and the interface conditions

$$[\mathbf{T}]\cdot\mathbf{n} = \rho v^2[\nabla u]\cdot\mathbf{n}, \qquad [u_t] = -v[\nabla u]\cdot\mathbf{n},$$
$$[\psi] = \langle\mathbf{T}\rangle\cdot\mathbf{n}([\nabla u]\cdot\mathbf{n}) + bv,$$

where Δ is the laplacian; $s_i^2 = \mu_i/\rho$ with μ_i the shear moduli; $\mathbf{T}_0$ and ψ_0 are constants.

Acknowledgment. The research discussed here was supported by the Army Research Office and the National Science Foundation.

REFERENCES

[1] Gurtin, M. E., Multiphase thermomechanics with interfacial structure. 1. Heat conduction and the capillary balance law, Arch. Rational Mech. Anal., **104**, 185–221 (1988).

[2] Gurtin, M. E. and A. Struthers, Multiphase thermomechanics with interfacial structure. 3. Evolving phase boundaries in the presence of bulk deformation, Arch. Rational Mech. Anal., **112**, 97–160 (1990).

[3] Alexander, J. I. D. and W. C. Johnson, Thermomechanical equilibrium solid-fluid systems with curved interfaces, J. Appl. Phys. **58**, 816–824 (1985).

[4] Johnson, W. C. and J. I. D. Alexander, Interfacial conditions for thermomechanical equilibrium in two-phase crystals, J. Appl. Phys. **59**, 2735–2746 (1986).

[5] Leo, P. H. and R. F. Sekerka, The effect of surface stress on crystal-melt and crystal-crystal equilibrium, Forthcoming.

[6] Gurtin, M. E. and I. Murdoch, A continuum theory of elastic material surfaces, Arch. Rational Mech. Anal., **57**, 291–323 (1975).

[7] Abeyaratne, R. and J. K. Knowles, On the driving traction acting on a surface of strain discontinuity in a continuum. J. Mech. Phys. Solids, **38**, 345–360 (1990).

[8] Gurtin, M. E., A mechanical theory for crystallization of a rigid solid in a liquid melt; melting-freezing waves, Arch. Rational Mech. Anal., **110**, 287–312 (1990).

[9] Gurtin, M. E., Simple equations for dynamic phase transitions, Forthcoming.

[10] Abeyaratne, R. and J. K. Knowles, Wave propagation in linear, bilinear, and trilinear elastic bars, Forthcoming.

[11] Truskinovsky, L., Kinks versus shocks, Shock Induced Transitions and Phase Structures in General Media (ed. R. Fosdick, E. Dunn and M. Slemrod) Springer-Verlag (1991).

A Central Limit Theorem for Extreme Sojourns of Diffusion Processes

Simeon M. Berman

Courant Institute of Mathematical Sciences
New York University

Let $X(t), t \geq 0$, be a diffusion process on the real line; and, for $u, t > 0$, let $L_t(u)$ be the sojourn time of $X(s), 0 \leq s \leq t$, above the level u, that is, the measure of the set $\{s: 0 \leq s \leq t, X(s) > u\}$. The main result is a central limit theorem for the random variable $L_t(u)$, for $t \to \infty$ and a class of functions $u = u(t) \to \infty$. The conditions in the hypothesis of the theorem are stated in terms of the coefficient functions in the infinitesmal generator of the process, namely, the coefficients of diffusion and drift, denoted as $a(x)$ and $b(x)$, respectively. The conditions that are employed imply, in particular, that there is a stationary probability distribution for this process. In the case of a constant level u, the validity of the central limit theorem was established long ago (Maruyama and Tanaka, 1957). More recently the author considered the case $u(t) \to \infty$. Let $S(x)$ be the scale function of the process, defined as

$$S(x) = \int_0^x exp(-2\int_0^y [b(z)/a(z)])dz.$$

In the case where $u(t)$ grows with t in such a way that $S(u(t)) \sim t$, for $t \to \infty$, it was shown (Berman, 1983,1988) that the random variable $[2b^2(u)/a(u)]L_t(u)$ has a limiting infinitely divisible distribution of a specific form. The present work deals with the case falling between the situations $S(u(t)) \sim t$ and $u(t) \equiv$ constant, namely, $S(u(t))/t \to 0$ for $t \to \infty$. It is shown, in this case, that $(L_t(u) - EL_t(u))/(VarL_t(u))^{1/2}$ has a limiting standard normal distribution, for any starting point of the process in the state space. Asymptotic forms for the normalizing functions $EL_t(u)$ and $(VarL_t(u))^{1/2}$ are derived in terms of the coefficient functions $a(x)$ and $b(x)$.

Here is the precise statement of the main result:

THEOREM: Let $X(t), t \geq 0$, be a diffusion process defined by the stochastic integral equation $X(t) - X(t') = \int_{t'}^{t} b(X(s))ds + \int_{t'}^{t} a^{1/2}(X(s))W(ds)$, $0 < t' < t$, where $W(s)$ is the adapted Brownian motion.

Research sponsored by the U.S. Army Research Office

The coefficients $a(x)$ and $b(x)$ are assumed to satisfy the following conditions:

$a(x)$ and $-b(x)$ are regularly oscillating for $x \to \infty$. (*For* the definition, see Berman (1982).) (1)

$$\lim_{x \to \infty} -xb(x)/a(x) = \infty. \tag{2}$$

Put

$$v(x) = 2b^2(x)/a(x), \tag{3}$$

and assume

$$\liminf_{u \to \infty} \frac{\inf(v(x): x \geq u)}{v(u)} > 0. \tag{4}$$

Let $m(x)$ be the density of the speed measure: $m(x) = (a(x)S'(x))^{-1}$; and assume that $\int_{-\infty}^{\infty} m(x)dx < \infty$. Let $u(t)$ be an increasing function such that

$$\lim_{t \to \infty} \frac{S(u(t))}{t} = 0 \tag{5}$$

and

$$\lim_{\delta \to 0} \limsup_{t \to \infty} \frac{\delta S(u(t))}{S(u(t\delta))} = 0. \tag{6}$$

Then there are explicitly constructed functions $u(t)$ and $\sigma(t)$, expressed in terms of the coefficients functions $a(x)$ and $b(x)$, such that $(L(t) - u(t))/\sigma(t)$ has a limiting standard normal distribution for $t \to \infty$, for any initial point of the process.

If $EL_t(u(t))$ and $Var\ L_t(u(t))$ are the mean and variance under the stationary distribution (which exists because the speed measure is finite) then a weak compactness argument shows that $(L(t) - EL_t(u(t)))/(Var\ L_t(u(t)))^{1/2}$ also has a standard normal limit.

The proof will be given in a complete paper to be published elsewhere.

REFERENCES

S.M. Berman (1982) Sojourns and extremes of a diffusion process on a fixed interval. *Adv. Appl. Probability* 14 811-832.

S.M. Berman (1983) High level sojourns of a diffusion process on a long interval. Z. Wahrscheinlichkeitztheorie verw. Gebiete 62 185-199.

S.M. Berman (1988) Extreme sojourns of diffusion processes. *Ann. Probability* 16 361-374.

G. Maruyama and H. Tanaka (1957) Some properties of one-dimensional diffusion processes. *Mem. Fac. Sci. Kyushu Univ. Ser. A* 11 117-141.

3-D SHAPE FROM A SHADED AND TEXTURAL SURFACE IMAGE †

Yoonsik Choe and R. L. Kashyap
School of Electrical Engineering, Purdue University
West Lafayette, Indiana 47907-1285

ABSTRACT. To recover 3-D structure from a natural scene image involving textures, neither the Shape-from-shading nor the Shape-from-texture analysis is enough, because both radiance and texture information coexist within the surface of a natural scene. A new 3-D texture model is developed by considering the scene image as the superposition of a random texture image and a smooth shaded image. The whole image is analyzed using a patch-by-patch process. Each patch is assumed as a tilted and slanted texture plane. A modified reflectance map function is applied to describe the deterministic part, and the Fractional Differencing Periodic model is chosen to describe the random texture, because of its good performance in texture synthesis and its ability to represent the coarseness and the pattern of the surface at the same time. An orthographical projection technique is developed to deal with this particular random model, which has a non-isotropically distributed texture pattern. For estimating the parameter, a hybrid method which uses both the least square and the maximum likelihood estimates is applied directly to the given intensity function. By using these parameters, the synthesized image is obtained and used to reconstruct the original image. The contribution of this research will be in combining shape-from-shading and Shape-from-texture techniques to extract 3-D structure and texture pattern features from a single natural scene image which contains both shade and texture in it.

INTRODUCTION. An important task in computer vision is the recovery of 3-D scene information from single 2-D images. 3-D analysis of an image can be broken down into two main categories, Shape-from-shading and Shape-from-texture. In Shape-from-shading technique, scene radiance information plays an important role to extract 3-D surface information from image data [6,15, 20]. On the other hand, in Shape-from-texture technique, the texture pattern instead of shading is used to extract 3-D structure. Since texture gradients behave like intensity gradients, the shape of a surface can be inferred from the pattern of a texture on the surface by applying statistical texture analysis [14,22,23].

However, for describing a natural scene image, both the above approaches have their own limitations. The Shape-from-shading technique is applicable only under the assumption that the surface is smooth and has constant albedo, while the Shape-from-texture technique requires the surface to be relatively complex so that texture information can be extracted. Thus, neither technique is suitable to recover 3-D structure information from a natural scene, because both radiance and texture information coexist within the surface of a natural scene. Therefore, a robust technique is needed to handle this shortcoming. Recently, the fractal scaling parameter was introduced to measure the coarseness of the surface, and applied to represent the natural scene surface [21]. However, this fractal model is not enough to represent the real 3-D texture image, because even though two surfaces are estimated to have the same fractal scales, these surfaces can have different texture patterns.

†This research is partially supported by the U.S. Army Research Office under contract DAAL03-89K-0032.

In this paper, a composite model of Shape-from-shading and Shape-from-texture is developed to represent a 3-D surface image considering the scene image as the superposition of a smooth shaded image and a random texture image, that is, the deterministic function $x(l_1,l_2)$ and the random function $y(l_1,l_2)$. Then, the orthographical projection is adapted to take care of the non-isotropic distribution function due to the slant and tilt of a 3-D texture surface. The Fractional Differencing Periodic (FDP) model given below is chosen to represent the random texture.

$$y(l_1,l_2) = (1 - 2\cos\omega_1 z_1^{-1} + z_1^{-2})^{-\frac{c}{2}} \cdot (1 - 2\cos\omega_2 z_2^{-1} + z_2^{-2})^{-\frac{d}{2}} \zeta(l_1,l_2)$$

Here z_i is the delay operator, corresponding to l_i, and $\zeta(\cdot,\cdot)$ is a white noise sequence. The advantage of the (FDP) model is that it can simultaneously represent the coarseness and the pattern of the 3-D texture surface with the fractional differencing parameters c, d and the frequency parameters ω_1, ω_2, respectively, and it has the property of being flexible enough to synthesize both long-term and short-term correlation structures of random texture depending on the values of the fractional differencing parameter c and d. (More detailed discussion on FDP model will be given in chapter 2.3.) Since the object is described by a model involving several free parameters and the values of these parameters are determined directly from its projected image, it is possible to extract 3-D information and texture pattern directly from the given intensity values of the image without any pre-processing. Thus, the cumulative error obtained from several pre-processing stages can be minimized. For estimating the parameters, a hybrid method which uses both the least squares and the maximum likelihood estimates is applied and the estimation and the synthesis are done in frequency domain based on the local patch analysis. By using this model, the integrability problem which might occur in spatial domain analysis can be avoided, because only one inverse Fourier transform needs to be taken at the end of procedure to get the whole image.

The organization of this paper is as follows. In Section 2 we introduce the image model $i(l_1,l_2)$ which is obtained by superposing the deterministic function $x(l_1,l_2)$ and the random function $y(l_1,l_2)$, and the relationship between different directions of 3-D surface. Section 2.1 gives a scheme for estimating the illumination direction. The modified reflectance map function $x(l_1,l_2)$, and the orthographically projected Fractional Differencing Periodic function $y(l_1,l_2)$ are introduced in sections 2.2-2.3. Section 3.1 outlines the estimation scheme for the parameters in the composite model. Section 3.2 discusses some simulation results carried out to demonstrate the performance of the proposed algorithm, followed by Section 4 which concludes the paper.

The detailed paper with the same title will appear in the IEEE Transactions on Pattern Analysis and Machine Intelligence, October 1991.

REFERENCES

[1] Box, G.E.P. and Jenkins, G.M., *Time Series Analysis: Forecasting and Control,* Holden-Day, 1969.

[2] Brillinger, D.R., *Time Series, Data Analysis and Theory, Expanded Edition,* Holden-Day, Inc., 1981.

[3] Eom, K.-B., "Robust Image Models with Application," *Ph.D. Thesis,* Purdue University, West Lafayette, IN, 1986.

[4] Ferrie, F.P. and Levine, M.D., "Where and Why Local Shading Analysis Works," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-11, No. 2, Feb. 1989, pp. 198-206.

[5] Frankot, R.T. and Chellappa, R., "A Method for Enforcing Integrability in Shape from Shading Algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 10, No. 4, July 1988, pp. 439-451.

[6] Horn, B.K.P. and Brooks, M.J. (ed.), *Shape from Shading,* MIT Press, 1989.

[7] Horn, B.K.P. and Brooks, M.J., "The Variational Approach to Shape from Shading," *Computer Vision, Graphics and Image Processing,* Vol. 33, 1986, pp. 174-188.

[8] Hosking, J.R.M., "Fractional Differencing," *Biometrika,* Vol. 68, 1981, pp. 165-176.

[9] Kanatani, K., "Detection of Surface Orientation and Motion from Texture by a Stereological Technique," *Artificial Intelligence,* Vol. 23, 1984, pp. 213-237.

[10] Kang, H. and Aggarwal, J.K., "Design of Two-Dimensional Recursive Filters by Interpolation," *IEEE Trans. Circuit and Systems,* Vol. CAS-24, 1977, pp. 281-291.

[11] Kashyap, R.L., "Image Models," in *Handbook of Pattern Recognition and Image Processing,* Academic Press, Inc., 1986, pp. 281-310.

[12] Kashyap, R.L. and Eom, K.-B., "Estimation in Long-Memory Time-Series Model," *Journal of Times Series Analysis,* Vol. 9, 1988, pp. 35-41.

[13] Kashyap, R.L. and Eom, K.-B., "Texture Boundary Detection Based on the Long Correlation Model," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-11, No. 1, Jan. 1989, pp. 58-67.

[14] Kender, J.R., "Shape from Texture: An Aggregation Transform that Maps a Class of Textures into Surface Orientation," *Proc. 6th IJCAI,* 1979, pp. 475-480.

[15] Lee, C.-H. and Rosenfeld, A., "Improved Methods of Estimating Shape from Shading Using the Light Source Coordinate System," *Artificial Intelligence,* Vol. 26, 1985, pp. 125-143.

[16] Mandelbrot, B.B. and Van Ness, J.W., "Fractional Brownian Motions, Fractional Noises and Applications," *SIAM Rev.,* Vol. 10, 1968, pp. 422-437.

[17] Pentland, A. and Kube, P., "On the Imaging of Fractal Surfaces," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-10, Sept. 1988, pp. 704-707.

[18] Pentland, A.P., "Finding the Illumination Direction," *J. Opt. Soc. Am.,* Vol. 72, April 1982, pp. 448-455.

[19] Pentland, A.P., "Fractal-Based Description of Natural Scenes," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-6, Nov. 1984, pp. 661-674.

[20] Pentland, A.P., "Local Shading Analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. PAMI-6, Mar. 1984, pp. 170-187.

[21] Pentland, A.P., "Shading Into Texture," *Artificial Intelligence,* Vol. 29, 1986, pp. 147-170.

[22] Stevens, K.A., "The Visual Interpretation of Surface Contours," *Artificial Intelligence,* Vol. 18, 1981, pp. 47-47.

[23] Witkin, A.P., "Recovering Surface Shape and Orientation from Texture," *Artificial Intelligence,* Vol. 17, 1981, pp. 17-45.

[24] Kashyap, R.L. and P.M. Lapsa, "Synthesis and Estimation of Random Fields Using Long-Correlation Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* Vol. 6, No. 6, 1984, pp. 800-808.

RECURRENCE RELATIONS, CONTINUED FRACTIONS AND TIME EVOLUTION IN MANY-PARTICLE SYSTEMS

M. Howard Lee
Department of Physics and Astronomy
University of Georgia
Athens, GA 30602, USA

ABSTRACT. The study of time and frequency dependent behavior in quantum many-particle systems represents one of the most significant developments in statistical physics in recent years. Fundamental approaches involve solving the Heisenberg equation of motion for a given dynamical variable and then evaluating an ensemble average at two different times. Most interesting and difficult regimes are long times and low frequencies where standard perturbative techniques become inapplicable. Recent advances have shown that recurrence relations and continued fractions provide sounder approaches to solving these problems. Progress made at the University of Georgia, supported by the ARO, will be described.

I. Physical Problem

We shall consider the following physical systems: (i) Coupled spins. (ii) Interacting electrons. (iii) Classical harmonic oscillator chains. The spin systems are models of magnetism. The electron gas is a model of metals as well as a model of celestial bodies, e.g., white dwarfs. The harmonic oscillator chains are models of lattice dynamics, e.g., phonons, and also of defects and impurities in solids. These physical systems are denoted by the Hamiltonian H. At the outset we shall assume that the Hamiltonian is Hermitian, $H^+ = H$, where + denotes Hermitian conjugation. Let A be a dynamical variable of interest. For example, $A = \underset{\sim}{s}_i$, where $\underset{\sim}{s}_i$ means the spin at site i; $A = \underset{\sim}{p}_i$, where $\underset{\sim}{p}_i$ is the momentum of the particle at site i; or $A = g_k = \Sigma_q \, a^+_q \, a_{q-k}$, where a^+_k and a_k are, respectively, the creation and annihilation operator at wave vector k. In general, H is a functional of the dynamical variable A, H = H(A). Since we are interested in the behavior of macroscopic bodies, the thermodynamic limit ($N \to \infty$, $V \to \infty$, but $N/V \to$ const, where N is the number of particles and V is the volume containing these particles) will always be implicitly implied.

Presented at the 9th Army Conference on Applied Mathematics and Computing, 18-21 June 1991, University of Minnesota, Minneapolis, MN.

II. Canonical Approach to Nonequilibrium Problems

Conceptually the canonical approach is very simple. One first obtains the time evolution of A by solving the equation of motion

$$\dot{A}(t) = i[H, A(t)] = iL\, A(t) \tag{1}$$

where [,] means a commutator or a Poisson bracket depending on whether H refers to a quantum or a classical system, and L is the Liouville operator. Given the solution for A(t), one next constructs the autocorrelation function of the following form:

$$\varphi(t) = < A(t)\, A(0) > \tag{2}$$

where one may take A(t=0) = A to be the initial condition, and the angular brackets to mean an ensemble averaging over all possible states of H in the sense of statistical physics. That is,

$$< \ldots > = \mathrm{Tr} \ldots e^{-\beta H} / \mathrm{Tr}\, e^{-\beta H}\,, \tag{3}$$

where Tr means a trace or a sum over the states of H, β is the inverse temperature.

The autocorrelation function $\varphi(t)$ is physically significant. It contains thermodynamic information such as irreversibility and ergodicity. The Laplace transform of $\varphi(t)$ is the scattering function $S(\omega)$, where ω is the frequency, which may be measured by means of, e.g., neutron beams or X-rays or laser. Hence, through $\varphi(t)$, one can determine what microscopic structures give rise to observed macroscopic properties. One can thereby trace macroscopic behavior (e.g., plasma oscillation) to its microscopic origin. A word

about the conventional approach. As we shall see, the canonical approach is not easy to realize. The conventional approach is to obtain the autocorrelation function $\varphi(t)$ directly by solving some phenomenological equations for it, e.g., the Langevin equation. Ordinarily one must employ approximate techniques such as mean-field, stochastic or other similar theories. While very useful and in many ways necessary, the solutions given by the conventional approach have difficulty of being linked to the basic problems posed through H.

Perhaps the best known realized example of the canonical approach is the solution of nearest-neighbor coupled classical linear harmonic oscillator chains.[1] One can obtain the time evolution of the momentum of a tagged particle from the canonical equations of motion and thereby the momentum autocorrelation function. The standard method is to subject H to a unitary transformation: $H \rightarrow \tilde{H} = U\ H\ U^{-1}$, where U is unitary operator which diagonalizes H. This process is equivalent to converting lattice coordinates of the oscillators into normal coordinates. There is a price to be paid for doing this transformation. If, for example, one wishes to follow delocalization of a perturbation initially imparted to the tagged particle, say at $t = 0$, it is virtually impossible to do so in the space of normal coordinates.

This kind of transformation is ad hoc. One must be able to find a unique transformation for each problem. Hence, it is not easy to extract common features of successful solutions. This author has developed a new canonical method which avoids the transformation route.[2] The solutions obtained from this new method possess certain dynamical features. They may be classified so as to provide a universal picture of dynamical processes.

III. Method of Recurrence Relations

The method of recurrence relations is a general method, developed in the early 1980s.[2] It is applicable to all Hermitian systems with both finite or infinite degrees of freedom. The formal solution to the time-evolution eq. (1) may be given as

$$A(t) = e^{iLt} A \tag{4}$$

where L is the Liouville operator. One can imagine that the time evolution of A describes a trajectory in a vector space. Let A(t) be a vector in this space and also $\|A\|$ denote the norm of A. If $\|A(t)\| = \|A\|$, the length of the vector A(t) is an invariant of time. Since H is Hermitian, the Bessel equality is satisfied. The dimensionality d of this space may be finite or infinite, depending on L. If d = 2, for example, A(t) represents a plane rotation in this space. The time evolution must necessarily be of oscillatory motion. As the dimensionality increases, the motion of A in this vector space evidently becomes more complex. The nature of the motion is, however, bound by the geometry of the vector space.

A linear vector space is spanned by its basis vectors, say $\{f_v\}$, v = 0,1,...,d-1. Let these basis vectors be orthogonal, i.e., $(f_v, f_{v'}) = 0$ if $v' \neq v$. Given these basis vectors, we can restate the qualitative statement made about the time evolution of A as follows:

$$A(t) = \sum_{v=0}^{d-1} a_v(t)\, f_v\,, \tag{5}$$

where $a_v(t)$ are some time-dependent real functions. The above orthogonal expansion is still without any physical content since the vector space has not been realized. But we shall see that given a realization of the vector space, there are two parameters--only two--which will completely describe the physical nature of A(t). They are the

dimensionality $d = \{f_v\}$ and the "hypersurfacity" $\sigma = \{\|f_v\|\}$ or, more usefully, $\sigma = \{\|f_v\|/\|f_{v-1}\|\}$.

If the vector space is an abstract one, one may obtain the basis vectors by the Gram-Schmidt orthogonalization process, sometimes known in physics as the projection operator technique of Mori-Zwanzig. If the space is a realized one, then the Gram-Schmidt process is not a natural choice. It is, in fact, a clumsy one. A space is realized if the inner product for the space is defined. Physical problems are by nature not cast in an abstract space but in some realized space. The appropriate inner product was discovered by Kubo.[3] If X and Y are vectors, the inner product of X and Y is given by

$$(X,Y) = \beta^{-1} \int_0^\beta du < e^{uH} X e^{-uH} > - <X><Y> , \tag{6}$$

where $\beta^{-1} = kT$, T temperature, k Boltzmann's constant. Observe that if $[X,H] = 0$,

$$(X,Y) = <XY> - <X><Y> , \tag{7}$$

which represents "fluctuations" in thermodynamics, e.g., susceptibility, specific heat. Through the above given inner product, one begins to see that the space realized by (6) is indeed physically meaningful. The inner product (6) is known as the Kubo scalar product (KSP). The connection between the autocorrelation function and the KSP is self-evident.

One can now find the basis vectors which span the physical space realized by the KSP. It was found that these basis vectors are connected by a recurrence relation,[4]

$$f_{v+1} = \dot{f}_v + D_v f_{v-1} , \quad 0 \le v \le d-1 , \tag{8}$$

where $\dot{f} = i[H,f]$, $D_v = \|f_v\|/\|f_{v-1}\|$, and the boundary conditions $f_{-1} \equiv 0$ and $D_0 \equiv 1$.

Equation (8) will be referred to as RR I. Now, of d basis vectors, there is always one degree of freedom. If one exercises that freedom by choosing $f_0 = A$, the dynamical variable, RR I implies that the remaining d-1 basis vectors can be obtained one by one: $f_0 \rightarrow f_1 \rightarrow f_2 \ldots$ This process continues until the final one which vanishes. This determines the dimensionality d if it is finite. If the process continues indefinitely, the realized space has infinite dimensions. Our choice $f_0 = A$ implies that $a_0(t=0) = 1$ and $a_v(t=0) = 0$, $v \geq 1$. Where is the physics contained here? It is in the hypersurfacity $\sigma = \{D_v\}$, which is a function of H and β.

Returning to the orthogonal expansion (5), we next focus on $\{a_v(t)\}$ the coefficients of expansion. We know at once from the reality and time reversal symmetry conditions that $a_v^*(t) = a_v(t)$ and $a_v(-t) = a_v(t)$. The equation of motion (1) and RR I imply that there is also a recurrence relation for $\{a_v(t)\}$. In fact, one finds that[4]

$$D_{v+1}\, a_{v+1}(t) = -\dot{a}_v(t) + a_{v-1}(t)\,, \qquad 0 \leq v \leq d-1\,, \tag{9}$$

with $a_{-1} \equiv 0$. Equation (9) will be referred to as RR II. Given these two recurrence relations, the orthogonal expansion (5) represents the solution of the equation of motion (1). Since a_v's are functions, they are physically measurable quantities. To illustrate this point, let us consider the simplest case, which is d = 2. For a two-dimensional vector space, according to our scheme: $d = \{f_0, f_1\}$ and $\sigma = \{D_1\}$, all other quantities being zero. Hence, from eq. (9), we obtain

$$D_1\, a_1 = -\dot{a}_0 \tag{10a}$$

$$-\dot{a}_1 + a_0 = 0\,. \tag{10b}$$

Equations (10a,b) are solved at once:

$$a_0 = \cos \omega t \tag{11a}$$

$$a_1 = \sin \omega t/\omega \,, \tag{11b}$$

where $\omega^2 = D_1$. Note that

$$a_0^{\,2} + D_1 a_1^{\,2} = 1 \tag{12}$$

which is a statement of Bessel's equality. This simplest example turns out to be none too trivial. It represents the basic structure of the dynamics for mean field or RPA theory!

IV. Formal Properties

There are a number of useful properties contained in the orthogonal expansion (5) now that our space is the physical space. Exercising one degree of freedom at hand, we choose as before $f_0 = A$, where A is the dynamical variable. Then,

$$a_0(t) = (A(t), A)/(A, A) = R(t) \,, \tag{13}$$

which we recognize as the relaxation function of linear response theory.[3] The memory function M(t) can be shown to be related to the relaxation function R(t).[2] Let $\tilde{a}_v(z) = \mathcal{J}[a_v(t)]$ where $\mathcal{J}$ is the Laplace transform operator. If $\mathcal{J}$ is applied to RR II, we obtain:

$$1 = \tilde{a}_0 + D_1 \tilde{a}_1 \tag{14a}$$

$$\tilde{a}_{v-1} = z\,\tilde{a}_v + D_{v+1}\,\tilde{a}_{v+1} \,, \quad 1 \le v \le d-1 \,. \tag{14b}$$

Equation (14a) represents the fluctuation-dissipation theorem in linear response theory.

The two equations (14a,b) imply that

$$\tilde{a}_0 = 1/z + D_1/z + D_2/z + \dots , \tag{15}$$

a continued fraction of Stieltjes, first derived in the statistical physics context by Mori.[5] By applying the inverse transform $\mathcal{F}^{-1}$ on (15), we can obtain $a_0(t)$. Given a_0, we can obtain a_1, a_2, etc., successively by the application of RR II.

Also, we note that the formal structure of RR II is restrictive. It forbids certain types of functions for $\{a_v(t)\}$. Excluded are, for example, the simple exponential and the entire class of the orthogonal polynomials. Allowed or admissible are the Gaussian, circular functions, hyperbolics, the Bessel functions of integer and half-integer orders, the hypergeometric function of an even argument, the elliptic functions.

The method of recurrence relations is distinguished from noncanonical approaches. It does not directly solve the equation of motion; instead it solves the equation of motion by finding admissible solutions. It requires two essential ingredients: the dimensionality and hypersurfacity. These two quantities allow a unique way of classifying physical solutions. The method is canonical in its approach to dynamics, i.e., both A(t) and $\{a_v(t)\}$ are obtained. Hence, the solutions are richer. The method has recovered, as far as we know, all the existing exactly solvable problems, usually much more simply.[6]

V. Physical Applications

One simple application is afforded by the problem of time evolution in a classical nearest-neighbor coupled harmonic oscillator monatomic chain. Let N be the number of atoms in the chain, taken to be an even number, which will be allowed to grow indefinitely. We shall impose periodic boundary conditions on the chain for simplicity. It has been shown[7] that for this problem, the two key quantities are:

$$d = N/2 + 1 \tag{16}$$

$$\sigma = \{2,2\} \qquad \text{if } N = 2 \tag{17a}$$

$$= \{2,1,1,2\} \qquad \text{if } N = 4 \tag{17b}$$

$$= \{211 \ldots 112\} \qquad \text{if } N < \infty \,. \tag{17c}$$

There is a physical dimension in σ, $\sqrt{k/m}$, where k is the spring constant and m the mass of an atom, but it has been set to unity.

The front-end symmetry in σ is remarkable (see 17c). If $d < \infty$ (i.e., $N < \infty$), the relaxation functions $a_v(t)$'s are all circular functions. It means that there is a finite recurrence time or Poincare cycle. If $N \to \infty$, hence, $d \to \infty$, the symmetry in σ is destroyed

$$\sigma = \{2111 \ldots 111 \ldots\} \,. \tag{17d}$$

It sets up irreversibility in the time evolution behavior. Ours is an example of irreversibility in a Hermitian system. The necessary (and probably sufficient) condition for irreversibility is thus $d \to \infty$ (as a result of $N \to \infty$). We have examples where $N \to \infty$ does not necessarily signify $d \to \infty$. But in this problem, d and N are simply related. See (16). It indicates that a perturbation imparted to a tagged oscillator atom propagates atom to atom until the last one and then it is reflected if N is finite. It never returns if N is not finite, inducing irreversibility.

If $N \to \infty$, using (17d) in (15), we obtain

$$\tilde{a}_0(z) = (z^2 + 4)^{-1/2} \,. \tag{18}$$

Hence,

$$a_v(t) = J_v(t) \,, \tag{19}$$

where J_v is the Bessel function of order $v = 0.1,2,...$ The square root singularity (18) indicates that there are two Riemann sheets in the plane of z. The physical significance of these sheets appears if one makes the mass of the tagged oscillator atom, say m_0, different from that of the rest. Let $s = m/m_0$. One then obtains,

$$\sigma = (2s\ 1\ 1\ 1\ \ldots) \,, \tag{20}$$

which differs from the hypersurface of a pure monatomic chain only in the first member. As a result,

$$\tilde{a}_0(z) = \frac{1}{pz + \sqrt{z^2+4}} \,, \tag{21}$$

where $p = s^{-1} - 1$. Except when $p = 0, \pm 1$, there is now a simple pole in addition to the square root singularity. If $p < 0$, the simple pole lies in the "physical" sheet; if $p > 0$, the pole lies in the "nonphysical" sheet. The two sheets are distinguished when one obtains $a_0(t)$ from (21) via the inverse transform. One takes the physical sheet only. The singular function (21) is closely related to a function which appears in the Joukowski transformation in the theory of aerofoils.

VI. Discussion

Space limitation does not permit us to give a detailed discussion of our work on a diatomic chain here. It suffices to mention that one can obtain d and σ as described for a

monatomic chain. If $N \to \infty$, $d \to \infty$ as before. The elements of the hypersurface are no longer constant (i.e., periodicity of 1), but now they form a set of a periodicity of two, i.e.,

$$\sigma = \{2a,b,b,a,a,b,b,a,a,...\} , \qquad (22)$$

where a and b are certain mass parameters. The analytic structure of the resultant $\tilde{a}_0(z)$ is evidently richer than that for a monatomic chain. There are in fact additional finite branch lines. One can obtain the autocorrelation function in various regimes of the mass parameter.

Other limiting cases (e.g., next n.n., constant-coupled h.o. chains) may also be studied in this manner. Delocalization of an excitation in these models can be straightforwardly determined by the method of recurrence relations. One particular advantage of this method is its ability to establish dynamic equivalence. We mention that such an equivalence between h.o. chains and a 2D quantum electron gas at long wavelengths was recently established.[7]

Acknowledgments

This work has been supported by the NSF and ARO/CRDEC. The material presented here is based on a seminar presented at the Department of Mathematical Physics, Leningrad State University, Leningrad, USSR in May of 1991. The author is grateful to Professors B. Pavlov and Y. N. Demkov for their hospitality while being a guest at the institute. Several co-workers have contributed to the developments of the method of recurrence relations including J. B. Hong, J. Florencio, R. Dekeyser and M. B. Yu.

References

1. A. A. Maradudin et al., Theory of Lattice Dynamics, Academic, N.Y., 1971.
2. M. H. Lee, Phys. Rev. B **26**, 1072 (1982).
3. R. Kubo, Rep. Prog. Phys. **29**, 255 (1966).
4. M. H. Lee, Phys. Rev. Lett. **49**, 1072 (1982).
5. H. Mori, Prog. Theor. Phys. **34**, 309 (1965).
6. M. H. Lee, J. Hong and J. Florencio, Physica Scripta **T19**, 498 (1987).
7. M. H. Lee, J. Florencio and J. Hong, J. Phys. **A22**, L331 (1989).
8. M. B. Yu, J. H. Kim and M. H. Lee, J. Luminesc. **45**, 144 (1990).

IMAGE SINGULARITIES OF GREEN'S FUNCTIONS FOR ANISOTROPIC ELASTIC HALF–SPACES AND BIMATERIALS*

T. C. T. Ting
Department of Civil engineering, Mechanics and Metallurgy
University of Illinois at Chicago
Box 4348, Chicago, IL 60680 USA

ABSTRACT Using Stroh's formalism simple explicit expressions of Green's functions for anisotropic elastic half–spaces and bimaterials subject to line forces and line dislocations are presented. One of the novel features is that, knowing the Green's function for an infinite space, Green's functions for half–spaces and bimaterials can be written down immediately with very little derivation. The other novel feature is the physical interpretations of Green's functions. The Green's function for a half–space consists of ten Green's functions for the infinite space. One of the ten Green's functions has its singularities located in the half–space where they are prescribed. The other nine represent image singularities which are located outside of the half–space not occupied by the material. The locations of the nine image singularities as well as the nature of the singularities are presented explicitly. For bimaterials which consists of two anisotropic half–spaces bonded together, there are nine image singularities each for the two materials. Again the locations and the nature of the singularities are presented explicitly. We also suggest graphical solutions for finding the locations of these singularities. Since the Green's function for an infinite space has a real form solution, this implies that Green's functions for half–spaces and bimaterials can have a real form solution. The image singularities for degenerate materials for which isotropic materials are a special case are discussed briefly. An anomaly is that the image singularities for degenerate materials are not simply line forces and line dislocations. Although the Green's functions obtained here are for line forces and line dislocations, the results can be applied to Green's functions for other types of singularities such as concentrated couples. In particular, the locations of image singularities presented here are independent of the type of singularity concerned.

EXTENDED SUMMARY. The Green's function for two–dimensional deformations of an infinite anisotropic elastic material subject to a line dislocation has been obtained in [1–5]. Further developments of Green's functions to include line forces were given in [6]. Green's functions for an infinite medium have also been found for anisotropic composite spaces [7] and for the more general case of angularly inhomogeneous anisotropic materials [8,9].

Green's functions for anisotropic half–spaces and bimaterials have been considered by many investigators (see [10–17], for example). In the earlier work Green's functions for half–spaces are obtained from the Green's function for an infinite space by adding a distribution of forces along the surface of the half–space so that the net surface traction vanishes. Likewise, Green's functions for bimaterials are obtained by adding a distribution of forces and dislocations along the interface to maintain the continuity of displacement and surface traction at the interface. With this approach the solution is not explicit in

*Supported by the U. S. Army Research Office. The full length paper is to appear in the Quarterly Journal of Mechanics and Applied Mathematics.

that the final solution requires integration of the distributed forces and dislocations along the free surface or the interface. Progress has been made recently in obtaining Green's functions. The simplest solutions of Green's functions for half–spaces and bimaterials appear to be the ones obtained by Suo [15] using the methods of analytical continuation. However, Suo did not give the solution in full, only in an abbreviated form. A breakthrough was made by Hwu and Yen [18] whose approach for finding Green's functions for an elliptic hole in an infinite anisotropic medium suggests that one can write down Green's functions for half–spaces and bimaterials immediately with little derivation. This is one purpose of this paper. The other purpose of this paper is to interpret physical meanings of the Green's functions so obtained.

The basic formalism of Stroh [2,19–21] is outlined in Section 2 and some identities which are needed in the sequel are presented in Section 3. The Green's function for an infinite space due to a line force and a line dislocation is given in Section 4 which serves as the bases for the rest of the paper. Section 5 deals with the surface Green's function for a half–space while the Green's function for a half–space in which the singularities are located inside the half–space are presented in Section 6. It is shown that the Green's function for a half–space contains nine image singularities. The locations and the nature of these image singularities are given explicitly. Graphical solutions of the locations of the image singularities are presented in Section 7. Section 8 studies Green's functions for bimaterials where it is shown that there are nine image singularities each for the two materials in the bimaterials. The locations and the nature of these image singularities are also given explicitly. In the last section we discuss briefly the image singularities of Green's functions for degenerate materials. We also discuss the generality of the methods presented here which can be applied to Green's functions for half–spaces and bimaterials due to other types of singularities.

REFERENCES

[1] J. D. Eshelby, W. T. Read, and W. Shockley, "Anisotropic elasticity with applications to dislocation theory," Acta Metall. 1, 251–259 (1953).

[2] A. N. Stroh, "Dislocations and cracks in anisotropic elasticity," Phil Mag. 3, 625–646 (1958).

[3] J. R. Willis, "Stress field produced by dislocations in anisotropic media," Phil. Mag. 21, 931–949 (1970).

[4] K. Malen, "A unified six–dimensional treatment of elastic Green's functions and dislocations," Phys. Status Solidi B 44, 661–672 (1971).

[5] R. J. Asaro, J. P. Hirth, D. M. Barnett and J. Lothe, "A further synthesis of sextic and integral theories for dislocations and line forces in anisotropic media," Phys. Status Solidi B 60, 261–271 (1973).

[6] D. M. Barnett and J. Lothe, "Line force loadings on anisotropic half–spaces and wedges," Phys. Norv. 8, 13–22 (1975).

[7] T. C. T. Ting, "Line forces and dislocations in anisotropic elastic composite wedges and spaces," Phys. Status Solidi B 146, 81–90 (1988).

[8] H. O. K. Kirchner, "Line defects along the axis of rotationally inhomogeneous media," Phil. Mag. A 55, 537–542 (1987).

[9] T. C. T. Ting, "Line forces and dislocations in angularly inhomogeneous anisotropic elastic wedges and spaces," Q. Appl. Math. **47**, 123–128 (1989).

[10] J. R. Willis, "Fracture mechanics of interface cracks," J. Mech. Phys. Solids. **19**, 353–368 (1971).

[11] J. Braekhus and J. Lothe, "Dislocations at and near planar interfaces," Phys. Status Solidi B **43**, 651–657 (1971).

[12] S. Nakahara and J. R. Willis, "Some remarks on interfacial dislocations," J. Phys. F: Metal Phys. **3**, L249–254 (1973).

[13] D. M. Barnett and J. Lothe, "An image force theorem for dislocations in anisotropic bicrystals," J. Phys. F. **4**, 1618–1635 (1974).

[14] V. K. Tewary, R. H. Wagoner and J. P. Hirth, "Elastic Green's function for a composite solid with a planar interface" J. Mater. Res. **4**, 113–123 (1989).

[15] Zhigang Suo, "Singularities, interfaces and cracks in dissimilar anisotropic media," Proc. R. Soc. Lon. A **427**, 331–358 (1990).

[16] Jianmin Qu and Qianqian Li, "Interfacial dislocation and its application to interface crack in anisotropic bimaterials," J. Elasticity **25** (1991), in press.

[17] Jianmin Qu, "Green's functions in anisotropic bimaterials," in Modern Theory of Anisotropic Elasticity and Applications, J. J. Wu, T. C. T. Ting, D. M. Barnett, eds. SIAM Pub., in press (1991).

[18] Chyanbin Hwu and Wen J. Yen, "Green's functions of anisotropic plates containing elliptic holes," Int. J. Solids Structures **27**, 1705–1719 (1991).

[19] A. N. Stroh, "Steady state problems in anisotropic elasticity," J. Math. Phys. **41**, 77–103 (1962).

[20] D. M. Barnett and J. Lothe, "Synthesis of the sextic and the integral formalism for dislocations, Greens functions and surface waves in anisotropic elastic solids," Phys. Nor., **7**, 13–19 (1973).

[21] P. Chadwick and G. D. Smith, "Foundations of the theory of surface waves in anisotropic elastic materials," Adv. Appl. Mech. **17**, 303–376 (1977).

[22] K. A. Ingebrigtsen and A. Tonning, "Elastic surface waves in crystals," Phys. Rev. **184**, 942–951 (1969).

[23] S. A. Gundersen, D. M. Barnett and J. Lothe, "Rayleigh wave existence theory. A supplementary remark," Wave Motion **9**, 319–321 (1987).

[24] T. C. T. Ting, "Some identities and the structure of $\mathbf{N}_1$ in the Stroh formalism of anisotropic elasticity," Q. Appl. Math. **46**, 109–120 (1988).

[25] J. Lothe and D. M. Barnett, "On the existence of surface–wave solutions for anisotropic half–spaces with free surface," J. Appl. Phys. **47** 428–433 (1976).

[26] P. Chadwick and T. C. T. Ting, "On the structure and invariance of the Barnett–Lothe tensors," Q. Appl. Math. **45**, 419–427 (1987).

[27] T. C. T. Ting, "The eigenvectors of the **S** matrix and their relations with line dislocations and forces in anisotropic elastic solids," in Micromechanics and Inhomogeneity, The Toshio Mura Anniversary Volume. Springer–Verlag, N.Y. 449–467 (1990).

[28] T. C. T. Ting, "On the orthogonal, Hermitian and positive definite properties of the matrices $i\mathbf{B}^{-1}\overline{\mathbf{B}}$ and $-i\mathbf{A}^{-1}\overline{\mathbf{A}}$ in anisotropic elasticity," J. Elasticity, in press.

[29] T. C. T. Ting, "Barnett–Lothe tensors and their associated tensors for monoclinic materials with the symmetry plane at $x_3 = 0$," J. Elasticity, in press.

[30] J. P. Hirth and J. Lothe, Theory of Dislocations. Wiley, N. Y. (1982).

[31] J. Dundurs, "Elastic interaction of dislocations with inhomogeneities," in Mathematical Theory of Dislocations. T. Mura, ed. ASME, N. Y. 70–115 (1969).

[32] T. C. T. Ting, "The anisotropic elastic wedge under a concentrated couple," Q. J. Mech. Appl. Math. **41**, 563–578 (1988).

THE COMPUTATION OF CRYSTALLINE MICROSTRUCTURE*

MITCHELL LUSKIN†, AND CHARLES COLLINS‡

Abstract. We describe a two-dimensional model of crystalline martensitic microstructure, and we present a new visualization of computational results for the finite element approximation of solutions to the variational problem with microstructure on a sequence of refined meshes.

1. Introduction. We present computational results for a two-dimensional model of crystalline martensitic microstructure which was proposed by Ericksen and James. This two-dimensional model has the property that the energy density is frame-indifferent and has two symmetry-related energy wells. Variational problems of this type do not always attain their minimum value for any admissible deformation [BJ2]. Rather, the solution must often be described in terms of a microstructure since the deformation gradients of minimizing sequences can have oscillations with amplitude which remains finite and with wavelength which converges to zero.

A mathematical theory of microstructure has been developed during the past several years to describe solutions to these variational problems [BJ1, ChKi, E1, E2, J, Ki, Ko]. This theory also gives a recipe for the evaluation of macroscopic properties of crystals with microstructure.

Computations for a three-dimensional model for InTl, a shape-memory alloy with symmetry-related (martensitic) variants, were given in [CoL]. These computations successfully obtained microstructure on the scale of the grid and the austenitic-martensitic interface. We have found that the computation of three-dimensional deformations for crystals with symmetry-related microstructure requires large amounts of computing time. We have developed the two-dimensional model so that we can more quickly experiment with our algorithms and so that we can more easily do computations which are relevant to the development of the continuum theory. For instance, the two-dimensional model has been used to study complex microstructure involving the mixture of more than two deformation gradients [Co].

The development of a theory to rigorously analyze the numerical approximation of microstructure in crystals was begun in [CoKiL, CoL] and has been further developed in [ChCo]. These concepts have also recently been applied to the numerical approximation of the fine scale structure of the magnetization field of some ferromagnetic materials [LM].

*This work is part of the Transitions and Defects in Ordered Materials Project and was supported in part by the ARO through grants DAAL03-88-K-0110 and DAAL03-89-G-0081, the Army High Performance Computing Research Center, the Cray Research Foundation, and by a grant from the Minnesota Supercomputer Institute.

†School of Mathematics, University of Minnesota, Minneapolis, Minnesota 55455.

‡Department of Mathematics, University of Michigan, Ann Arbor, Michigan 48109.

2. Two-dimensional model. The bulk energy of a two-dimensional crystal is modeled by

$$\mathcal{E}(y) = \int_{\Omega} \phi(\nabla y(x), \theta)\, dx$$

where $\Omega \subset \mathbf{R}^2$ is the reference configuration for the crystal, $y(x) : \Omega \to \mathbf{R}^2$ is the deformation, θ is the temperature, and ϕ is the energy density.

We use the energy density proposed by Ericksen and James which is given by

$$\phi(F, \theta) = \kappa_1(\text{Trace}\, C - 2)^2 + \kappa_2 c_{12}^2 + \kappa_3 \left(\left(\frac{c_{11} - c_{22}}{2} \right)^2 - \epsilon^2 \right)^2$$

where

$$C = F^T F = \begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$$

is the Cauchy-Green strain tensor; κ_1, κ_2, and κ_3 are elastic moduli; and ϵ is the transformation strain.

The energy density ϕ is frame-indifferent, i.e.,

$$\phi(RF, \theta) = \phi(F, \theta)$$

for any rotation R and ϕ has the symmetry group of the square, i.e.,

$$\phi(FR_\eta, \theta) = \phi(F, \theta) \qquad \text{for } \eta = \pi/2, \pi, 3\pi/2$$

where R_η is the rotation matrix with angle η

$$R_\eta = \begin{pmatrix} \cos\eta & -\sin\eta \\ \sin\eta & \cos\eta \end{pmatrix}.$$

The energy density $\phi(F, \theta)$ attains its minimum value at the symmetry-related deformation gradients $F^T F = U_0^2$ or $F^T F = U_1^2$ where

$$U_0 = \begin{pmatrix} \sqrt{1-\epsilon} & 0 \\ 0 & \sqrt{1+\epsilon} \end{pmatrix} \qquad \text{and} \qquad U_1 = \begin{pmatrix} \sqrt{1+\epsilon} & 0 \\ 0 & \sqrt{1-\epsilon} \end{pmatrix}.$$

We note that U_0 and U_1 are symmetry-related since

$$U_0 = R_{-\pi/2} U_1 R_{\pi/2}.$$

The unstressed austenitic phase is represented in this model by the deformation gradient $F = I$ where I is the identity matrix, and the unstressed martensitic phase is

represented by the symmetry-related deformation gradients $F = U_0$ and $F = U_1$ which represent different "variants" of the martensite. We assume that the temperature θ is held fixed in the body below the transition temperature between the austenitic phase and the martensitic phase. At such a temperature the unstressed martensitic phase should be a global minimum of the energy density while the unstressed austenitic phase should be a local minimum of the energy density. Since the purpose of the two-dimensional model is to provide a model to test the effectiveness and efficiency of our algorithms for the computation of martensitic microstructure, we have simplified our energy density by removing the local minimum at unstressed austenitic phase $F = I$.

We chose the values of the material coefficients to resemble the elastic moduli and transformation strain for InTl which has a cubic austenitic phase and a tetragonal martensitic phase. To simulate the properties of the three-dimensional energy density for InTl proposed by Ericksen which has a cubic symmetry group [E1, E2], we have constructed the two-dimensional energy density ϕ with the symmetry group of the square and we have used the moduli

$$\kappa_1 = 10, \quad \kappa_2 = 3, \quad \kappa_3 = 1, \quad \epsilon = .1.$$

The transformation strain ϵ is taken to be larger than the physical strain to enable numerical computations to be done on a coarser grid. We have also represented the tetragonal martensitic phase by a two-dimensional rectangular phase.

For an unstressed solid the above model allows the existence of interfaces which separate different variants of martensite. These interfaces or "twin lines" are given by lines across which the deformation is continuous, but across which the deformation gradient is discontinuous. A continuous deformation $y(x)$ exists such that

$$\begin{aligned} \nabla y(x) &= U_0 \qquad \text{where } x \cdot n > 0, \\ \nabla y(x) &= R_\zeta U_1 \qquad \text{where } x \cdot n < 0 \end{aligned}$$

if and only if

$$R_\zeta U_1 = U_0 + a \otimes n \tag{1}$$

for some angle ζ and some vector a. The set of solutions to (1) is given by ζ such that $\cos\zeta = \sqrt{1-\epsilon^2}$, $\sin\zeta = \pm\epsilon$, and

$$a = \sqrt{2}\epsilon\left(\sqrt{1-\epsilon}, \pm\sqrt{1+\epsilon}\right), \qquad n = \frac{\sqrt{2}}{2}(1, \mp 1).$$

Thus, there are two possible families of parallel interfaces.

For $\beta(s)$ taking only the values 0 and 1 the continuous deformation

$$y(x) = U_0 x + a\int_0^{x\cdot n} \beta(t)dt,$$

has discontinuous deformation gradients on lines orthogonal to n since

$$\nabla y(x) = U_0 \qquad \text{where } \beta(x \cdot n) = 0,$$
$$\nabla y(x) = R_\zeta U_1 \qquad \text{where } \beta(x \cdot n) = 1.$$

It follows that $\phi(\nabla y(x)) = 0$ for all $x \in \Omega$.

We recall that for any 2×2 nonsingular matrices V_1 and V_2 such that the eigenvalues of $(V_1 V_2^{-1})^T (V_1 V_2^{-1})$ satisfy $\lambda_1 < 1 < \lambda_2$, there exists a rotation R and nonzero vectors b and m such that [BJ1]

$$RV_1 = V_2 + b \otimes m.$$

In particular, there exists a rotation R and nonzero vectors b and m such that

$$RI = U_0 + b \otimes m. \tag{2}$$

Now (2) implies that the two-dimensional model allows a continuous deformation with an interface separating a region of austenite $F = RI$ from a region of martensite $F = U_0$. However, the three-dimensional model of InTl does not allow an interface to separate a region of austenite from a region containing a single variant of martensite, and this is confirmed by experimental observations. Rather, for the three-dimensional model an interface can separate regions of austenite and martensite only if the martensitic region is a fine-scale mixture of more than one variant of martensite [BJ1], and this is also confirmed by experimental results. By constructing our two-dimensional energy density so that $F = I$ is not a local minimum, we have eliminated the possibility of a spurious interface between regions of austenite and of single-variant martensite for a two-dimensional unstressed crystal.

The goal of our computations is to compute the displacement, y_h, which minimizes the bulk energy among all admissible finite element displacements on a mesh with length scale h. The deformation on the boundary of the body is constrained to equal

$$y(x) = \left(\frac{1}{2}U_0 + \frac{1}{2}R_\zeta U_1\right) x, \qquad x \in \partial\Omega. \tag{3}$$

There does not exist a deformation which has the minimum energy $\mathcal{E} = 0$ and which satisfies the boundary conditions [BJ2]. Rather, if $\mathcal{E}(y_n) \to 0$ as $n \to \infty$ for a sequence of deformations satisfying the boundary conditions, then the amplitude of the oscillations of $\nabla y_n(x)$ remains finite as $n \to \infty$, but the wavelength of the oscillations becomes arbitrarily small. The "microstructure" solution to this problem is unique, though, and is given by the mixture of U_0 and $R_\zeta U_1$ in equal proportions [BJ2].

We can give an analytic treatment of the microstructure for this problem by defining

$$\beta(x) = \begin{cases} 0 & \text{if } 2m < x < 2m+1, \text{ for } m \text{ an integer} \\ 1 & \text{if } 2m-1 < x < 2m, \text{ for } m \text{ an integer} \end{cases}$$

and

$$\beta_\delta(x) = \beta(x/\delta).$$

Then

$$y_\delta(x) = U_0 x + a \int_0^{x \cdot n} \beta_\delta(t) dt$$

has minimum energy ($\mathcal{E}(y_\delta) = 0$), but $y_\delta(x)$ does not satisfy the boundary conditions (3). However,

$$y_\delta(x) \to \left(\frac{1}{2}U_0 + \frac{1}{2}R_\zeta U_1\right) x$$

uniformly as $\delta \to 0$. It follows that we can modify $y_\delta(x)$ near the boundary to construct a deformation, $\tilde{y}_\delta(x)$, which satisfies the boundary conditions (3) and such that

$$\mathcal{E}(\tilde{y}_\delta) \to 0 \qquad \text{as } \delta \to 0.$$

The scale of the microstructure for $\tilde{y}_\delta(x)$ is δ.

3. Two-dimensional computations. We used continuous, bilinear finite elements, and we developed an optimization algorithm based on the conjugate gradient algorithm [Co]. The crystal was oriented so that the lines of discontinuity of the deformation gradient are diagonal to the mesh—the most difficult test.

In each local element, we evaluate the deformation gradient $F = \nabla y_h$ at the center, and we shade the area in the local element to display the function

$$\psi(F) = \frac{\|F^T F - U_0^2\|}{\|F^T F - U_0^2\| + \|F^T F - U_1^2\|},$$

where $\|A\| = \sum A_{ij}^2$. In Figure 1 we see the results of our numerical experiments with the boundary conditions (3) where $\sin\zeta = -\epsilon$ for mesh lengths $h = 1/16$, $h = 1/32$, $h = 1/48$, and $h = 1/64$. The element is white when $F^T F \approx U_0^2$, the element is black when $F^T F \approx U_1^2$, and the element is colored varying shades of gray to denote the distance of the element deformation gradient to one of the energy wells.

Our algorithm has successfully computed a microstructure on the scale of each successively finer mesh. Since the computed microstructure is not completely regular, we have actually computed a local minimum of the finite element optimization problem. Nevertheless, our computed local minimum has small enough energy so that it exhibits the microstructure of the global minimum.

The number of local minima becomes arbitrarily large as the mesh is refined since we can have local minima with oscillations of arbitrarily small wavelength and since the number of possible "defects" where the microstructure is irregular can become arbitrarily large.

The video [CoLR] shows the path of our conjugate gradient algorithm to a local minimum. The microstructure organizes itself so that the energy density is small in disjoint regions. As these regions coalesce or approach the boundary, the unique microstructure that is compatible with the boundary conditions is chosen throughout the entire crystal.

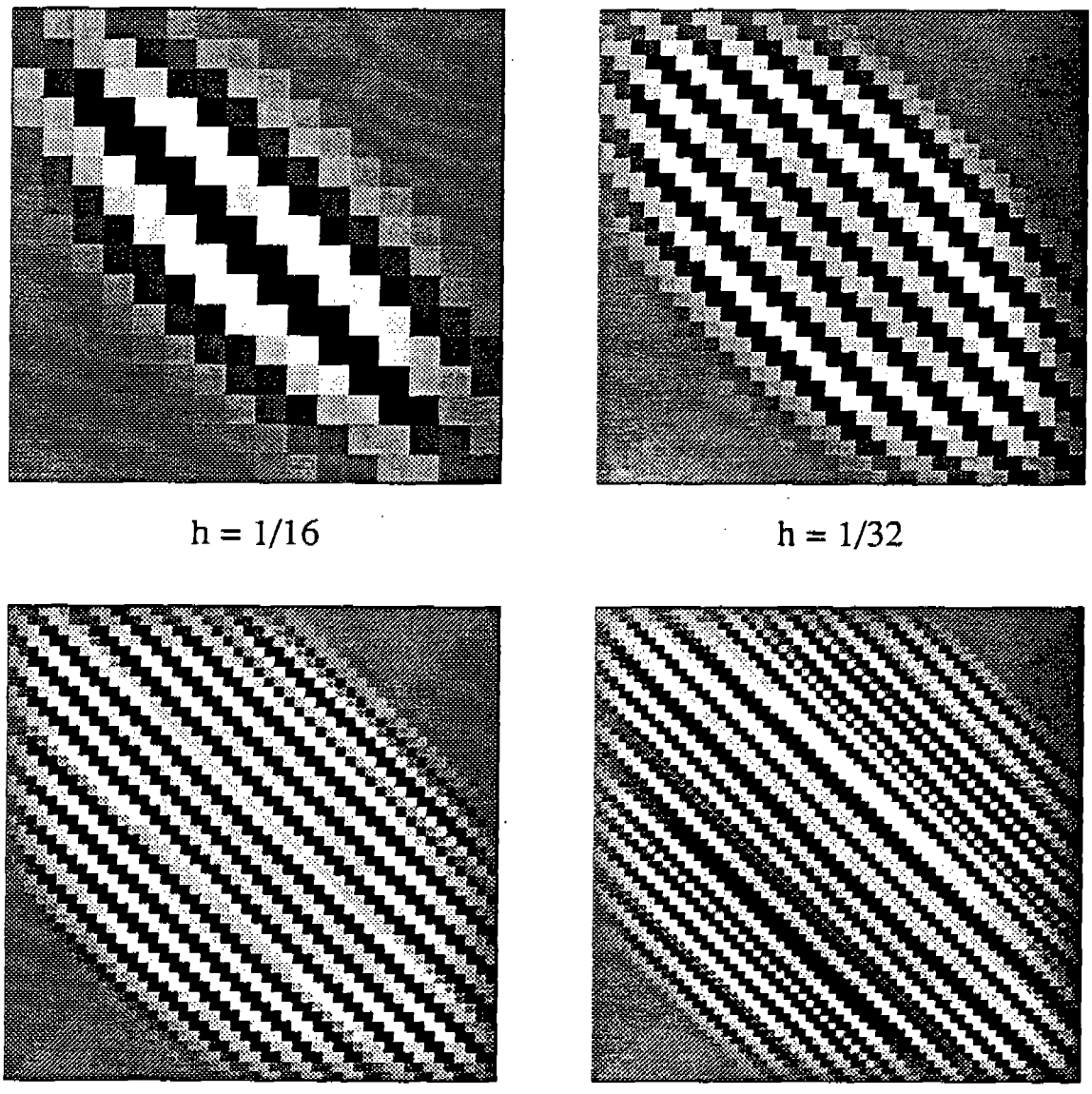


Figure 1. Deformation gradient of $y_h(x)$ for meshes with scales $h = 1/16$, $h = 1/32$, $h = 1/48$, and $h = 1/64$. The deformation gradient at the midpoint of each local element is displayed. The local element is shaded according to the value of $\psi(F)$.

REFERENCES

[BJ1] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[BJ2] J. M. Ball and R. D. James, *Experimental tests of a theory of fine microstruture and the two-well problem*, preprint.

[ChCo] M. Chipot and C. Collins, *Numerical approximations in variational problems with potential wells*, SIAM J. Numer. Anal. (to appear).

[ChKi] M. Chipot and D. Kinderlehrer, *Equilibrium configurations of crystals*, Arch. Rational Mech. Anal, 103 (1988), pp. 237–277.

[Co] C. Collins, *Computation of twinning*, in *in Microstructure and Phase Transitions*, IMA Volumes in Mathematics and its Applications (James, Kinderlehrer, and Luskin, eds.), Springer-Verlag, New York, to appear, pp..

[CoKiL] C. Collins, D. Kinderlehrer, and M. Luskin, *Numerical approximation of the solution of a variational problem with a double well potential*, SIAM J. Numer. Anal., 28 (1991), pp. 321–332.

[CoL1] C. Collins and M. Luskin, *The computation of the austenitic-martensitic phase transition*, in *Partial Differential Equations and Continuum Models of Phase Transitions, Lecture Notes in Physics 344, M. Rascle, D.Serre, and M. Slemrod, eds.*, Springer-Verlag, 1989, pp. 34–50.

[CoL2] C. Collins and M. Luskin, *Optimal order estimates for the numerical approximation of the solution of a variational problem with a double well potential*, Math. Comp. (to appear).

[CoLR] C. Collins, M. Luskin, and J. Riordan, *Computational images of crystalline microstructure*, in *AMS Video Series*, 1991.

[E1] J. L. Ericksen, *Some constrained elastic crystals*, in *Material Instabilities in Continuum Mechanics and Related Problems, J. M. Ball, ed.*, Oxford University Press, 1987, pp. 119–137.

[E2] J. L. Ericksen, *Constitutive theory for some constrained elastic crystals*, Int. J. Solids and Structures, 22 (1986), pp. 951–964.

[J] R. James, *Basic principles for the improvement of shape-memory and related materials*, in *Smart Materials, Structures, and Mathematical Issues, C. Rogers, ed.*, Technomic Publishing Co., 1989.

[Ki] D. Kinderlehrer, *Remarks about equilibrium configurations of crystals*, in *Material Instabilities in Continuum Mechanics and Related Problems, J. M. Ball, ed.*, Oxford University Press, 1987, pp. 217–242.

[Ko] R. Kohn, *Relaxation of a double-well energy*, Continuum Mechanics and Thermodynamics (to appear, 1991).

[L] M. Luskin, *Numerical analysis of microstructure for crystals with a nonconvex energy density*, in *The Metz Days Surveys, 1989-90*, Pitman Research Notes in Mathematics, M. Chipot and J. Saint Jean Paulin, eds., Longman Company, UK, to appear.

[LM] M. Luskin and L. Ma, *Analysis of the finite element approximation of microstructure in micromagnetics*, SIAM J. Numer. Anal. (to appear).

On Dynamical Aspects of a Phase Transition Problem *

Hiroaki Fujimoto
Harumi Hattori
Department of Mathematics
West Virginia University
Morgantown, WV 26506

Abstract

In this note we discuss a dynamical systems approach to a phase transition problem based on the Korteweg theory of capillarity. We consider the existence of a global solution to show that we have a dynamical system. We discuss the stability and bifurcation analysis of stationary solutions and then we study the connecting orbit problems in the semiflow. The connection matrix is a useful tool to discuss qualitative aspects of the dynamical behavior of solutions. We also dicuss the slowly varying solutions and preliminary numerical results for this are given.

1 Introduction.

In this note we study dynamical aspects of the following system of parabolic equations

$$\begin{aligned} p_t &= \nu p_{xx} - \eta q_{xx} + \sigma(q) - P, \\ q_t &= p_{xx}, \end{aligned} \tag{1.1}$$

where $x \in [0,1]$ and the boundary conditions are given by

$$\begin{aligned} p_x(0,t) = 0, \quad p(1,t) = 0, \\ q_x(0,t) = 0, \quad q_x(1,t) = 0, \end{aligned} \tag{1.2}$$

and the initial data are given by

$$p(x,0) = f(x), \quad q(x,0) = g(x). \tag{1.3}$$

The above system is derived from an equation

$$u_{tt} = \sigma(u_x)_x + \nu u_{xxt} - \eta u_{xxxx} \tag{1.4}$$

*This work was supported by Army Grant DAAL 03-89-G-0088.

with boundary conditions

$$u(0,t) = 0, \quad \sigma(u_x(1,t)) + \nu u_{xt}(1,t) - \eta u_{xxx}(1,t) = P, \tag{1.5}$$

$$u_{xx}(0,t) = 0, \quad u_{xx}(1,t) = 0, \tag{1.6}$$

by setting $p = \int_1^x u_t\,dx$ and $q = u_x$. Equation (1.4) models a bar which goes through a phase transition. The boundary conditions (1.5) show that the bar is under a soft loading device. The boundary conditions (1.6) are the natural boundary conditions for the corresponding variational problem. The terms with the coefficients ν and η are called viscosity and capillarity terms, respectively. In what follows, we assume that σ is given by Fig. 1.1. In this figure $(0, \alpha^*]$ and $[\beta^*, \infty)$ are called the α-phase and the β-phase, respectively. They correspond to the different phases of the material.

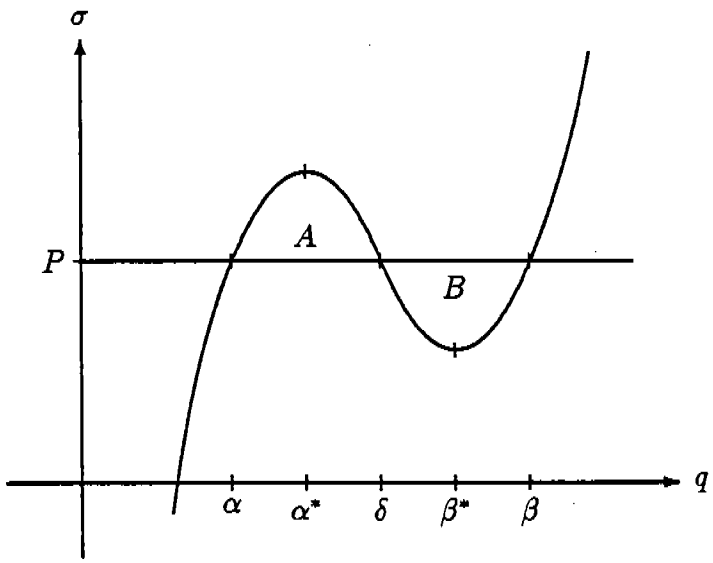


Figure 1.1

The capillarity term was first considered by Korteweg [6]. Recently, various effects of this term have been discussed, for example, in [1], [2], [3], [7], [8],

In what follows, we discuss first the existence of global solutions, the stability of stationary solutions and the bifurcation diagram, the connecting orbit problems in the semiflow, and the slow motions. We omit most of the proofs, as they will appear in [5].

2 Existence of a global solution.

We state the theorem establishing the existence of a global solution to (1.1) and (1.2).

Theorem 2.1 *Suppose $\sigma \in C^3$ and that $(f(x), g(x)) \in H^1(0,1)$ and are compatible with the boundary conditions (1.2). Then there exists a unique global solution $(p,q) \in H^2(0,1)$ for (1.1) through (1.3).*

Proof: We define the operator A by

$$A\begin{pmatrix} p \\ q \end{pmatrix} = \begin{pmatrix} \nu p_{xx} - \eta q_{xx} \\ p_{xx} \end{pmatrix}$$

and show that A with the boundary conditions (1.2) is an infinitesimal generator of a compact analytic semigroup in $L^2(0,1)$. It should be mentioned that since the boundary conditions are not typical, they cause some difficulty in estimating the eigenvalues and the resolvent. Since σ is a nonlinear functions of q, the semigroup in L^2 is not enough. Nevertheless, it is possible to show that if the initial data are in $H^1(0,1)$ and satisfy the boundary conditions, then for the following iteration

$$\begin{aligned} p_t^{(n+1)} &= \nu p_{xx}^{(n+1)} - \eta q_{xx}^{(n+1)} + \sigma(q^{(n)}) - P, \\ q_t^{(n+1)} &= p_{xx}^{(n+1)}, \\ p^{(0)} = f(x), &\quad q^{(0)} = g(x), \end{aligned}$$

there is a contraction mapping in $H^1(0,1)$. This gives local existence. Now, we use the following equality

$$E(p,q)(t) + \int_0^t \int_0^1 \nu p_{xx}^2(x,s)\, dxds = E(p,q)(0), \tag{2.1}$$

where

$$E(p,q)(t) = \int_0^1 \{\frac{1}{2}p_x^2 + W(q) - Pq + \frac{\eta}{2}q_x^2\}(x,t)\, dx, \tag{2.2}$$

as the *a priori* estimate for the H^1 norm of (p,q) so that the continuation argument is possible, and hence we can show the existence of a global solution.

□

3 Stability and bifurcation analysis.

Here, we discuss the stability of stationary solutions and their bifurcations.

Lemma 3.1 *The constant solutions $(0,\alpha)$, $(0,\beta)$, and $(0,\delta)$ are stationary solutions for all values of $\eta > 0$. Furthermore, their indices are $\mathrm{h}((0,\alpha)) = \mathrm{h}((0,\beta)) = \Sigma^0$. Namely, they are dynamically stable.*

Next, consider the eigenvalue problem corresponding (1.1) and (1.2):

$$\begin{aligned} \eta q_{xx} - (\nu\lambda + \sigma'(\delta))q &= -\lambda p, \\ p_{xx} &= \lambda q. \end{aligned} \tag{3.1}$$

Lemma 3.2 *The eigenvalues of (3.1) cross the origin from left to right of the imaginary axis at* $\eta = -\sigma'(\delta)/(n\pi)^2$ *as we decrease* η. *Furthermore zero eigenvalues are simple.*

Lemma 3.3 *If* $\sigma \in C^3$, $\sigma'(\delta) < 0$, *and* $\sigma'''(\delta) > 0$, *then there is a supercritical pitchfork bifurcation at* $\eta = -\sigma'(\delta)/(n\pi)^2$. *Furthermore, if* $\sigma(u+\delta)/u < \sigma'(\delta)$ *for* $\alpha - \delta \leq u \leq \beta - \delta$ *except at* $u = 0$, *then there is no secondary bifurcation along the non-constant stationary solutions.*

Theorem 3.4 *There exists a global compact attractor* $\mathbf{A}$ *for (1.1) and (1.2).*

The above lemmas and the theorem imply

Lemma 3.5 *If* $\eta > -\sigma'(\delta)/\pi^2$, *the stationary solution* $(0, \delta)$ *has one dimensional unstable manifold or equivalentely the index is* $\mathrm{h}((0,\delta)) = \Sigma^1$.

Combining the above lemmas and the theorem we have

Theorem 3.6 *If* $\sigma \in C^3$, $\sigma'(\delta) < 0$, *and* $\sigma'''(\delta) > 0$, *then the following holds:*

(i) For $-\sigma'(\delta)/(n\pi)^2 < \eta < -\sigma'(\delta)/((n-1)\pi)^2$, $(0,\delta)$ *is a nondegenerate stationary solution and has an n-dimensional unstable manifold.*

(ii) If $M(k^{\pm})$ *denote the non-constant stationary solutions which arise from the bifurcation point* $\eta = -\sigma'(\delta)/(k\pi)^2$, *then* $M(k^{\pm})$ *are non-degenerate and have k-dimensional unstable manifolds.*

4 Connecting orbit problems.

We now discuss the connecting orbit problem in the semiflow. In the semiflow the connecting orbit means the solutions connecting two stationary solutions, namely,

$$\lim_{t\to-\infty}(p,q)(t) = \text{a stationary solution},$$
$$\lim_{t\to\infty}(p,q)(t) = \text{another stationary solution}.$$

To simplify the notation let $M(0^+) = (0,\alpha)$, $M(0^-) = (0,\beta)$, and $M(n) = (0,\delta)$. Then, we have

Theorem 4.1 *Given a collection* $\{j^*, j+1^*, j+2^*, ..., j+r^* \mid * = +\text{or}-\}$ *and* $\epsilon > 0$, *there exists a solution* $(p(t), q(t))$ *of (1.1) and (1.2) and a sequence* $t_1 > t_2 > ... > t_{r-1}$ *such that*

$$\lim_{t\to-\infty}(p(t),q(t)) = M(j+r^*), \quad \lim_{t\to\infty}(p(t),q(t)) = M(j^*), \tag{4.1}$$

and

$$d(M(j+i^*), (p(t_i), q(t_i))) < \epsilon. \tag{4.2}$$

Furthermore,

$$cl(C(M(j+r^*), M(j^*))) \cap cl(C(M(j+i^*), M(j+s^*))) \neq \emptyset, \text{ for } 0 \leq s \leq r.$$

This theorem establishes that there is always a connecting orbit from a stationary solution with higher dimensional unstable manifold to a stationary solution with lower one. To prove this we apply the connection matrix to the global compact attractor A whose homology index is Σ^0. When there are $(2n+1)$ stationary solutions, we can show that the connection matrix is given by

$$\begin{array}{c} \\ M(0^*) \\ M(1^*) \\ M(2^*) \\ \vdots \\ M(n) \end{array} \begin{array}{c} \begin{array}{ccccc} M(0^*) & M(1^*) & M(2^*)\cdots & & M(n) \end{array} \\ \begin{pmatrix} 0 & B_1 & 0 & \cdots & 0 \\ & 0 & B_2 & \cdots & \vdots \\ & & 0 & & \\ & & & & B_n \\ & & & & 0 \end{pmatrix} \end{array},$$

where

$$B_k = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad k = 1, \cdots, (n-1)$$

and

$$B_n = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

5 Slow motions

In Carr and Pego [4] they have show that if $A = B$ in Fig. 1.1, then for

$$u_t = \eta u_{xx} - \sigma(u) + P \tag{5.1}$$

there exist slowly varying solutions which are actually exponentially slow. Since our system is similar to 5.1), we expect that there are slowly varying solutions.

In this section we choose P so that areas A and B in Fig. 1.1 are equal. We shall showa numerical comparison of (p, q) in (1.1) and (p_o, q_o) satisfying

$$\begin{aligned} \nu q_{ot} &= \eta q_{oxx} - \sigma(q_o) + P, \\ p_o &= \int_1^x \int_0^s q_{ot}(r,t)\, dr ds \end{aligned} \tag{5.2}$$

with the boundary conditions for q_o given by (1.2b). Notice that p_o satisfies the boundary conditions (1.2a).

First we derive (5.2). If the motion is slow, it should reflect as small values in p_t. Therefore, for an approximate solution we drop p_t term from the first equation in (1.1) and use $q_t = p_{xx}$ to obtain the first equation of (5.2). This equation is a well known semilinear parabolic equation with bistable nonlinearity for which the dynamics are well understood. Then, from $q_t = p_{xx}$ we obtain the second equation of (5.2). We denote the solution to

(5.2) by (p_o, q_o) and study how the difference $(p - p_o, q - q_o)$ behaves. Set $\bar{p} = p - p_o$ and $\bar{q} = q - q_o$. Then, $(\bar{p}, \bar{q})$ satisfies

$$\begin{aligned} \bar{p}_t &= \nu \bar{p}_{xx} - \eta \bar{q}_{xx} + \sigma(\bar{q} + q_o) - \sigma(q_o) - \frac{\partial}{\partial t} \int_1^x \int_0^s q_{ot}\, dr ds, \\ \bar{q}_t &= \bar{p}_{xx}. \end{aligned} \tag{5.3}$$

Since q_{ot} is small for the slow motions, this encourages the numerical comparison of (p, q) and (p_o, q_o).

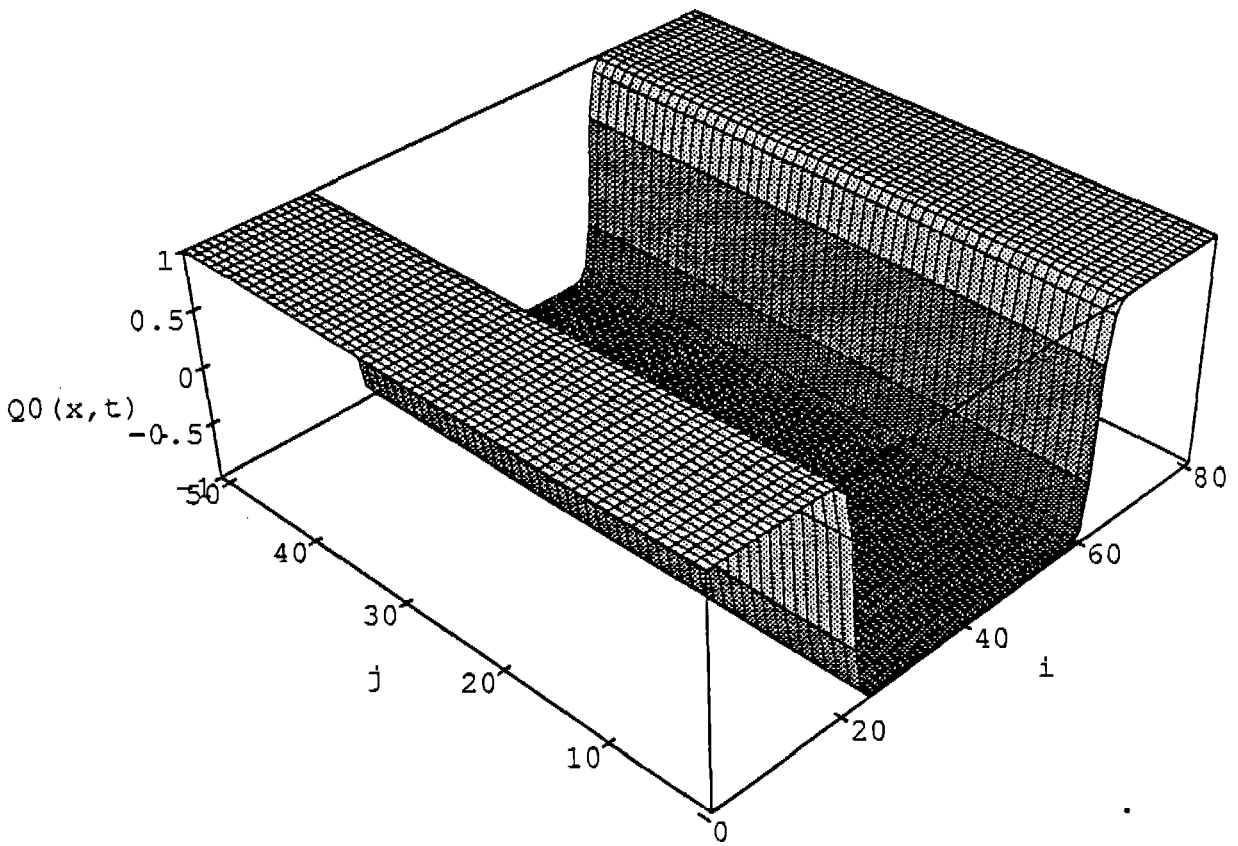


Figure 5.1 The values of q_o.

We now report the preliminary numerical results. We take $\sigma(q) = q^3 - q$, $P = 0$, $\nu = 1.0$, and $\eta = 0.0001$. In Figure 5.1 the values of $q_o(x,t)$ with the initial data $q_o(x,0) = 0.1\cos(2\pi x)$ are given. In Figure 5.2 the values of q with

$$(p,q)(x,0) = (\int_1^x \int_0^s q_o(r,0)_t \, dr ds, q_o(x,0))$$

are given. In these figures i and j denote x and t variables, respectively. The x variable ranges from 0 to 80 and the t variable ranges from 100 to 150. The difference between q and q_o is given in Figure 5.3. It is interesting to see that the difference is very small. Further details of computation will appear in a future publication.

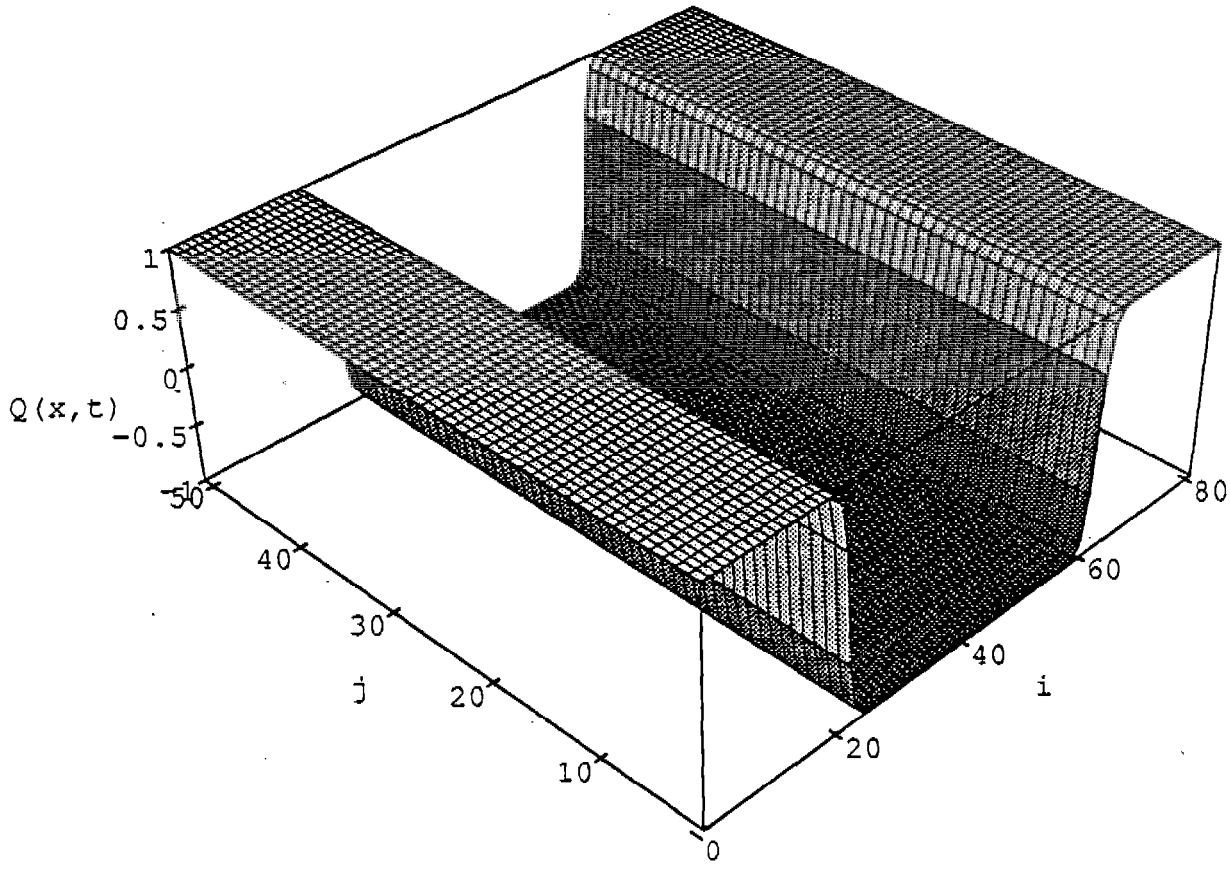


Figure 5.2 The values of q.

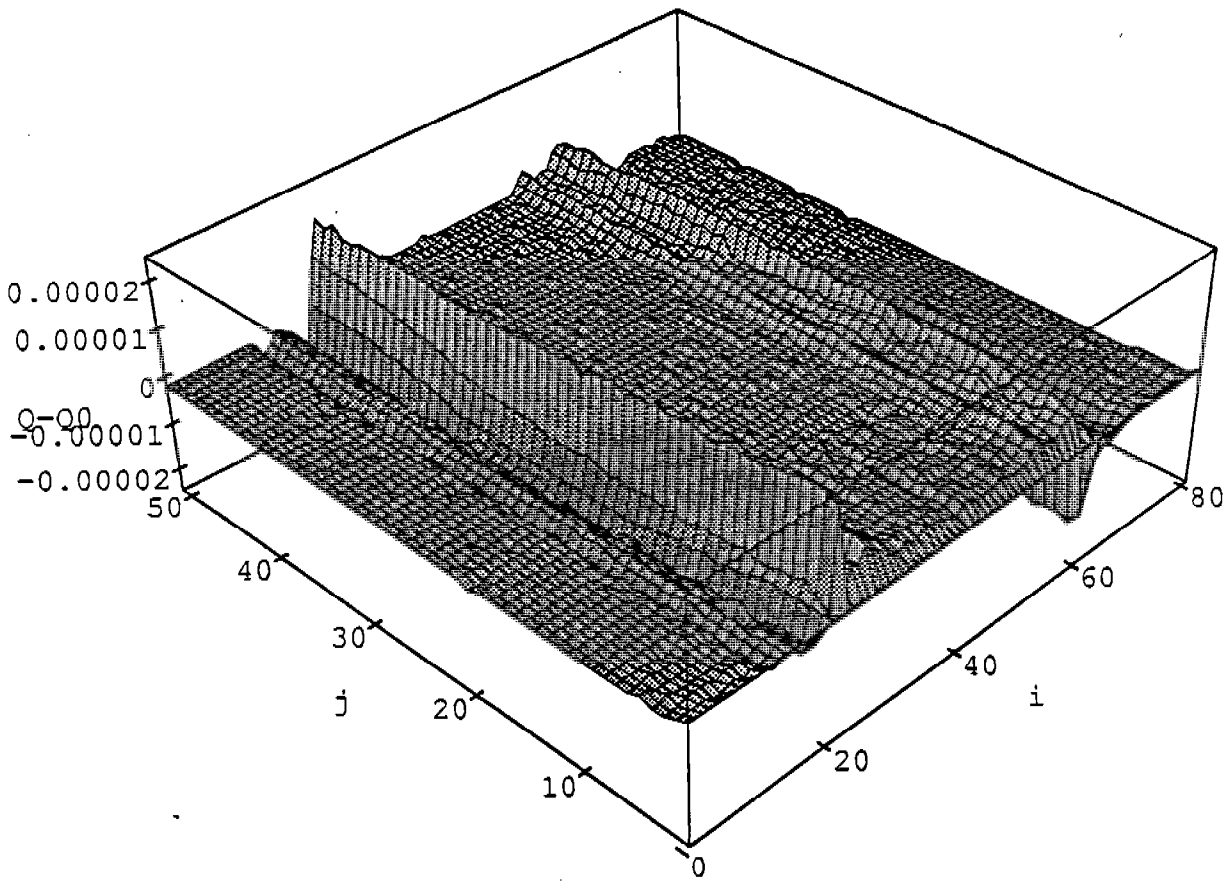


Figure 5.3 The difference between q and q_o.

References

[1] Andrews, G. and J.M. Ball, Asymptotic behaviour and change of phase in one-dimensional nonlinear viscoelasticity, J. Diff. Eqns. 44 (1982), 306-341.

[2] Carr, J., M.E. Gurtin, and M. Slemrod, One dimensional structured phase transitions under prescribed loads, J. Elasticity 15 (1985), 133-142.

[3] Carr, J., M.E. Gurtin, and M. Slemrod, Structured phase transitions on a finite interval, Arch. Rat. Mec. Anal. 86 (1984), 317-351.

[4] Carr, J. and R.L. Pego, Metastable patterns in solutions of $u_t = \epsilon^2 u_{xx} - f(u)$, Comm. Pure Appl. Math. 42 (1989), 523-576.

[5] Hattori, H. and C. Mischaikow, A dynamical systems approach to a phase transition problem, to appear in J. Diff. Eqns.

[6] Korteweg, Sur la forme que prennent les équations des mouvement des fluides si l'on tient comple des forces capillaires par des variations de densite, Arch. Neerl. Sci. Exactes Nat. Ser. II 6 (1901), 1-24.

[7] Serrin, J., Phase transition and interfacial layers for van der Waals fluids, in "Proceedings of SAFA IV Conference, Recent Methods in Nonlinear Analysis and Applications, Naples, 1980" (A. Camfora, S. Rionero, C. Sbordone, C. Trombetti, Eds.)

[8] Slemrod, M., Admissiblity criteria for propagating phase boundaries in a van der Waals fluid, Arch. Rat. Mech. Anal. 81 (1983), 301-315.

Energy Dissipation in an Elastic Material Containing a Mobile Phase Boundary Subjected to Concurrent Dynamic Pulses[1]

Jiehliang Lin and Thomas J. Pence
Department of Metallurgy, Mechanics and Materials Science,
Michigan State University,
East Lansing, MI 48824-1226

Abstract: We consider the energetic behavior of a phase boundary that is subjected to concurrent dynamic pulses, one from each side, in the event that the phase boundary motion is maximally dissipative. The total energy loss is contrasted to that which would occur if the two pulses were not concurrent.

1. Introduction

Stress-induced phase transitions in solids can be modeled in a continuum elasticity setting by means of strain energy density functions that are not convex. In one spatial dimension this gives rise to stress-strain laws that are not monotonic [1975E]. Using this framework, it has been shown that an acoustic pulse impinging on a pre-existing stationary phase boundary within such a material gives rise to both a reflected pulse and a transmitted pulse [1991P]. A notable feature of the purely elastic theory is that it alone is not sufficient to determine the outcome of such a pulse/phase boundary encounter, allowing instead for a family of possible solutions; in fact this family can be parametrized by the speed at which the phase boundary moves during the encounter. This state of affairs, however, can be rendered well-posed by augmenting the theory with an additional criterion specific to phase boundary motion which has the effect of singling out one member of the family of possible solutions. These include criteria which capture kinetic effects [1987T], [1990G], [1991A], impedance effects [1991P], dissipative effects [1980J], [1986H], [1991PP], or other phenomena not accounted for by the purely elastic theory [1983H], [1991T].

Understanding the large-time asymptotic dynamics of any such process is complicated by the geometric increase in the number of pulses with time due to the spawning of both a reflected and a transmitted pulse at each pulse/phase boundary encounter of the reverberation process [1991L]. In addition, any such process will in general eventually give rise to a situation in which pulses impinge on the phase boundary from both the front and the back. Our purpose here is to consider this concurrent pulse problem. In the next section we state the problem and display the family of solutions as parametrized by the phase boundary speed. Then, following [1991PP], we determine in Sections 3 and 4 the particular solution that is maximally dissipative. We then consider the following question:

How does the total energy loss for a concurrent pulse problem governed by a maximum dissipation rate criterion (M.D.C.) compare to the combined energy loss for two subsidiary problems: one involving only the pulse which impinges from the front (governed by M.D.C.), and the other involving only the pulse which impinges from the back (also governed by M.D.C.)?

1. Supported by the U.S. Army Research Office under contract DAAL03-89-G-0089.

In Sections 5 and 6 we show that the former is greater than the latter in the event that both incoming pulses are of the same sign (with respect to the ambient), but that the latter is greater than the former in the event that they are of opposite signs.

2. Families of Solutions to the Concurrent Pulse Problem

Let τ, γ and v denote respectively stress, strain and particle velocity. Following [1991P], we consider a layer, 0<x<h, composed of an elastic material whose stress-strain behavior in one dimension is given by

$$\tau = \hat{\tau}(\gamma) \equiv \left\{ \begin{array}{ll} c^2\gamma & \text{for } 0 \le \gamma \le \gamma_M \\ \hat{\tau}_u(\gamma) & \text{for } \gamma_M \le \gamma \le \gamma_m \\ c^2\gamma + d & \text{for } \gamma \ge \gamma_m \end{array} \right\}, \qquad \hat{\tau}(-\gamma) = -\hat{\tau}(\gamma), \tag{2.1}$$

where c and d are constants and $\hat{\tau}_u(\gamma)$ is a smooth decreasing function that renders $\hat{\tau}(\gamma)$ continuous. The layer is assumed to be initially pre-stressed in equilibrium, so that $v=0$, with a single phase boundary at $x=s_o$ separating high strain phase with $\gamma=\gamma_b$ in $x<s_o$ from low strain phase with $\gamma=\gamma_a$ in $x>s_o$. The strain values γ_a and γ_b are taken to be the well known Maxwell strains which have the geometrical interpretation of cutting off equal areas on the stress strain curve (Fig. 1). An immediate consequence is that the initial configuration is one of minimum energy [1975E] for the prevail-

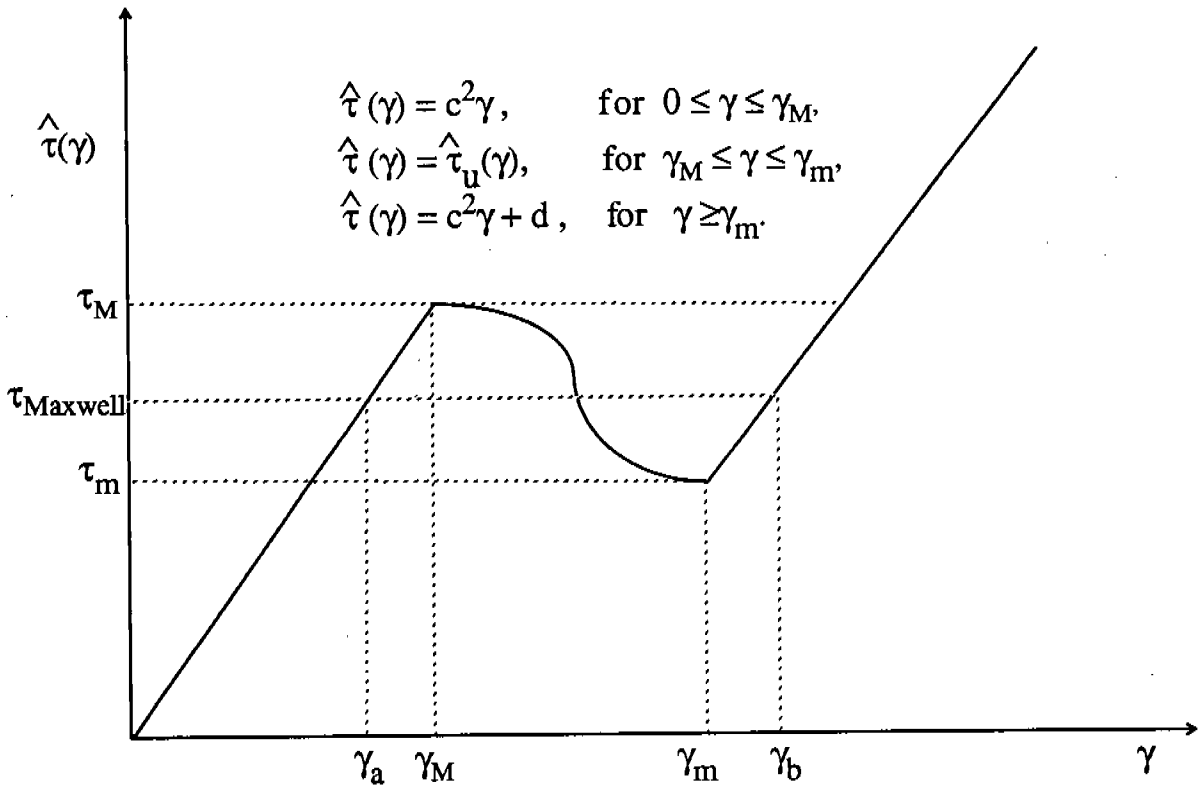


Fig. 1. Stress-strain contstitutive response as described by (2.1). The descending portion of such a constitutive response function is associated with unstable material behavior [1975E]. The strain intervals: $[-\gamma_M,\gamma_M]$, $[\gamma_M,\gamma_m]$, $[\gamma_m, \infty)$, correspond respectively to a low strain phase, an unstable phase and a high strain phase.

ing boundary conditions governing the initial equilibrium configuration. Any subsequent change in the boundary conditions will give rise to changes in the strain and velocity fields governed by the equations[2]

$$v_x - \gamma_t = 0, \qquad \hat{\tau}'(\gamma)\gamma_x - v_t = 0. \tag{2.2}$$

In particular, we consider a loading condition at x=0 that gives rise to a square wave pulse with strain $\gamma_b+\Delta\gamma_1$ over a time interval t_b, and a loading condition at x=h that gives rise to a square wave pulse with strain $\gamma_a+\Delta\gamma_2$ over a time interval t_a. We shall not concern ourselves with the specific loading conditions needed to generate these pulses, nor with restrictions upon $\Delta\gamma_1$ and $\Delta\gamma_2$ necessary to ensure compatibility with (2.1) other than to note that these issues can be treated in a systematic fashion [1991P]. According to (2.2), each pulse will travel toward the phase boundary with speed c; furthermore the right moving pulse has width ct_b and particle velocity given by $-c\Delta\gamma_1$, while the left moving pulse has width ct_a and particle velocity given by $c\Delta\gamma_2$. The encounter of such a right moving pulse with the phase boundary is treated in [1991P] on the assumption that the encounter ends before the arrival of any pulse from the other side. Our purpose here, however, is to study such a concurrent encounter. There are four generic cases: (*rr*), (*rl*), (*lr*), (*ll*), where (*rr*) denotes the case where the *right moving* pulse (with strain increment $\Delta\gamma_1$) encounters the phase boundary first and also terminates last, (*rl*) denotes the case where the *right moving* pulse encounters the phase boundary first, but the encounter with the *left moving* pulse terminates last, and the remaining two cases are defined accordingly. For the remainder of this section, and also for Section 3, we shall restrict attention to the (*rl*) case. There are then three distinct interaction periods: Π_1 in which only the right moving pulse encounters the phase boundary, Π_{con} in which both pulses encounter the phase boundary concurrently, and Π_2 in which only the left moving pulse encounters the phase boundary. Figure 2 diagrams these encounters in the (x,t)-plane. According to this figure, the following additional assumptions are also implicit in our treatment: (A1) the phase boundary remains at rest unless acted on by a pulse, (A2) phase transitions take place only by movement of the pre-existing phase boundary, and (A3) the phase boundary velocity is constant during each of the three interaction periods and these three phase boundary velocities obey

$$-c < \dot{s}_1 < c, \qquad -c < \dot{s}_{con} < c, \qquad -c < \dot{s}_2 < c. \tag{2.3}$$

Further discussion of these issues can be found in [1991P]. In addition we have depicted the phase boundary as coming to rest after the complete encounter has ended, in which case the fields return to their initial conditions on each side of the since displaced phase boundary.

In Figure 2, the subscripts $T1$ and $R1$ denote the fields in the transmitted and reflected pulses associated with interaction period Π_1. In addition, the (x,t)-domain with combined incoming and reflected pulse during the interaction period Π_1 is denoted by subscript $S1$. A similar convention is followed for subscripts $T2$, $R2$, and $S2$ for the pulses associated with interaction period Π_2. The fields associated with the combination of $T1$ and the incoming pulse characterized by $\Delta\gamma_2$ is denoted by subscript $T1i2$. Finally, there are four additional (x,t)-domains associated with pulses

2. Primes and subscripts denote differentiation in the usual fashion. Note also that we have taken the density to be equal to one in $(2.2)_2$.

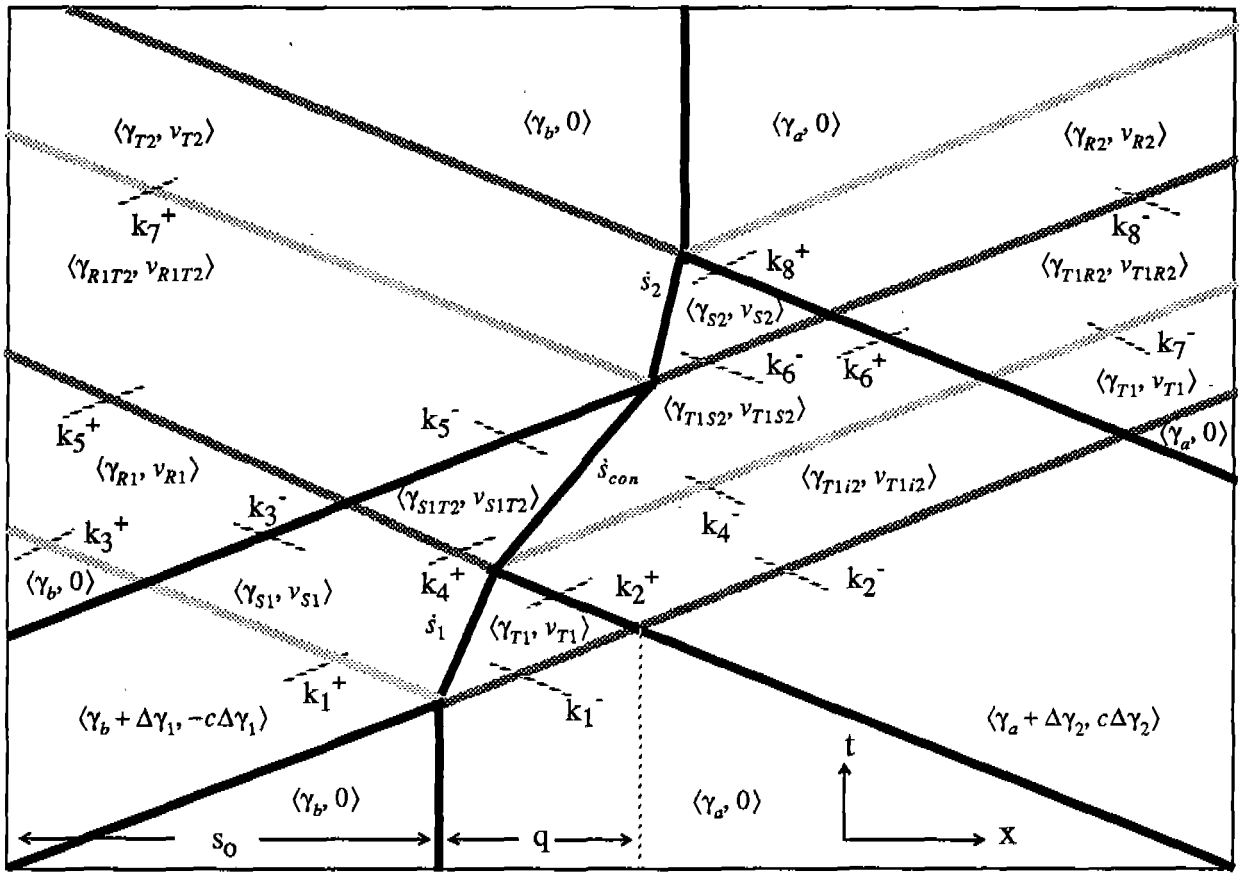


Fig. 2. Concurrent encounter of a right moving shear pulse, a left moving shear pulse and a phase boundary. The shear strain γ and velocity v in these incoming pulses and the generated pulses are denoted by $\langle\gamma,v\rangle$. The characteristic curves are indicated by dashed line segments.

that arise as a consequence of the concurrent interaction period Π_{con}; these are denoted by the subscripts *S1T2*, *R1T2*, *T1S2*, and *T1R2*. A consequence of (A3), (2.1) and (2.2) is that the value of strain and particle velocity are individually constant on the individual (x,t)-domains associated with the 11 symbols *T1*, *R1*, *S1*, *T1i2*, *T2*, *R2*, *S2*, *S1T2*, *R1T2*, *T1S2*, and *T1R2*. The corresponding 22 unknown values for strain and velocity, in conjunction with the three unknown phase boundary velocities $\dot{s}_1$, $\dot{s}_{con}$, and $\dot{s}_2$, comprise the unknown quantities in the complete encounter problem. Relations connecting these 25 unknown values to the parameters c, γ_b, γ_a, $\Delta\gamma_1$, $\Delta\gamma_2$ which characterize the material, initial conditions, and loading conditions follow from the theory of Riemann invariants as applied to (2.1), (2.2). In particular, this gives that v-$c\gamma$ is constant on any line segment with slope $\frac{dx}{dt}=c$ in the (x,t)-plane provided that it does not cross the phase boundary. Similarly $v+c\gamma$ is constant on all line segments with slope $\frac{dx}{dt}=-c$ that do not cross the phase boundary. Each of these Riemann invariant conditions generates 8 algebraic equations relating $\{\gamma,v\}$ pairs between contiguous (x,t)-domains; the associated connecting line segments are denoted by $K_1^+,...K_8^+$ and $K_1^-,...,K_8^-$ in Figure 2. Across the phase boundary, the two Rankine-Hugoniot conditions

$$[\![v]\!] = -\dot{s}[\![\gamma]\!], \qquad [\![\tau]\!] = -\dot{s}[\![v]\!], \tag{2.4}$$

associated with (2.1), (2.2) are required to hold. These give rise to an additional 6 algebraic equations, 2 for each of the 3 interaction periods Π_1, Π_{con}, and Π_2. Thus in total there are 22 equations relating the 25 unknown values. Regarding the three phase boundary velocities as parameters, the 22 equations are linear in the 22 strain and particle velocities. The resulting 22x22 coefficient matrix is nonsingular provided that none of the three phase boundary velocities $\dot{s}_1$, $\dot{s}_{con}$, and $\dot{s}_2$, take on the values c or $-c$. Hence (2.3) ensures that the system can be inverted. Certain simplifications are achieved in the resulting problem due to various uncouplings (i.e. zero blocks in the coefficient matrix). For example $\{\gamma_{S1},v_{S1}\}$ and $\{\gamma_{T1},v_{T1}\}$ can be found from the 2 Riemann invariant conditions associated with K_I^+ and K_I^-, along with the 2 Rankine-Hugoniot conditions associated with interaction period Π_1. The resulting 22 field quantities are thus found to be given by:

$$
\begin{aligned}
\gamma_{S1} &= \gamma_b + \Delta\gamma_1 - \frac{(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1+c)}, & v_{S1} &= -c\Delta\gamma_1 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1+c)},\\
\gamma_{R1} &= \gamma_b - \frac{(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1+c)}, & v_{R1} &= -\frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1+c)},\\
\gamma_{T1} &= \gamma_a + \Delta\gamma_1 + \frac{(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1-c)}, & v_{T1} &= -c\Delta\gamma_1 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1-c)},\\
\gamma_{T1i2} &= \gamma_a + \Delta\gamma_1 + \Delta\gamma_2 + \frac{(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1-c)}, & v_{T1i2} &= -c\Delta\gamma_1 + c\Delta\gamma_2 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_1}{2\,(\dot{s}_1-c)}\\
\gamma_{S1T2} &= \gamma_b + \Delta\gamma_1 + \Delta\gamma_2 - \frac{(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}+c)}, & v_{S1T2} &= -c\Delta\gamma_1 + c\Delta\gamma_2 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}+c)},\\
\gamma_{T1S2} &= \gamma_a + \Delta\gamma_1 + \Delta\gamma_2 + \frac{(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}-c)}, & v_{T1S2} &= -c\Delta\gamma_1 + c\Delta\gamma_2 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}-c)},\\
\gamma_{R1T2} &= \gamma_b + \Delta\gamma_2 - \frac{(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}+c)}, & v_{R1T2} &= c\Delta\gamma_2 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}+c)},\\
\gamma_{T1R2} &= \gamma_a + \Delta\gamma_1 + \frac{(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}-c)}, & v_{T1R2} &= -c\Delta\gamma_1 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_{con}}{2\,(\dot{s}_{con}-c)},\\
\gamma_{S2} &= \gamma_a + \Delta\gamma_2 + \frac{(\gamma_b-\gamma_a)\,\dot{s}_2}{2\,(\dot{s}_2-c)}, & v_{S2} &= c\Delta\gamma_2 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_2}{2\,(\dot{s}_2-c)},\\
\gamma_{R2} &= \gamma_a + \frac{(\gamma_b-\gamma_a)\,\dot{s}_2}{2\,(\dot{s}_2-c)}, & v_{R2} &= -\frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_2}{2\,(\dot{s}_2-c)},\\
\gamma_{T2} &= \gamma_b + \Delta\gamma_2 - \frac{(\gamma_b-\gamma_a)\,\dot{s}_2}{2\,(\dot{s}_2+c)}, & v_{T2} &= c\Delta\gamma_2 - \frac{c\,(\gamma_b-\gamma_a)\,\dot{s}_2}{2\,(\dot{s}_2+c)}.
\end{aligned}
\tag{2.5}
$$

The phase boundary velocities $\dot{s}_1$, $\dot{s}_{con}$, and $\dot{s}_2$, are undetermined by the above procedure and can be regarded as parametrizing all possible solutions. In addition to (2.3), various additional restrictions upon the phase boundary velocities will arise due to the requirement that the strain values in (2.5) remain confined to the intervals associated with the different branches of the stress-strain relation (2.1). The net effect of these considerations is to generate additional inequality constraints beyond (2.3) on the phase boundary velocities. The totality of inequality constraints are not mutually exclusive provided that $\Delta\gamma_1$ and $\Delta\gamma_2$ are sufficiently small. If, however, $\Delta\gamma_1$ and $\Delta\gamma_2$ are large,

then mutual exclusivity may prevail (see [1991P]). We shall henceforth assume that we are dealing with values of $\Delta\gamma_1$ and $\Delta\gamma_2$ which do not give rise to mutual exclusivity so that three non-empty parametrization intervals, $\mathfrak{I}_1$, $\mathfrak{I}_{con}$, and $\mathfrak{I}_2$, exist for the three phase boundary velocities.

3. Maximally Dissipative Solutions for the (*rl*)-case

As mentioned in the Introduction, the freedom to determine the phase boundary velocity allows the theory to accommodate additional requirements upon conditions which govern phase boundary motion. We shall in what follows examine one possible operative condition, namely a *maximum dissipation rate criterion* (M.D.C.). As is well known, the motion of a phase boundary gives rise to a change in the total mechanical energy stored in the mechanical fields [1980J]. In particular, if $\gamma^+=\gamma(s(t)+,t)$ and $\gamma^-=\gamma(s(t)-,t)$ are the strains directly adjacent to the phase boundary, then the energy loss rate, or dissipation rate, is given by

$$D(t) = \dot{s}(t)\left(\int_{\gamma^-}^{\gamma^+}\hat{\tau}(\gamma)\,d\gamma-\frac{1}{2}\left(\hat{\tau}(\gamma^+)+\hat{\tau}(\gamma^-)\right)(\gamma^+-\gamma^-)\right). \tag{3.1}$$

For the concurrent pulse problem, with strains as given in (2.5), one finds that the dissipation rate during the three interaction periods are given by

$$\begin{aligned} D_1 &= -\frac{1}{2\dot{s}_1}\{c^2(c^2-\dot{s}_1^2)\{(\gamma_{T1}-\gamma_a)^2-(\gamma_{S1}-\gamma_b)^2\}\},\\ D_{con} &= -\frac{1}{2\dot{s}_{con}}\{c^2(c^2-\dot{s}_{con}^2)\{(\gamma_{T1S2}-\gamma_a)^2-(\gamma_{S1T2}-\gamma_b)^2\}\},\\ D_2 &= -\frac{1}{2\dot{s}_2}\{c^2(c^2-\dot{s}_2^2)\{(\gamma_{S2}-\gamma_a)^2-(\gamma_{T2}-\gamma_b)^2\}\}, \end{aligned} \tag{3.2}$$

where use has been made of the special equal area property of the Maxwell strains γ_a and γ_b which characterize the initial configuration.

The maximum dissipation rate criterion is equivalent to the entropy rate admissibility criterion of Dafermos [1973D], which, in the present setting, selects solutions which have the property of maximizing $D(t)$. Hence, entering (3.2) with (2.5) and differentiating with respect to the appropriate phase boundary velocity, one obtains the following implicit equations for $\dot{s}_1$, $\dot{s}_{con}$, and $\dot{s}_2$:

$$\Delta\gamma_1 = \frac{(\gamma_b-\gamma_a)\,c^3\dot{s}_1}{(\dot{s}_1^2-c^2)^2}, \qquad \Delta\gamma_1+\Delta\gamma_2 = \frac{(\gamma_b-\gamma_a)\,c^3\dot{s}_{con}}{(\dot{s}_{con}^2-c^2)^2}, \qquad \Delta\gamma_2 = \frac{(\gamma_b-\gamma_a)\,c^3\dot{s}_2}{(\dot{s}_2^2-c^2)^2}. \tag{3.3}$$

Each of the equations (3.3) admits a unique solution obeying (2.3) which we shall denote by $\dot{s}_1^{(md)}$, $\dot{s}_{con}^{(md)}$, and $\dot{s}_2^{(md)}$. It can be shown that these values indeed maximize D_1, D_{con}, and D_2 with respect to all $\dot{s}_1$, $\dot{s}_{con}$, and $\dot{s}_2$ obeying (2.3). Even though it may or may not be the case that $\dot{s}_1^{(md)} \in \mathfrak{I}_1$, $\dot{s}_{con}^{(md)} \in \mathfrak{I}_{con}$, and $\dot{s}_2^{(md)} \in \mathfrak{I}_2$; we shall assume that these inclusions hold for the remainder of this communication.[3] We now introduce normalized phase boundary velocities and pulse

3. These inclusions will hold if both $\Delta\gamma_1$ and $\Delta\gamma_2$ are sufficiently small [1991PP].

strain increments as follows:

$$\dot{\tilde{s}} = \frac{\dot{s}}{c}, \qquad \Delta\tilde{\gamma} = \frac{\Delta\gamma}{(\gamma_b - \gamma_a)}, \tag{3.4}$$

where subscripted and superscripted quantities such as $\dot{\tilde{s}}_{con}^{(md)}$ are defined in the obvious fashion by these same normalizations. Then the maximum values for D_1, D_{con}, and D_2, which will be denoted by $D_1^{(md)}$, $D_{con}^{(md)}$, and $D_2^{(md)}$, are given by

$$D_1^{(md)} = c^3(\gamma_b - \gamma_a)^2 \chi(\dot{\tilde{s}}_1^{(md)}), \qquad D_{con}^{(md)} = c^3(\gamma_b - \gamma_a)^2 \chi(\dot{\tilde{s}}_{con}^{(md)}), \qquad D_2^{(md)} = c^3(\gamma_b - \gamma_a)^2 \chi(\dot{\tilde{s}}_2^{(md)}), \tag{3.5}$$

$$\chi(\dot{\tilde{s}}) \equiv \frac{\dot{\tilde{s}}^2(1+\dot{\tilde{s}}^2)}{2(1-\dot{\tilde{s}}^2)^2}, \tag{3.6}$$

where

$$\dot{\tilde{s}}_1^{(md)} = \dot{\tilde{S}}(\Delta\tilde{\gamma}_1), \qquad \dot{\tilde{s}}_{con}^{(md)} = \dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2), \qquad \dot{\tilde{s}}_2^{(md)} = \dot{\tilde{S}}(\Delta\tilde{\gamma}_2), \tag{3.7}$$

and $\dot{\tilde{S}}(\Delta\tilde{\gamma})$ is defined for all real $\Delta\tilde{\gamma}$ as the unique root, within the interval $-1 < \dot{\tilde{S}} = \dot{\tilde{S}}(\Delta\tilde{\gamma}) < 1$, to the equation

$$\dot{\tilde{S}}^4 - 2\dot{\tilde{S}}^2 - \frac{1}{\Delta\tilde{\gamma}}\dot{\tilde{S}} + 1 = 0, \qquad ((\Delta\tilde{\gamma} = 0) \Rightarrow (\dot{\tilde{S}} = 0)). \tag{3.8}$$

4. Maximally Dissipative Solutions for the Concurrent Pulse in General

The (rr), (lr), and (ll)-cases can be treated in a similar fashion. In all cases, formulae $(3.5)_2$ and $(3.7)_2$ hold during the genuinely concurrent part of the encounter. If and when a portion of the encounter only involves the right moving pulse with strain increment $\Delta\gamma_1$ then $(3.5)_1$ and $(3.7)_1$ hold, whereas $(3.5)_3$ and $(3.7)_3$ govern those portions of any encounters that involve only the left moving pulse with strain increment $\Delta\gamma_2$. In order to determine which of the four possible cases is that which occurs, let

$$q = \frac{h}{2} - s_o. \tag{4.1}$$

Then one finds that the four cases occur according to

$$\begin{aligned}
&(rl): \quad q > 0, \quad ct_a(c+\dot{s}_1^{(md)})(c-\dot{s}_{con}^{(md)}) - ct_b(c+\dot{s}_1^{(md)})(c+\dot{s}_{con}^{(md)}) + 2q(c-\dot{s}_1^{(md)})(c+\dot{s}_{con}^{(md)}) > 0,\\
&(rr): \quad q > 0, \quad ct_a(c+\dot{s}_1^{(md)})(c-\dot{s}_{con}^{(md)}) - ct_b(c+\dot{s}_1^{(md)})(c+\dot{s}_{con}^{(md)}) + 2q(c-\dot{s}_1^{(md)})(c+\dot{s}_{con}^{(md)}) < 0,\\
&(lr): \quad q < 0, \quad ct_a(c-\dot{s}_2^{(md)})(c-\dot{s}_{con}^{(md)}) - ct_b(c-\dot{s}_2^{(md)})(c+\dot{s}_{con}^{(md)}) + 2q(c+\dot{s}_2^{(md)})(c-\dot{s}_{con}^{(md)}) < 0,\\
&(ll): \quad q < 0, \quad ct_a(c-\dot{s}_2^{(md)})(c-\dot{s}_{con}^{(md)}) - ct_b(c-\dot{s}_2^{(md)})(c+\dot{s}_{con}^{(md)}) + 2q(c+\dot{s}_2^{(md)})(c-\dot{s}_{con}^{(md)}) > 0.
\end{aligned} \tag{4.2}$$

5. Energy Loss for the Maximally Dissipative Solution in the (*rl*)-case

In this section we begin the examination of the question raised in the Introduction. As shown in [1991L], this particular question arises in the study of the large time asymptotic dynamics of an acoustic reverberation process in which all interactions are governed by the maximum dissipation rate criterion (M.D.C.). For the (*rl*) case discussed in Sections 2 and 3, let $t_1^{(md)}$, $t_{con}^{(md)}$, and $t_2^{(md)}$ denote the time duration of the encounters associated with the interaction periods Π_1, Π_{con}, and Π_2. These quantities are given in terms of $\dot{s}_1^{(md)}$, $\dot{s}_{con}^{(md)}$, and $\dot{s}_2^{(md)}$ as

$$t_1^{(md)} = \frac{2q}{c+\dot{s}_1^{(md)}}, \qquad t_{con}^{(md)} = \frac{ct_b(c+\dot{s}_1^{(md)}) - 2q(c-\dot{s}_1^{(md)})}{(c+\dot{s}_1^{(md)})(c-\dot{s}_{con}^{(md)})},$$
$$t_2^{(md)} = \frac{ct_a(c+\dot{s}_1^{(md)})(c-\dot{s}_{con}^{(md)}) - ct_b(c+\dot{s}_1^{(md)})(c+\dot{s}_{con}^{(md)}) + 2q(c-\dot{s}_1^{(md)})(c+\dot{s}_{con}^{(md)})}{(c+\dot{s}_1^{(md)})(c-\dot{s}_{con}^{(md)})(c+\dot{s}_2^{(md)})}, \tag{5.1}$$

so that the total energy loss for the complete encounter process governed by (M.D.C.) is

$$\Delta E^{(md)} = D_1^{(md)} t_1^{(md)} + D_{con}^{(md)} t_{con}^{(md)} + D_2^{(md)} t_2^{(md)}. \tag{5.2}$$

We now turn to consider the energy loss which would accompany two subsidiary problems.

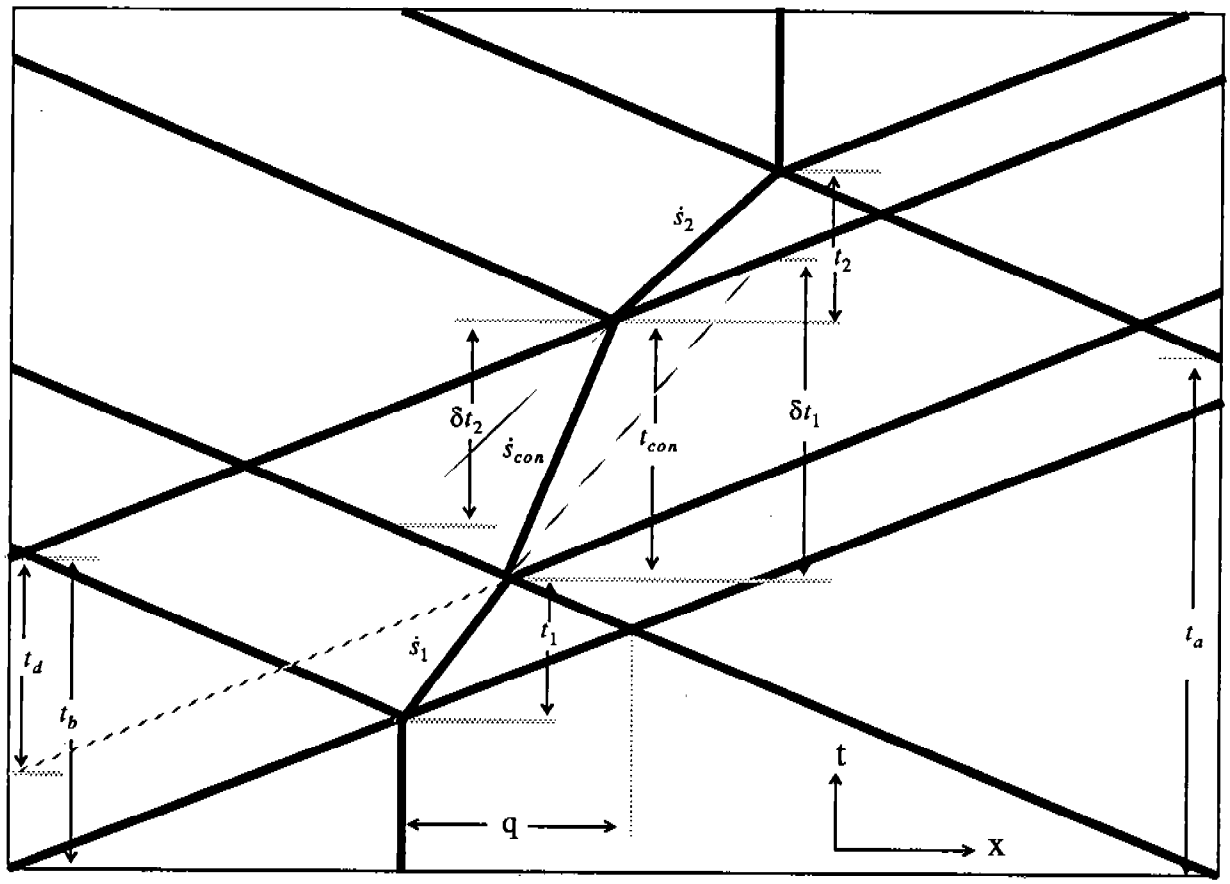


Fig. 3. The (*rl*) concurrent encounter.

The first problem is that in which only the right moving pulse, associated with strain increment $\Delta\gamma_1$, impinges upon the phase boundary. The encounter dynamics are again assumed to be governed by (M.D.C.). The phase boundary velocity and dissipation rate for this problem can simply be found by setting $\Delta\gamma_2=0$ in the previous development; consequently they are given by $\dot{s}_1^{(md)}$ and $D_1^{(md)}$. Similarly, the second problem is that in which only the left moving pulse, associated with strain increment $\Delta\gamma_2$, impinges upon the phase boundary with encounter dynamics governed by (M.D.C.). Hence the phase boundary velocity and dissipation rate in this problem are given by $\dot{s}_2^{(md)}$ and $D_2^{(md)}$. It is, however, important to note that the time duration of the encounters are *not* given by $t_1^{(md)}$ and $t_2^{(md)}$, but rather are each of a longer duration due to the additional interaction time which was taken by the concurrent pulse in the original problem. These additional interaction times will be denoted respectively by $\delta t_1^{(md)}$ and $\delta t_2^{(md)}$ and are given by (see Figure 3):

$$\begin{aligned}\delta t_1^{(md)} &= \frac{c t_b\,(c+\dot{s}_1^{(md)}) - 2q\,(c-\dot{s}_1^{(md)})}{(c+\dot{s}_1^{(md)})\,(c-\dot{s}_1^{(md)})},\\ \delta t_2^{(md)} &= \frac{c t_b\,(c+\dot{s}_1^{(md)})\,(c+\dot{s}_{con}^{(md)}) - 2q\,(c-\dot{s}_1^{(md)})\,(c+\dot{s}_{con}^{(md)})}{(c+\dot{s}_1^{(md)})\,(c-\dot{s}_{con}^{(md)})\,(c+\dot{s}_2^{(md)})},\end{aligned} \tag{5.3}$$

so that the total energy losses in the two subsidiary problems are given by

$$\Delta E^{(md)}_{\Delta\gamma_1 \text{only}} = D_1^{(md)}\,(t_1^{(md)} + \delta t_1^{(md)}), \qquad \Delta E^{(md)}_{\Delta\gamma_2 \text{only}} = D_2^{(md)}\,(t_2^{(md)} + \delta t_2^{(md)}). \tag{5.4}$$

Consequently, the difference in the energy loss between the original problem and the combined energy loss for the two subsidiary problems, is given by

$$\Upsilon_{(rl)} \equiv \Delta E^{(md)} - (\Delta E^{(md)}_{\Delta\gamma_1 \text{only}} + \Delta E^{(md)}_{\Delta\gamma_2 \text{only}}) = D_{con}^{(md)}\,t_{con}^{(md)} - D_1^{(md)}\,\delta t_1^{(md)} - D_2^{(md)}\,\delta t_2^{(md)}. \tag{5.5}$$

In order to develop a simple expression for $\Upsilon_{(rl)}$ it is convenient to introduce

$$t_d \equiv t_b - \frac{2q\,(c-\dot{s}_1^{(md)})}{c\,(c+\dot{s}_1^{(md)})} > 0, \tag{5.6}$$

where $t_d > 0$ follows either from $t_{con}^{(md)} > 0$ or else from its interpretation as a 'projected time' given in Figure 3. Substituting from (5.1), (5.3) into (5.5) and using (3.5), (3.6), (5.6) yields

$$\Upsilon_{(rl)} = \frac{1}{2} c^3 t_d (\gamma_b - \gamma_a)^2 \Phi(\dot{\bar{s}}_1^{(md)}, \dot{\bar{s}}_{con}^{(md)}, \dot{\bar{s}}_2^{(md)}), \tag{5.7}$$

where

$$\Phi(\dot{\bar{s}}_1^{(md)}, \dot{\bar{s}}_{con}^{(md)}, \dot{\bar{s}}_2^{(md)}) = \frac{\dot{\bar{s}}_{con}^{(md)2}\,(1+\dot{\bar{s}}_{con}^{(md)2})}{(1-\dot{\bar{s}}_{con}^{(md)2})^2\,(1-\dot{\bar{s}}_{con}^{(md)})} - \frac{\dot{\bar{s}}_1^{(md)2}\,(1+\dot{\bar{s}}_1^{(md)2})}{(1-\dot{\bar{s}}_1^{(md)2})^2\,(1-\dot{\bar{s}}_1^{(md)})} - \frac{\dot{\bar{s}}_2^{(md)2}\,(1+\dot{\bar{s}}_2^{(md)2})\,(1+\dot{\bar{s}}_{con}^{(md)})}{(1-\dot{\bar{s}}_2^{(md)2})^2\,(1+\dot{\bar{s}}_2^{(md)})\,(1-\dot{\bar{s}}_{con}^{(md)})}. \tag{5.8}$$

In view of (3.7) we define

$$\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) \equiv \Phi(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1), \dot{\tilde{S}}(\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2), \dot{\tilde{S}}(\Delta\tilde{\gamma}_2)) = \Phi(\dot{\bar{s}}_1^{(md)}, \dot{\bar{s}}_{con}^{(md)}, \dot{\bar{s}}_2^{(md)}). \tag{5.9}$$

Thus the question posed in the Introduction reduces, in the (rl)-case, to a determination of the sign of $\hat{\Phi}$. Now

$$\hat{\Phi}(0,\Delta\tilde{\gamma}_2) = 0, \qquad \hat{\Phi}(\Delta\tilde{\gamma}_1,0) = 0, \tag{5.10}$$

where, for example, $(5.10)_1$ follows from (5.8) since $\Delta\tilde{\gamma}_1 = 0$ implies $\dot{\tilde{s}}_1^{(md)} = 0$, $\dot{\tilde{s}}_{con}^{(md)} = \dot{\tilde{s}}_2^{(md)}$; while a similar argument gives $(5.10)_2$. Let partial derivatives of $\hat{\Phi}$ be denoted by numerical subscripts, e.g. $\hat{\Phi}_1(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2) = \frac{\partial}{\partial\Delta\tilde{\gamma}_1}\hat{\Phi}(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2)$, then (5.10) gives $\hat{\Phi}(0,0) = 0$, $\hat{\Phi}_1(0,0) = 0$, $\hat{\Phi}_2(0,0) = 0$, $\hat{\Phi}_{11}(0,0) = 0$, $\hat{\Phi}_{22}(0,0) = 0$, while (3.7), (3.8), (5.8), (5.9) gives $\hat{\Phi}_{12}(0,0) = 2$, so that the origin is neither a maximum nor a minimum and

$$\hat{\Phi}(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2) = 2\Delta\tilde{\gamma}_1\Delta\tilde{\gamma}_2 + O((\Delta\tilde{\gamma})^3). \tag{5.11}$$

We have numerically calculated $\hat{\Phi}(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2)$ for various pairs $(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2)$ and display the results in tabular form and also in the contour plot of Figure 4.

These numerical results suggest that $\hat{\Phi} > 0$ in the first and third quadrants, whereas $\hat{\Phi} < 0$ in the second and fourth quadrants. The final task of this section will be to show that this is indeed the case. To this end, we obtain by virtue of (3.7), (3.8) and (5.8) that

$$\hat{\Phi}(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2) = \Delta\tilde{\gamma}_1\hat{W}_1(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2) + \Delta\tilde{\gamma}_2\hat{W}_2(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2), \tag{5.12}$$

where

$$\hat{W}_1(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2) = \bar{W}_1(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1),\dot{\tilde{S}}(\Delta\tilde{\gamma}_1+\Delta\tilde{\gamma}_2)), \qquad \hat{W}_2(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2) = \bar{W}_2(\dot{\tilde{S}}(\Delta\tilde{\gamma}_1+\Delta\tilde{\gamma}_2),\dot{\tilde{S}}(\Delta\tilde{\gamma}_2)), \tag{5.13}$$

Table. Values of $\hat{\Phi}(\Delta\tilde{\gamma}_1,\Delta\tilde{\gamma}_2)$ near the origin.

$\Delta\tilde{\gamma}_2$ \ $\Delta\tilde{\gamma}_1$	-1.000	-0.800	-0.600	-0.400	-0.200	0.000	0.200	0.400	0.600	0.800	1.000
1.000	-0.878	-0.910	-0.851	-0.670	-0.377	0.000	0.412	0.817	1.200	1.558	1.895
0.800	-0.624	-0.635	-0.620	-0.516	-0.302	0.000	0.343	0.685	1.007	1.310	1.594
0.600	-0.437	-0.426	-0.410	-0.359	-0.223	0.000	0.269	0.541	0.799	1.041	1.268
0.400	-0.278	-0.265	-0.246	-0.215	-0.142	0.000	0.188	0.384	0.571	0.746	0.910
0.200	-0.133	-0.126	-0.115	-0.097	-0.065	0.000	0.098	0.206	0.311	0.408	0.500
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
-0.200	0.116	0.107	0.095	0.078	0.051	0.000	-0.095	-0.231	-0.377	-0.517	-0.647
-0.400	0.210	0.192	0.169	0.136	0.086	0.000	-0.158	-0.418	-0.751	-1.095	-1.419
-0.600	0.285	0.258	0.224	0.178	0.110	0.000	-0.194	-0.515	-0.988	-1.569	-2.162
-0.800	0.346	0.311	0.267	0.209	0.127	0.000	-0.215	-0.562	-1.076	-1.793	-2.669
-1.000	0.396	0.354	0.301	0.233	0.140	0.000	-0.229	-0.590	-1.110	-1.830	-2.818

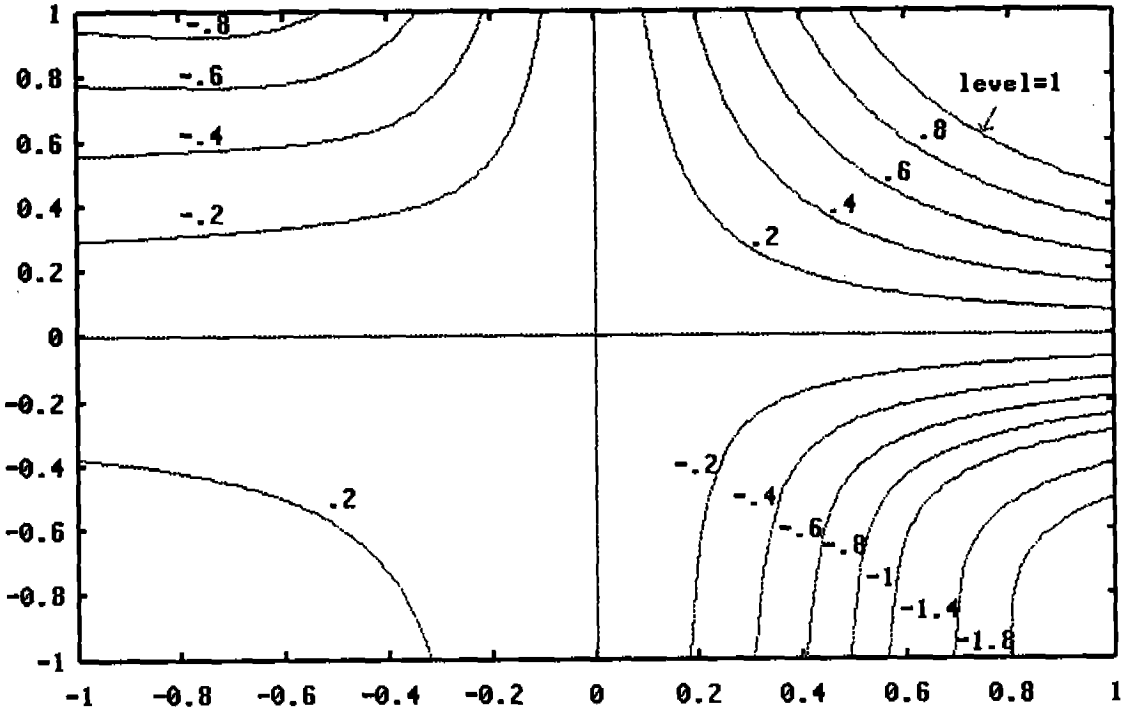


Fig. 4. Contour plot of $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ near the origin.

with

$$W_1(\overset{\sim(md)}{s_1}, \overset{\sim(md)}{s_{con}}) = \frac{\overset{\sim(md)}{s_{con}}(1+\overset{\sim(md)^2}{s_{con}})}{(1-\overset{\sim(md)}{s_{con}})} - \frac{\overset{\sim(md)}{s_1}(1+\overset{\sim(md)^2}{s_1})}{(1-\overset{\sim(md)}{s_1})},$$

$$W_2(\overset{\sim(md)}{s_{con}}, \overset{\sim(md)}{s_2}) = \frac{\overset{\sim(md)}{s_{con}}(1+\overset{\sim(md)^2}{s_{con}})}{(1-\overset{\sim(md)}{s_{con}})} - \frac{\overset{\sim(md)}{s_2}(1+\overset{\sim(md)^2}{s_2})(1+\overset{\sim(md)}{s_{con}})}{(1+\overset{\sim(md)}{s_2})(1-\overset{\sim(md)}{s_{con}})}. \tag{5.14}$$

It may be verified from (5.14) that

$$\begin{pmatrix} W_1(\overset{\sim md}{s_1}, \overset{\sim md}{s_{con}}) < 0, & -1 < \overset{\sim md}{s_{con}} < \overset{\sim md}{s_1} < 1, \\ W_1(\overset{\sim md}{s_1}, \overset{\sim md}{s_{con}}) = 0, & -1 < \overset{\sim md}{s_{con}} = \overset{\sim md}{s_1} < 1, \\ W_1(\overset{\sim md}{s_1}, \overset{\sim md}{s_{con}}) > 0, & -1 < \overset{\sim md}{s_1} < \overset{\sim md}{s_{con}} < 1, \end{pmatrix}, \quad \begin{pmatrix} W_2(\overset{\sim md}{s_{con}}, \overset{\sim md}{s_2}) < 0, & -1 < \overset{\sim md}{s_{con}} < \overset{\sim md}{s_2} < 1, \\ W_2(\overset{\sim md}{s_{con}}, \overset{\sim md}{s_2}) = 0, & -1 < \overset{\sim md}{s_{con}} = \overset{\sim md}{s_2} < 1, \\ W_2(\overset{\sim md}{s_{con}}, \overset{\sim md}{s_2}) > 0, & -1 < \overset{\sim md}{s_2} < \overset{\sim md}{s_{con}} < 1, \end{pmatrix}, \tag{5.15}$$

which in view of (3.7) and (3.8) gives

$$\begin{pmatrix} \hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0, & \text{if } \Delta\tilde{\gamma}_2 < 0, \\ \hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = 0, & \text{if } \Delta\tilde{\gamma}_2 = 0, \\ \hat{W}_1(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0, & \text{if } \Delta\tilde{\gamma}_2 > 0, \end{pmatrix}, \quad \begin{pmatrix} \hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) < 0, & \text{if } \Delta\tilde{\gamma}_1 < 0, \\ \hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = 0, & \text{if } \Delta\tilde{\gamma}_1 = 0, \\ \hat{W}_2(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) > 0, & \text{if } \Delta\tilde{\gamma}_1 > 0, \end{pmatrix}. \tag{5.16}$$

Hence (5.12) and (5.16) yield

$$\begin{aligned}
\hat{\Phi}(\Delta\bar{\gamma}_1, \Delta\tilde{\gamma}_2) &< 0, \quad \text{if } \Delta\bar{\gamma}_1 < 0,\ \Delta\tilde{\gamma}_2 > 0 \text{ or if } \Delta\gamma_1 > 0,\ \Delta\gamma_2 < 0, \\
\hat{\Phi}(\Delta\bar{\gamma}_1, \Delta\tilde{\gamma}_2) &= 0, \quad \text{if } \Delta\tilde{\gamma}_1 = 0 \text{ or if } \Delta\gamma_2 = 0, \\
\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) &> 0, \quad \text{if } \Delta\tilde{\gamma}_1 < 0,\ \Delta\tilde{\gamma}_2 < 0 \text{ or if } \Delta\gamma_1 > 0,\ \Delta\gamma_2 > 0.
\end{aligned} \tag{5.17}$$

6. Energy Loss for the Maximally Dissipative Concurrent Pulse in General

The energy losses for the *(rr)*, *(lr)*, and *(ll)*-cases can be determined in a corresponding way. For the *(rr)*-case one finds that $\Upsilon_{(rr)}$, the energy loss difference analogous to $\Upsilon_{(rl)}$, is given by

$$\Upsilon_{(rr)} = D_{con}^{(md)} t_{con}^{(md)} - D_1^{(md)} \Omega t_1^{(md)} - D_2^{(md)} \Omega t_2^{(md)}, \tag{6.1}$$

where $\Omega t_1^{(md)}$ and $\Omega t_2^{(md)}$ are displayed in Figure 5. We then find that

$$\Upsilon_{(rr)} = \frac{1}{2} c^3 t_a (\gamma_b - \gamma_a)^2 \Psi(\overset{\cdot(md)}{s_1}, \overset{\cdot(md)}{s_{con}}, \overset{\cdot(md)}{s_2}), \tag{6.2}$$

where

$$\Psi(\overset{\cdot(md)}{s_1}, \overset{\cdot(md)}{s_{con}}, \overset{\cdot(md)}{s_2}) = \Phi(-\overset{\cdot(md)}{s_2}, -\overset{\cdot(md)}{s_{con}}, -\overset{\cdot(md)}{s_1}). \tag{6.3}$$

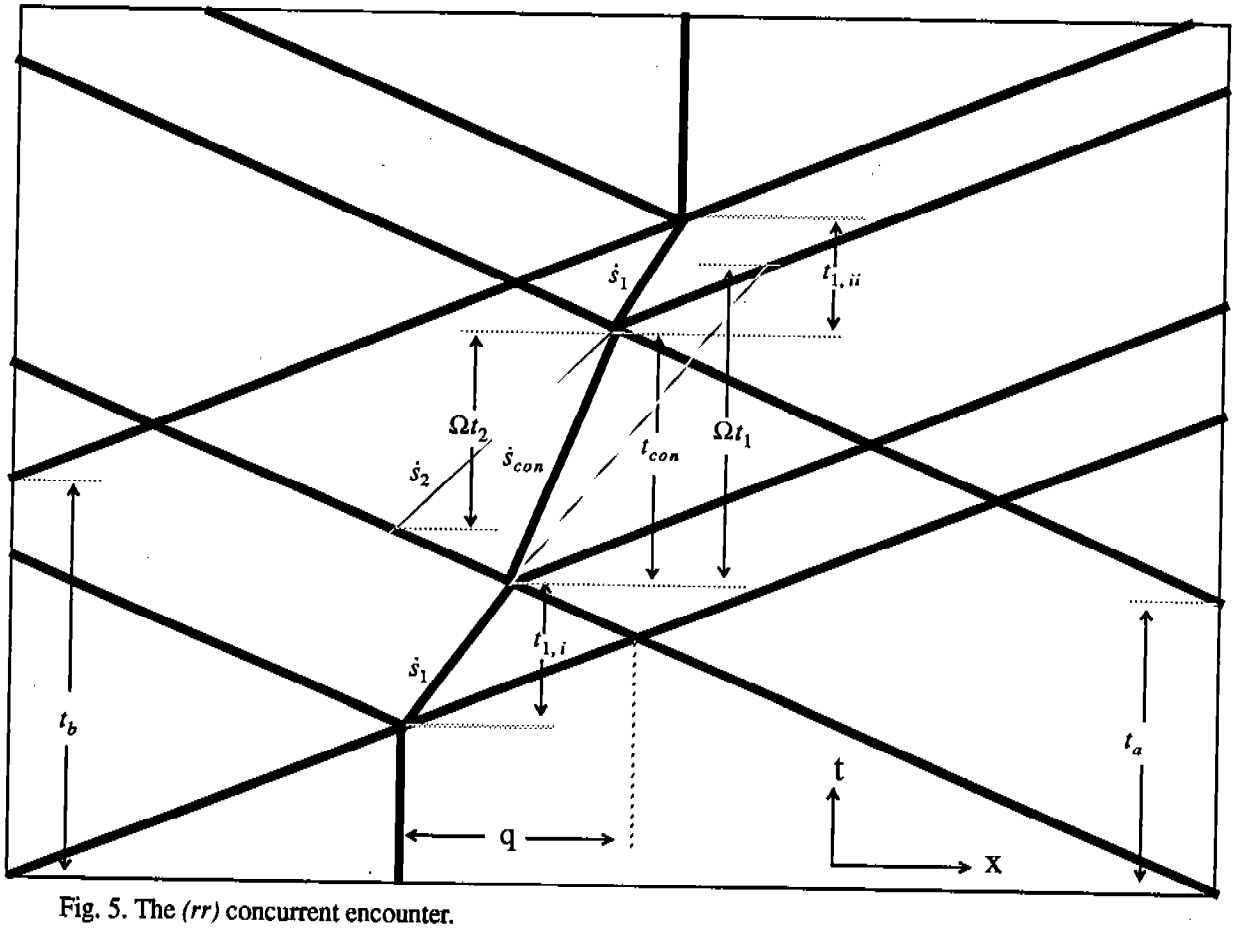


Fig. 5. The *(rr)* concurrent encounter.

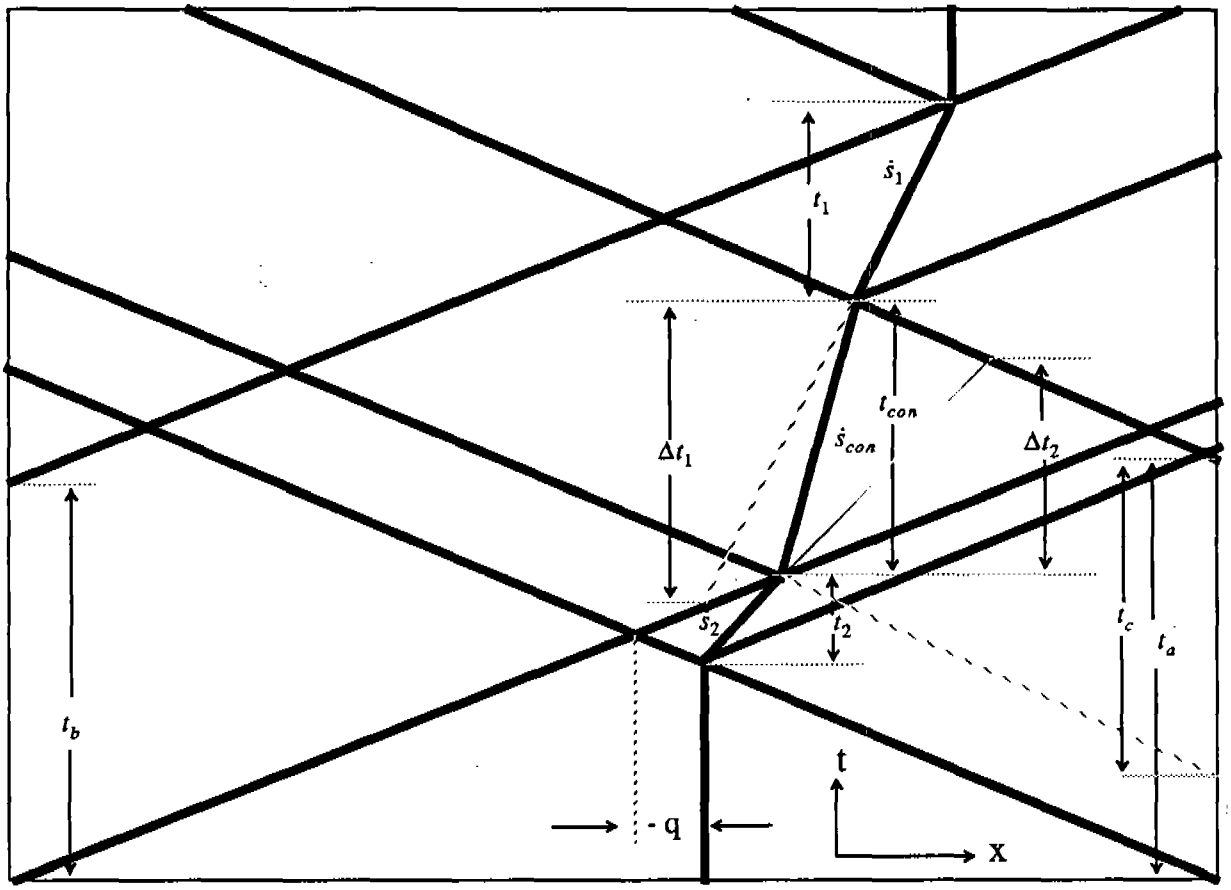


Fig. 6. The *(lr)* concurrent encounter.

Hence defining $\hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ analogous to $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ and using (6.3), (3.7), (3.8) gives

$$\hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2) = \hat{\Phi}(-\Delta\tilde{\gamma}_2, -\Delta\tilde{\gamma}_1) \tag{6.4}$$

indicating that reflection symmetry about the diagonal line $\Delta\tilde{\gamma}_1 + \Delta\tilde{\gamma}_2 = 0$ transforms the contour plot in Figure 4 for $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ into a contour plot for $\hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$. In particular, $\hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ and $\hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2)$ are each positive in the first and third quadrants, and are each negative in the second and fourth quadrants. With this background, we now summarize our findings for all of the concurrent pulse cases:

$$\begin{aligned}
(rl): &\quad \Upsilon_{(rl)} = \tfrac{1}{2}c^3 t_d (\gamma_b - \gamma_a)^2 \hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2), \\
(rr): &\quad \Upsilon_{(rr)} = \tfrac{1}{2}c^3 t_a (\gamma_b - \gamma_a)^2 \hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2), \\
(lr): &\quad \Upsilon_{(lr)} = \tfrac{1}{2}c^3 t_c (\gamma_b - \gamma_a)^2 \hat{\Psi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2), \\
(ll): &\quad \Upsilon_{(ll)} = \tfrac{1}{2}c^3 t_b (\gamma_b - \gamma_a)^2 \hat{\Phi}(\Delta\tilde{\gamma}_1, \Delta\tilde{\gamma}_2).
\end{aligned} \tag{6.5}$$

In the (*lr*)-case, another projected time $t_c \equiv t_a + \dfrac{2q\,(c + \dot{s}_2^{(md)})}{c\,(c - \dot{s}_2^{(md)})}$ has been introduced (Figure 6).

Thus we conclude that *the concurrent pulse encounter suffers the greater energy loss in the event that both incoming pulses are of the same sign, whereas the concurrent pulse encounter suffers the lesser energy loss in the event that the incoming pulses are of opposite sign.*

References

[1973D] Dafermos, C.M., The entropy rate admissibility criterion for solutions of hyperbolic conservation laws, *J. Diff. Eqs.* **14**, 202-212.

[1975E] Ericksen,J.L., Equilibrium of bars, *J. Elasticity* **5**, 191-201.

[1980J] James, R.D., The propagation of phase boundaries in elastic bars, *Arch. Rational Mech. Anal.* **73**, 125-158.

[1983H] Hagan, R. and M. Slemrod, The viscosity-capillarity criterion for shocks and phase transitions, *Arch. Rational Mech. Anal.* **83**, 333-361.

[1986H] Hattori, H., The Riemann problem for a van der Waals fluid with entropy rate admissibility criterion-isothermal case, *Arch. Rational Mech. Anal.* **92**, 247-263.

[1987T] Truskinovsky, L., Dynamics of nonequilibrium phase boundaries in a heat conducting nonlinear elastic medium, *J. Appl. Math. Mech.* (PMM) **51** (1987) 777-784.

[1990G] Gurtin, M.E. and A. Struthers, Multiphase thermomechanics with interfacial structure 3. evolving phase boundaries in the presence of bulk deformation, *Arch. Rational Mech. Anal.* **112**, 97-160.

[1991A] Abeyaratne, R. and J.K. Knowles, Kinetic relations and the propagation of phase boundaries in solids, *Arch. Rational Mech. Anal.* **114**, 119-154.

[1991L] Lin, J. and T.J. Pence, On the energy dissipation due to wave ringing in non-elliptic elastic materials, in review.

[1991P] Pence, T.J., On the encounter of an acoustic shear pulse with a phase boundary in an elastic material: reflection and transmission behavior, *J. Elasticity* **25**, 31-74.

[1991PP] Pence, T.J., On the encounter of an acoustic shear pulse with a phase boundary in an elastic material: energy and dissipation, *J. Elasticity* **26**, 95-146.

[1991T] Truskinovsky, L., Kinks vs. shocks, to appear in *Shock Induced Transitions and Phase Structures in General Media* (R. Fosdick, E. Dunn and M. Slemrod, eds.) Springer-Verlag.

A Unified Representation for Some Combinatorial Optimization Problems

Wing Shing Wong
AT&T Bell Laboratories
Holmdel, NJ, 07733

Abstract

In this short note, we list a number of combinatorial optimization problems, among them the Traveling Salesman Problem and the Graph Partitioning Problem, that can be represented by a common matrix formulation. This formulation was used previously by Brockett to study certain geometric matching problems. Although this unified representation does not necessarily imply the existence of a unified efficient algorithm to solve all these problems, it may provide useful insights for a better understanding of the structure of these problems.

1. Introduction

Several recent papers ([1-3]) have investigated the idea of using gradient flows on $\mathbf{SO(n)}$, the special orthogonal group, to provide a new mechanism for solving certain combinatorial optimization problems. The cost functions for these optimization problems can be formulated in the following forms:

$$tr C^T \Theta \tag{1}$$

$$tr C^T \Theta^T S \Theta \tag{2}$$

As an interesting side observation, note that the set of functions of type (1) or (2) are pivotal elements in the representation theory of $\mathbf{SO(n)}$.

In this short note, we list some combinatorial optimization problems that can be formulated by such cost functions. The list includes the Assignment Problem (AP), the Traveling Salesman Problem (TSP), the Graph Partitioning Problem and some routing optimization problems. Although this unified representation does not necessarily imply the existence of a unified efficient algorithm to solve all these problems, it may provide useful

insights for a better understanding of the structure of these problems. In Section 5, we summarize some results of embedding the AP and the TSP in $\mathbf{SO(n)}$. The results provide a new perspective on some local search techniques as applied to these problems.

2. The AP, TSP, and Extended Traveling Salesman Problems

The Assignment Problem is a well known combinatorial optimization problem with polynomial time complexity (see [4, 5] for more information.) It can be formulated in the matrix form as:

$$\min_{P \in \mathbf{P(n)}} tr\ C^T P\ , \tag{3}$$

where $\mathbf{P(n)}$ denote the set of n by n permutation matrices. There is an interesting connection between the Assignment Problem and the Geometric Matching Problem as was pointed out in [3].

The Traveling Salesman Problem (TSP) is an *NP-hard* problem. The problem can be formulated as:

$$\min_{P \in \mathbf{T(n)}} tr\ C^T P\ , \tag{4}$$

where $\mathbf{T(n)}$ stands for the subset of $\mathbf{P(n)}$ consisting of irreducible matrices, that is, matrices with no non-trivial invariant subspaces. Elements in $\mathbf{T(n)}$ are also known as directed tours, with the interpretation that P_{ij} equals to 1 if and only if there is a directed arc from node i to node j.

Define S_n to be an n by n matrix of the form:

$$\begin{bmatrix} 0, & 1, & 0, & \cdots & 0 \\ 0, & 0, & 1, & \cdots & 0 \\ 0, & 0, & 0, & \cdots & 0 \\ & \cdots & \cdots & & \\ & \cdots & \cdots & & \\ 0, & 0, & 0, & \cdots & 1 \\ 1, & 0, & 0, & \cdots & 0 \end{bmatrix}$$

It is easy to verify that $P^T S_n P$ is an irreducible permutation matrix if P is any permutation matrix. If the cycle $(i_1 \cdots i_n)$ denotes the order in which the nodes are visited by a tour, T, define P by

$$P_{j,k} = \begin{cases} 1 & \text{if } k = i_j \\ 0 & \text{otherwise} \end{cases}$$

Then $P^T SP$ is the matrix representation of T. Hence, if we let $S = S_n$, then the TSP can be reformulated as:

$$\min_{P \epsilon \mathbf{P(n)}} tr\ C^T P^T SP \tag{5}$$

There are many combinatorial optimization problems that can be considered as extensions to the TSP. Consider the following example of a Two Traveling Salesmen Problem with $2n$ cities. The problem is similar to the TSP, but the cost function in this case is defined by the total sum of two weighted tours, and the tours are restricted by the condition that each tour visits n cities and every city is visited by one and only one tour. We can formulate this problem similar to the TSP, by using a slightly different S. Define:

$$\hat{S}_{2n} = \begin{bmatrix} S_n & 0_n \\ 0_n & S_n \end{bmatrix}.$$

Let $S = \hat{S}_{2n}$. Then this problem can be formulated as:

$$\min_{P \epsilon \mathbf{P(2n)}} tr\ C^T P^T SP$$

It is clear that one can generalize this concept to other extensions of the TSP.

It is interesting to observe that while the AP which is of polynomial complexity has a linear cost function representation, the TSP and other $NP-hard$ problems require a second order representation.

3. Graph Partitioning Problems

Let $G = (V, E)$ be a fully connected graph with n nodes and weighted edges defined by a weight matrix C. Let n_1 and n_2 be two positive integers such that their sum is equal to n. The Generalized Graph Partitioning Problem is to find a partition of V into two subsets with n_1 and n_2 nodes such that the sum of the weights on the cut edges (that is, edges with their endpoints in different subsets of the partition) are minimized. This is a well known $NP-$ *complete* problem with many good heuristic solutions, including the Kernighan-Lin and simulated annealing.

We can formulate this problem in the following matrix formulation. Define $I_{j,k}$ to be an j by k matrix by

$$I_{j,k} = \begin{bmatrix} 1, & 1, & \cdots & 1 \\ 1, & 1, & \cdots & 1 \\ & \cdots & \cdots & \\ 1, & 1, & \cdots & 1 \end{bmatrix}$$

Define

$$S = \begin{bmatrix} 0_{n_1} & , I_{n_1,n_2} \\ I_{n_2,n_1}, & 0_{n_2} \end{bmatrix}$$

It is easy to see that the Graph Partitioning Problem can be formulated as:

$$\min_{P \epsilon \mathbf{P(n)}} tr\ C^T P^T S P$$

4. Optimization Problems in Network Routing

Harshavardhana [6] has shown that for certain optimal node assignment problems associated with *loop-free routing*, the problem can be formulated in the form of (5), where C represents a cost matrix for connecting different nodes in a network and S is the adjacency matrix defining the connectivity graph.

5. Embedding In the Orthogonal Group

Since $\mathbf{P(n)}$ is a subset of $\mathbf{O(n)}$, we can also embed the optimization problems defined by (3) and (5) as optimization problems on $\mathbf{O(n)}$, that is:

$$\min_{P \epsilon \mathbf{O(n)}} tr\ C^T \Theta \ , \tag{6}$$

or

$$\min_{P \epsilon \mathbf{O(n)}} tr\ C^T \Theta^T S_n \Theta \ . \tag{7}$$

Since $\mathbf{O(n)}$ contains elements of the form DP, where D is a diagonal matrix with diagonal values 1 or -1, (we will call such a matrix an H-matrix,) it is clear that the global minimum of (6) and (7) will not be a global minimum in $\mathbf{O(n)}$ in general. Moreover, for the Assignment Problem, the set of critical points of $trC^T\Theta$ is of the form

$$\{\Theta \ \epsilon \ \mathbf{O(n)},\ \Theta \ = \ \Theta^T,\ \Theta C \ = \ C\Theta\} \ .$$

This shows that in general, the global minimum for (3) is not even a local minimum of (5). To remedy this problem, one can reformulate the embedding of AP as:

$$\min_{P \epsilon \mathbf{O(n)}} tr\ \alpha(\Theta) \ = \ \min_{P \epsilon \mathbf{O(n)}} tr\ C^T(\Theta \circ \Theta) \ , \tag{8}$$

where we denote the Schur-Hadamard product between two matrices by $M \circ N \equiv (M_{ij} N_{ij})$. Since $(DP) \circ (DP) \ = \ P \circ P$ for any diagonal matrix D with diagonal value 1 or -1, in this formulation we can restrict the domain of optimization to the connected component of $\mathbf{O(n)}$ with determinant 1, namely, $\mathbf{SO(n)}$. Hence, the problem defined by (8) is equivalent to:

$$\min_{P \epsilon \mathbf{SO(n)}} tr\ \alpha(\Theta) \ . \tag{9}$$

Notice that for the problem defined by (9), finding an optimal point that is defined by an H-matrix will immediately lead to an optimal point defined by a permutation matrix.

Furthermore, this formulation possesses other nice properties. In particular, the following theorem is proven in [3]:

Theorem 1: If Θ is an H-matrix, then it is a critical point of α as a function on $\mathbf{SO(n)}$. Moreover, Θ is a non-degenerate local minimum of α if and only if the permutation matrix that has the same zero entries as Θ is a non-degenerate $2-opt$ solution of the corresponding Assignment Problem. Moreover, there exists an H-matrix that achieves the global minimum value.

For the TSP, one can show that a critical point of $C^T\Theta^T S\Theta$ satisfies

$$C^T\Theta^T S\Theta = \Theta^T S\Theta C^T .$$

This implies that $\Theta C^T\Theta^T$ is a circulant matrix. This ia a very restrictive conclusion. The following is a better way to embed the TSP in $\mathbf{SO(n)}$:

$$\min_{P\epsilon \mathbf{SO(n)}} tr\ \beta(\Theta) = \min_{P\epsilon \mathbf{SO(n)}} tr\ C^T(\Theta\circ\Theta)^T S(\Theta\circ\Theta) . \tag{10}$$

Let T be a tour. We define a neighborhood of T in the following way: If T contains four distinct arcs from node p to q, q to r, p' to q', and q' to r', (see the following figure), then the tour obtained from T by removing these four arcs and joining p to q', q' to r, p' to q, and q to r' is an element in the neighborhood of T. A tour that is locally optimal for the TSP in this definition of a neighborhood is called a *weak 4-opt* solution.

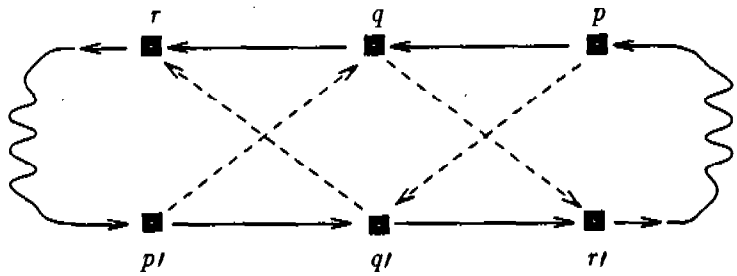
We can prove a result similar to Theorem 1 for the TSP by using the embedding defined in (10). Before stating this result, we observe that there is no loss in generality if we assume for the TSP that the cost function C satisfies the following two properties:

A1. All entries of C are non-negative.

A2. The diagonal elements of C dominate in the sense: $C_{i,i} \geq 2C_{j,k}$ for any i and $j \neq k$.

Then, by a straightforward computation, one can show the following result holds.

Theorem 2: If Θ is an H-matrix, then it is a critical point of β as a function on $\mathbf{SO(n)}$. Moreover, Θ is a non-degenerate local minimum of β if and only if the permutation matrix that has the same zero entries as Θ is a non-degenerate *weak 4-opt* solution of the corresponding Traveling Salesman Problem.



Forming A New Tour

REFERENCES

[1] Brockett, R. W., "Dynamical Systems that Sort Lists and Solve Linear Programming Problems," *Proc. 27th IEEE Conference on Decision and Control*, (1988) 799-803.

[2] Brockett, R. W., "A Geometrical Matching Problem," *J. of Linear Algebra and Its Applications*, 122 (1989) 761-777.

[3] Brockett, R. W., and W. S. Wong, "A Gradient Flow for the Assignment Problem," in *New Trends in Systems Theory*, Birkhauser, Boston, (1991) 170-177.

[4] Spivey, W. A., and R. M. Thrall, *Linear Optimization*, Holt, Rinehart and Winston, New York, NY, (1970).

[5] Burkard, R. E., "Traveling Salesman and Assignment Problems: A Survey," in *Discrete Optimization I*, North-Holland, Amsterdam, (1979).

[6] Harshavardhana, P. "Design and Analysis of Nonhierarchical Node-by-Node Routing Virtual Circuit Networks", *Proc. of IEEE Global Telecommunications Conference*, (1989) 1434-1439.

CLASSIFICATION OF FINITE DIMENSIONAL FILTERS FROM LIE ALGEBRAIC POINT OF VIEW *

Stephen S.-T. Yau

Department of Mathematics, Statistics and Computer Sciences
University of Illinois at Chicago
Box 4348, M/C 249
Chicago, IL. 60680

Abstract

Ever since the technique of the Kalman-Bucy filter was popularized, there has been an intense interest in finding new classes of finite dimensional recursive filters. In the late seventies, the concept of the estimation algebra of a filtering system was introduced. It has proven to be an invaluable tool in the study of nonlinear filtering problems. In 1990, the present author considered a general class of nonlinear filtering systems which include both Kalman-Bucy filtering systems and Benes filtering systems as special cases. A simple algebraic necessary and sufficient condition was established for an estimation algebra of this class of filtering systems to be finite dimensional. Consequently the present author has rigorously constructed a new class of finite dimensional filters which include both Kalman-Bucy filters and Benes filters as special cases. In 1991, Chiou and the present author have shown that the above new class of finite dimensional filters are the most general filters from Lie algebraic point of view.

§1. Introduction

The basic approach to non-linear filtering theory was via " innovation method ", originally proposed by Kailath ca. 1967 and subsequently rigorously developed by Fujisaki, Kallianpur and Kunita [FKK] in 1972. The difficulty with this approach is that the innovations process is not, in general, explicitly computable (except in the well-known Kalman-Bucy case). To circumvent this difficulty, it was independently proposed by Brockett-Clark [*BC*], Brockett [Br_1], Mitter [*Mi*] that the construction of the filter be divided into two parts : (*i*) a universal filter which is the evolution equation describing the unnormalized conditional density, the Duncan-Mortensen-Zakai (D-M-Z) equation and (*ii*) a state-output map, which depends on the statistics to be computed, where the state of the filter is the unnormalized conditional density. Their idea of using estimation algebra to construct finite dimensional nonlinear filters was motivated from the Wei-Norman approach of using Lie algebraic ideas to solve time varying linear differential equations. Let f be the drift term of the filtering system and Ω is the matrix whose (i,j)-entry is $\partial f_j/\partial x_i - \partial f_i/\partial x_j$. Tam, Wong and Yau [$TWY_1$] considered a class of filtering systems having the property that the drift

* Supported by the U.S. Army Research Office.

term f of the state evolution equation is a gradient vector field. In 1990, Yau $[Ya_1, Ya_2]$ considered a class of filtering systems having the property that all the (i,j)-entry of Ω are constants. He derived a single necessary and sufficient condition for an estimation algebra of this general class of filtering system to be finite dimensional. In particular, the Mitter conjecture that for finite dimensional estimation algebra the observation $h(x)$ has to be a degree one polynomial was proven. As an important consequence of these algebraic results, he constructed finite dimensional filters explicitly and rigorously for such a filtering system. Note also that the method used in $[Ya_1]$ computes the fundamental solution of the D-M-Z equation and hence it also solves filtering problem with non-Gaussian initial conditions. Perhaps the break through in the subject is that recently Chiou and Yau proves rigorously that from Lie algebraic point of view the finite dimensional filters constructed by Yau is the most general filter if the state space dimension is not more than two.

§2.The filtering problem considered and the basic concepts

The filtering problem considered here is based on the following observation model :

$$(2.1) \qquad \begin{cases} dx(t) = f(x(t))dt + g(x(t))dv(t) & x(0) = x_0, \\ dy(t) = h(x(t))dt + dw(t) & y(0) = 0, \end{cases}$$

in which x, v, y, and w, are respectively, R^n, R^p, R^m and R^m valued processes, and v and w have components which are independent, standard Brownian processes. We further assume that $n = p, f, h$ are C^∞ smooth, and that g is an orthogonal matrix. We will refer to $x(t)$ as the state of the system at time t and to $y(t)$ as the observation at time t.

Let $\rho(t,x)$ denotes the conditional probability density of the state given the observation $y(s) : 0 \le s \le t$. It is well known (see $[DM]$ for example) that $\rho(t,x)$ is given by normalizing a function, $\sigma(t,x)$, which satisfies the following Duncan-Mortensen-Zakai equation:

$$(2.2) \qquad d\sigma(t,x) = L_0\sigma(t,x)dt + \sum_{i=1}^{m} L_i\sigma(t,x)dy_i(t), \quad \sigma(0,x) = \sigma_0,$$

where $L_0 = \frac{1}{2}\sum_{i=1}^{n} \frac{\partial^2}{\partial x_i^2} - \sum_{i=1}^{n} f_i \frac{\partial}{\partial x_i} - \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} - \frac{1}{2}\sum_{i=1}^{m} h_i^2$ and for $i = 1, \ldots, m$, L_i is the zero degree differential operator of multiplication by h_i. (If ρ is a vector, we use the notation ρ_i to represent the i^{th} component of ρ. σ_0 is probability density of the initial point, x_0).

Equation (2.2) is a stochastic partial differential equation. In real applications, we are interested in constructing robust state estimators from observed sample paths with some property of robustness. Davis $[Da]$ studied this problem and proposed some robust algorithms. In our case, his basic idea reduces to defining a new unnormalized density

$$\xi(t,x) = exp(-\sum_{i=1}^{m} h_i(x)y_i(x))\sigma(t,x)$$

It is easy to show that $\xi(t,x)$ satisfies the following time varying partial differential equation

$$\frac{\partial \xi}{\partial t}(t,x) = L_0\xi(t,x) + \sum_{i=1}^{m} y_i(t)[L_0,L_i]\xi(t,x) + \frac{1}{2}\sum_{i=1}^{m} y_i^2(t)[[L_0,L_i],L_i]\xi(t,x), \tag{2.3}$$
$$\xi(0,x) = \sigma_0$$

where $[\cdot,\cdot]$ is the Lie bracket as described by the following definition.

Definition : Let X and Y are differential operators, the Lie bracket of X and Y , $[X,Y]$, is defined by

$$[X,Y]\xi = X(Y\xi) - Y(X\xi)$$

for any C^∞ function ξ.

The objective of constructing a robust finite-dimensional filter to (2.1) is equivalent to finding a smooth manifold M and complete C^∞ vector fields μ_i on M and C^∞ functions ν on $M \times R \times R^n$ and ω_i's on R^m, such that $\xi(t,x)$ can be represented in the form:

$$\frac{dz}{dt}(t) = \sum_{i=1}^{k} \mu_i(z(t))\omega_i(y(t)), \quad z(0) \in M \tag{2.4}$$
$$\xi(t,x) = \nu(z(t),t,x)$$

We shall use the Wei-Norman approach to construct a finite-dimensional filter for (2.1). Before we can achieve that, we need to introduce the concept of the estimation algebra of (2.1) and examine its algebraic structure.

Definition : The estimation algebra E of a filtering system (2.1), is defined to be the Lie algebra generated by $\{L_0, L_1, \cdots, L_m\}$, or, $E = < L_0, L_1, \cdots, L_m >_{L.A.}$

§3. Construction of general finite dimensional filters and Mitter conjecture

Let Ω be an $n \times n$ matrix whose (i,j)-entry $\partial f_j/\partial x_i - \partial f_i/\partial x_j$ are constants for all i,j. In this section, we shall assume the filtering system (2.1) has the property that Ω is a skew symmetric constant matrix.

We first observe the following Theorem

Theorem 1 : $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = c_{ij}$ are constants for all i and j if and only if $(f_1, \cdots, f_n) = (l_1, \cdots, l_n) + (\frac{\partial \phi}{\partial x_1}, \cdots, \frac{\partial \phi}{\partial x_n})$ where $l_1, \cdots, l_n$ are all polynomials of degree one and ϕ is a C^∞ function.

Observe that in Theorem 1 above if $\phi \equiv 0$ on R^n, then we are in the situation of Kalman-Bucy filtering system. If $(l_1, \cdots, l_n) \equiv 0$, then we have the Bene's filtering system as special case.

The following theorem was proven by Ocone in 1981.

Theorem (Ocone) Let E be a finite dimensional estimation algebra. If a function ξ is in E, then ξ is a polynomial of degree less than or equal to two.

One of the contribution of Mitter was to conjecture that $h_1, \cdots, h_m$ are polynomials of degree at most one if the estimation algebra is finite dimensional. This conjecture has recently been proven by the author [Ya_1] and plays the most fundamental role in the classification of finite dimensional estimation algebra.

Theorem 2 [Ya_1] Let E be a finite dimensional estimation algebra of (2.1) satisfying $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = c_{ij}$ where c_{ij} are constants for all $1 \leq i,j \leq n$. Then $h_1, \cdots, h_m$ are polynomials of degree at most one.

The argument used to prove Theorem 2 can also be used to prove the following useful theorem.

Theorem 3 [Ya_1] Let $F(x_1, \cdots, x_n)$ be a C^∞ function on R^n. Suppose that there exists a path $c : R \longrightarrow R^n$ and $\delta > 0$ such that $\lim_{t\to\infty} \|c((t)\| = \infty$ and $\lim_{t\to\infty} \sup_{B_\delta(c(t))} F = -\infty$, where $B_\delta(c(t)) = \{x \in R^n : \|x - c(t)\| < \delta\}$. Then there are no C^∞ functions $f_1, f_2, \cdots, f_n$ on R^n such that

$$\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 = F$$

For many applications, the following corollary is more convenient.

Corollary [Ya_1] Let $F(x_1, \cdots, x_n)$ be a polynomial on R^n. Suppose that there exists a polynomial path $c : R \longrightarrow R^n$ such that $\lim_{t\to\infty} \|c(t)\| = \infty$ and $\lim_{t\to\infty} F \circ c(t) = -\infty$.

Then there are no C^∞ functions $f_1, \cdots, f_n$ on R^n satisfying the equation

$$\sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 = F$$

Definition : Let E be an estimation algebra of (2.1) satisfying $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = c_{ij}$ where c_{ij} are constants for all $1 \leq i, j \leq n$. If E is finite dimensional, then the matrix

$$H = [\nabla h_1, \nabla h_2, \cdots, \nabla h_n] \tag{3.1}$$

,where we use ∇h_i to denote the column vector $(\frac{\partial h_i}{\partial x_1}, \cdots, \frac{\partial h_i}{\partial x_n})^T$, is a constant matrix in view of Theorem 2. H is called the observation matrix of (2.1).

The following result provides a single characterization of when the dimension of an estimation algebra is finite.

Let

$$D_i = \frac{\partial}{\partial x_i} - f_i$$

and

$$\eta = \sum_{i=1}^{n} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{j=1}^{m} h_j^2.$$

Then

$$L_0 = \frac{1}{2}(\sum_{i=1}^{n} D_i^2 - \eta).$$

Theorem 4 [Ya_1] Let E be an estimation algebra of (2.1) satisfying $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = c_{ij}$ where c_{ij} are constants for all $1 \leq i, j \leq n$.

(i) If η is a polynomial of degree at most two, and h_i, $1 \leq i \leq m$ are polynomials, then E is finite dimensional and has a basis consisting of $E_0 = L_0$, differential operators $E_1, \cdots, E_p$ (for some p) of the form

$$\textstyle\sum_{j=1}^{n} \alpha_{ij} D_j + \beta_i$$

where $\alpha_{ij}'s$ are constants and $\beta_i's$ are affine in x, and zero degree differential operators $E_{p+1}, \cdots, E_q, 1$ (for some $q > p$) where $E_i's$ are affine in x for $p+1 \leq i \leq q$. Moreover the quadratic part of $\eta - \sum_{i=1}^{m} h_i^2$ is positive semi-definite.

(ii) Conversely, if E is finite dimensional, then $h_1, \cdots, h_m$ are affine in x, i.e., the observation matrix has rank n (in particular $m \geq n$), then η is a polynomial of degree at most two.

In his talk at the International Congress of Mathematics, Brockett [Br_3] proposed to classify all finite dimensional estimation algebras. The following Theorem gives an important step towards the complete classification of finite dimensional estimation algebras.

Theorem 5 [Ya_1] Let E be an estimation algebra of (2.1) satisfying $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = c_{ij}$ where all c_{ij} are constants for all $1 \leq i, j \leq n$. Suppose $m \geq n$ and the observation matrix is a constant matrix with full rank. If E is finite dimensional, then it is of dimension $2n+2$ with basis given by $1, x_1, \cdots, x_n, D_1, \cdots, D_n$ and L_0.

Definition : Suppose X is a differential operator, ρ_0 is in the domain of X, r is a continuous function, and $R(t) = \int_0^t r(s)ds$. We denote by $e^{R(t)X}\rho_0$ the solution at time t of the following equation

$$\frac{\partial \rho}{\partial t}(t,x) = r(t)X\rho(t,x), \qquad \rho(0,x) = \rho_0(x)$$

if it is well-defined.

For $1 \leq i \leq n$, $e^{tD_i}\rho_0(x)$ can be expressed in the form :

$$e^{tD_i}\rho_0(x) = \rho_0(x_1, \cdots, x_i + t, \cdots, x_n) \;\; e^{-\int_0^t f_i(x_1, \cdots, x_i + t - s, \cdots, x_n)ds}.$$

Hence, we can extend easily the definition of $e^{tD_i}\rho_0(x)$ to $e^{tD_i}\rho_0(t,x)$.

Now we shall construct finite dimensional filters explicitely via the Wei-Norman approach.

Theorem 6 [Ya_1] Let E be an estimation algebra of (2.1) satisfying $\frac{\partial f_j}{\partial x_i} - \frac{\partial f_i}{\partial x_j} = c_{ij}$ where c_{ij} are constants for all $1 \leq i, j \leq n$. Suppose further that $m \geq n$ and the observation matrix has full rank, then $\eta = \sum_{j=1}^n a_{ij}x_ix_j + \sum_{i=1}^n b_ix_i + d$ where a_{ij}, b_i and d are constants for all $1 \leq i, j \leq n$ and the robust Duncan-Mortensen-Zakai equation (2.3) has a solution for all $t \geq 0$ of the form :

$$\xi(t,x) = e^{T(t)}e^{r_n(t)x_n} \cdots e^{r_1(t)x_1}e^{s_n(t)D_n} \cdots e^{s_1(t)D_1}e^{tL_0}\sigma_0 \tag{3.2}$$

where $T(t), r_1(t), \cdots, r_n(t), s_1(t), \cdots, s_n(t)$ satisfy the following ordinary differential equations (3.3), (3.4) and (3.5).

For $1 \leq i \leq n$

$$\frac{ds_i}{dt}(t) = r_i(t) + \sum_{i=1}^n s_j(t)c_{ji} + \sum_{k=1}^n h_{ki}y_k(t) \tag{3.3}$$

where $h_k(x) = \sum_{j=1}^n h_{kj}x_j + e_k$ for $1 \leq k \leq m$; h_{ki} and e_k are constants.

For $1 \leq j \leq n$

$$\frac{dr_j}{dt}(t) = \frac{1}{2}\sum_{i=1}^n s_i(t)a_{ij} \tag{3.4}$$

and

$$\frac{dT}{dt}(t) = \frac{1}{2}\sum_{i=1}^{n} r_i^2(t) - \frac{1}{2}\sum_{i=1}^{n} s_i^2(t)(\sum_{j=1}^{n} c_{ij}^2 - \frac{1}{2}a_{ii})$$

$$+ \sum_{1\le i<k\le n} s_i(t)s_k(t)(\sum_{j=1}^{n} c_{ij}c_{jk} + \frac{1}{2}a_{ik}) - \sum_{i,j=1}^{n} s_i(t)s_j(t)c_{ij} + \sum_{i=1}^{n} r_i(t) \tag{3.5}$$

$$- \sum_{j=2}^{n}\sum_{i=1}^{j} s_j(t)c_{ij} + \frac{1}{2}\sum_{i=1}^{n} s_i(t)b_i + \frac{1}{2}\sum_{i=1}^{m} y_i^2(t)\sum_{j=1}^{n} h_{ij}^2$$

§4. Classification of finite dimensional estimation algebras

The concept of estimation algebra was proven to be an invaluable tool in the study of non-linear filtering problems. So the problem of classifying all finite dimensional estimation algebras is extremely important. The following theorem is a consequence of $[TWY_1]$ and $[DTWY]$.

Theorem 7 : Suppose that the state space of the filtering system (2.1) is of dimension one. If the estimation algebra E is of finite dimensional, then either
(i) E is a real vector space of dimension 4 with basis $1, x, D = \frac{\partial}{\partial x} - f$ and $L_0 = \frac{1}{2}(D^2 - \eta)$
or (ii) E is a real vector space of dimension 2 with basis 1 and L_0
or (iii) E is a real vector space of dimension 1 with basis L_0.

Definition : The estimation algebra E, of a filtering problem (2.1), is said to be the estimation algebra with maximal rank if $x_i + c_i$ is in E for all $1 \le i \le n$ where c_i is a constant.

The following theorem due to Chiou and the author classifies all finite dimensional estimation algebras with maximal rank if $n = 2$. The novelty of the theorem is that there is no assumption on the drift term of the nonlinear filtering system.

Theorem 8 $[CY]$ Suppose that the state space of the filtering system (2.1) is of dimension two. If E is the finite dimensional estimation algebra with maximal rank, then the drift term f must be affine vector field plus gradient vector field and E is a real vector space of dimension 6 with basis given by $1, x_1, x_2, D_1, D_2$ and L_0.

Therefore from the Lie algebraic point of view, we have shown that the finite dimensional filters we constructed in §3 above are the most general finite dimensional filters.

References

[BC] R.W. Brockett and J.M.C. Clark, The geometry of the conditional density equation, in Analysis and Optimization of Stochastic Systems, O.L.R. Jacob et al, Academic Press, New York, pp. 299-309, 1980.

[Br_1] R.W. Brockett, Remarks on finite dimensional nonlinear estimation in *Analyse des Systems*. Asterisque, Vol. 75-76, pp. 47-55, 1980.

[Br_2] R.W. Brockett, Nonlinear systems and nonlinear estimation theory, in the Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Willems, eds, Reidel, Dordrecht, 1981.

[Br_3] R.W. Brockett, Nonlinear Control Theory and Differential Geometry, Proceedings of the International Congress of Mathematicians, (1983), pp. 1357-1386.

[CY] W.-L. Chiou and S. S.-T. Yau, Finite dimensional filters with nonlinear drift II : Brockett's problem on classification of finite dimensional estimation algebra. (preprint 1991)

[Da] M.H.A. Davis, On a multiplicative functional transformation arising in nonlinear filtering theory, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 125-139.

[DM] M.H.A. Davis and S.I. Marcus, An introduction to nonlinear filtering, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J.S. Willems, eds., Reidel, Dordrecht, 1981.

[$DTWY$] R.T. Dong, L.F. Tam, W.S. Wong and S. S.-T. Yau, Structure and classification theorems of finite dimensional exact estimation algebras, SIAM J. Control and Optimization Vol. 29, No.4 (1991), pp.866-877.

[FKK] M. Fujisaki, G. Kallianpur and H. Kunita, Stochastic differential equations for the nonlinear filtering problems, Osaka J. of Math., Vol. 1, pp.19-40,1972.

[Mi] S.K. Mitter, Filtering theory and Quantum fields, in *Analyse des Systems*, Asterisque, Vol. 75-76, pp. 199-205, 1980.

[TWY_1] L.F. Tam, W.S. Wong and S. S.-T. Yau, On a necessary and sufficient condition for finite dimensionality of estimation algebras, SIAM J. Control and Optimization, Vol. 28, No. 1 (1990), pp.173-185.

[TWY_2] L.F. Tam, W.S. Wong and S. S.-T. Yau, Recent results on finite dimensional exact estimation algebra, Proceedings of the 28th Conf. on Decision and Control, Tampa, Florida, Dec. (1989), pp. 2574-2575.

[Ya_1] S. S.-T. Yau, Finite dimensional filters with nonlinear drift I : A class of filters including both Kalman-Bucy filters and Benes filters. (preprint 1990)

[Ya_2] S. S.-T. Yau, Recent results on nonlinear filtering, New class of finite dimensional filters, Proceedings of the 29th Conf. on Decision and Control, Honolulu, Hawaii, Dec. (1990), pp. 231-233.

[YC] S. S.-T. Yau and W.-L. Chiou, Recent results on classification of finite dimensional estimation algebras : Dimension of state space ≤ 2, Proceedings of the 30th Conf. on Decision and Control, Brighton, England, Dec 11-13 (1991) (to appear)

An Accurate Algorithm for Minimal Partial Realizations

Adam W. Bojanczyk, Tong J. Lee, Franklin T. Luk
School of Electrical Engineering, Cornell University, Ithaca, New York 14853

ABSTRACT

We present a simple matrix representation of the Berlekamp-Massey algorithm for the minimum partial realizations problem, and show how pivoting can be added to the algorithm to improve numerical accuracy of the method.

1. Introduction

The problem of minimal realization of linear dynamical systems from input/output data has much practical importance. Many of realization procedures that have been developed rely on the solution of a Hankel system of linear equations. The Berlekamp-Massey (BM) algorithm [1], [6] is a fast Hankel linear system solver which originated in the field of coding theory. The algorithm is little known in the scientific computing community. One reason for its obscurity may be that the algorithm seems to lack a natural representation in matrix forms. Attempts to alleviate this situation can be found in Kung [4] and Jonckheere and Ma [3]. In this paper, we give a related but perhaps simpler, way to present the algorithm. We show how our presentation leads to a pivoting strategy that improves the numerical accuracy of the computation. What is more, unlike other pivoting schemes for Hankel and Toeplitz matrices, our new algorithm never requires more than $O(n^2)$ operations.

Throughout this paper, unless otherwise stated, all matrices are $n \times n$ and all vectors have n elements. Wherever convenient, we will use upper case Latin letters to denote matrices, lower case Latin letters to denote vectors, and lower case Greek letters to denote scalars. A Hankel matrix H has the form:

$$H = \begin{pmatrix} \eta_1 & \eta_2 & \cdots & \eta_n \\ \eta_2 & \eta_3 & \cdots & \eta_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_n & \eta_{n+1} & \cdots & \eta_{2n-1} \end{pmatrix}, \tag{1.1}$$

and we are interested in solving the matrix equation:

$$Hx = b. \tag{1.2}$$

By rearranging its columns or rows, the Hankel system can be transformed into a Toeplitz system. For example, we can re-order the columns of H from the last to the first, and get the Toeplitz system of equations:

$$\begin{pmatrix} \eta_n & \eta_{n-1} & \cdots & \eta_1 \\ \eta_{n+1} & \eta_n & \ddots & \eta_2 \\ \vdots & \vdots & \ddots & \vdots \\ \eta_{2n-1} & \eta_{2n-2} & \cdots & \eta_n \end{pmatrix} \begin{pmatrix} x_n \\ x_{n-1} \\ \vdots \\ x_1 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}. \tag{1.3}$$

Similarly, we define the Yule-Walker problem for the Hankel matrix to be:

$$\begin{pmatrix} \eta_1 & \eta_2 & \cdots & \eta_i \\ \eta_2 & \eta_3 & \cdots & \eta_{i+1} \\ \vdots & \vdots & \ddots & \vdots \\ \eta_i & \eta_{i+1} & \cdots & \eta_{2i-1} \end{pmatrix} y = - \begin{pmatrix} \eta_{i+1} \\ \eta_{i+2} \\ \vdots \\ \eta_{2i} \end{pmatrix}. \tag{1.4}$$

Many algorithms for solving (1.3) have been proposed, but most of them, e.g., Levinson [5], may fail to calculate an accurate solution if the Toeplitz matrix has ill-conditioned principal submatrices. Interestingly, our numerical experiments indicate that our new algorithm may still work very well under these circumstances. There is much recent interest to introduce pivoting to Toeplitz algorithms; see, e.g., [2].

This paper is organized as follows. Section 2 describes how one solves a Hankel matrix equation via the BM algorithm. Section 3 explains how the BM algorithm triangularizes a Hankel matrix that is strongly nonsingular. Section 4 presents our new numerical pivoting strategy. Section 5 considers the case of a general Hankel matrix. The last three sections contain examples that detail our numerical experience.

2. Solving a Hankel Matrix Equation

In Section 3, we will show how the Berlekamp-Massey algorithm constructs an upper triangular matrix R to reduce a Hankel matrix H to a lower triangular matrix L:

$$HR = L. \tag{2.1}$$

The triangular matrix R also has a unit diagonal. From this factorization, the Hankel system (1.2) can be easily solved. Now,

$$(HR)^T = L^T \quad \text{and} \quad H^T = H$$

imply that

$$R^T H = L^T.$$

Multiplying both sides of (1.2) by R^T, we get

$$L^T x = R^T b. \tag{2.2}$$

Hence we first apply R^T to b, and then solve the triangular system (2.2). So, if the factorization (2.1) is available, a total of n^2 multiplications is required to solve (1.2). It is worthwhile to point out here that the matrix R needs not be upper triangular. Even if R were a dense matrix, one could still solve (1.2) via (2.2), albeit at a cost of an additional $n^2/2$ multiplications. When we introduce pivoting in Section 4, we may destroy the triangularity of R.

3. Hankel Matrix Triangularization

For this section and the next, we assume that the Hankel matrix H is strongly nonsingular, i.e., that all its principal minors are nonzero. This assumption simplifies our presentation, and will be removed in Section 5. For convenience, we need a "shift-down" matrix:

$$Z = \begin{pmatrix} 0 & \cdots & 0 & 0 \\ & & & 0 \\ & I_{n-1} & & \vdots \\ & & & 0 \end{pmatrix},$$

where I_{n-1} is the identity matrix of order $(n-1)$. Thus,

$$Z \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ x_1 \\ \vdots \\ x_{n-2} \\ x_{n-1} \end{pmatrix}.$$

Note that

$$HZ = Z^T H + \begin{pmatrix} & & & -\eta_{n+1} \\ & 0_{n-1} & & \vdots \\ & & & -\eta_{2n-1} \\ \eta_{n+1} & \cdots & \eta_{2n-1} & 0 \end{pmatrix}. \tag{3.1}$$

We now show how the BM algorithm computes columns of the two matrices R and L of (2.1) recursively. We proceed by induction, and use the usual notation representing the columns of the three matrices:

$$H = (h_1, h_2, \cdots, h_n),$$

$$R = (r_1, r_2, \cdots, r_n),$$

$$L = (l_1, l_2, \cdots, l_n).$$

The first two columns of the matrices R and L are readily available:

$$r_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad , \quad r_2 = \begin{pmatrix} -\eta_2/\eta_1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{3.2}$$

and

$$l_1 = h_1 \quad , \quad l_2 = h_2 - (\eta_2/\eta_1)\, h_1. \tag{3.3}$$

Hence the top element of l_2 equals zero. Suppose now that the four columns r_j, r_{j+1}, l_j and l_{j+1} have already been computed, and that

$$H\,(\,r_j \quad r_{j+1}\,) = (\,l_j \quad l_{j+1}\,). \tag{3.4}$$

Let us denote the elements of r_{j+1} and l_{j+1} by

$$r_{j+1} = \begin{pmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_j \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad l_{j+1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_{n-j} \end{pmatrix}.$$

From the assumption of strong nonsingularity we are assuming that $\lambda_1 \neq 0$. Also, let

$$\hat{r}_{j+2} = Z r_{j+1}, \tag{3.5}$$

and

$$\hat{l}_{j+2} = H \hat{r}_{j+2}. \tag{3.6}$$

That is,

$$H\,(\,r_j \quad r_{j+1} \quad \hat{r}_{j+2}\,) = (\,l_j \quad l_{j+1} \quad \hat{l}_{j+2}\,).$$

The new vector $\hat{l}_{j+2}$ is easy to compute. From (3.1), it is seen that

$$\hat{l}_{j+2} = Z^T l_{j+1} + \xi e_n, \tag{3.7}$$

where

$$\xi = \eta_{n+1}\rho_1 + \eta_{n+2}\rho_2 + \cdots + \eta_{n+j}\rho_j + \eta_{n+j+1}, \tag{3.8}$$

and e_n denotes the last column of the $n \times n$ identity matrix. In words, the vector $\hat{l}_{j+2}$ is formed by "upshifting" each element of l_{j+1} by one slot and placing the scalar ξ in the n-th position. *A picture is worth a thousand words!* Hence we get

$$(\,l_j \quad l_{j+1} \quad \hat{l}_{j+2}\,) = \begin{pmatrix} 0 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 0 \\ \times & 0 & \lambda_1 \\ \times & \lambda_1 & \lambda_2 \\ \times & \lambda_2 & \lambda_3 \\ \vdots & \vdots & \vdots \\ \times & \lambda_{n-j-1} & \lambda_{n-j} \\ \times & \lambda_{n-j} & \xi \end{pmatrix}. \tag{3.9}$$

We now zero out the two leading nonzero elements of $\hat{l}_{j+2}$ by using appropriate multiples of the leading elements of l_j and l_{j+1}. That is, we post-multiply the $n \times 3$ matrix of (3.9) by the two 3×3 elimination matrices $E_1^{(0)}$ and $E_2^{(0)}$, where

$$E_1^{(0)} = \begin{pmatrix} 1 & 0 & m_1^{(0)} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad , \quad E_2^{(0)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & m_2^{(0)} \\ 0 & 0 & 1 \end{pmatrix}, \tag{3.10}$$

and $m_1^{(0)}$, $m_2^{(0)}$ denote the corresponding multipliers. Finally,

$$(\, l_j \quad l_{j+1} \quad l_{j+2} \,) \leftarrow (\, l_j \quad l_{j+1} \quad \hat{l}_{j+2} \,) E_1^{(0)} E_2^{(0)}, \tag{3.11}$$

and

$$(\, r_j \quad r_{j+1} \quad r_{j+2} \,) \leftarrow (\, r_j \quad r_{j+1} \quad \hat{r}_{j+2} \,) E_1^{(0)} E_2^{(0)}. \tag{3.12}$$

Note that the $(j+2)$-nd component of r_{j+2} is nonzero. Hence from the strong nonsingularity assumption, the $(j+2)$-nd component of l_{j+2} is also nonzero. Thus the multipliers are well-defined.

Let us perform an operation count. The time-consuming steps include the calculation of the inner product ξ (j multiplications), and the multiplication of a scalar into the four vectors r_j, r_{j+1}, l_j and l_{j+1} ($2n$ multiplications). Hence a total of $5n^2/2$ multiplications is required to compute the decomposition (2.1).

4. Pivoting

One may have noted that the magnitude of two multipliers m_1 and m_2 of (3.10) can be arbitrarily large. In response, we propose a simple scheme of eliminating the leading nonzero element of either l_j or $\hat{l}_{j+2}$, using either $E_1^{(0)}$ or $E_1^{(1)}$, respectively, where

$$E_1^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ m_1^{(1)} & 0 & 1 \end{pmatrix}. \tag{4.1}$$

The important point is that either $m_1^{(0)}$ or $m_1^{(1)}$ must be at most one in absolute value, in order to keep the overall process stable. We thus choose either $E_1^{(0)}$ or $E_1^{(1)}$ to achieve a better numerical accuracy. Our approach is somewhat similar to a pairwise pivoting scheme commonly used in systolic computing. Similarly, to eliminate the other nonzero element, we would choose among $E_2^{(0)}$,$E_2^{(1)}$, $E_2^{(2)}$ or $E_2^{(3)}$, where

$$E_2^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & m_2^{(1)} & 1 \end{pmatrix} \quad , \quad E_2^{(2)} = \begin{pmatrix} 1 & m_2^{(2)} & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad , \quad E_2^{(3)} = \begin{pmatrix} 1 & 0 & 0 \\ m_2^{(3)} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{4.2}$$

The location (i,j) of the multiplier represents the non-zero leading entry of column j is to be eliminated by that of column i.

The column updating proceeds essentially as before:

$$(\, l_j \quad l_{j+1} \quad l_{j+2} \,) \leftarrow (\, l_j \quad l_{j+1} \quad \hat{l}_{j+2} \,) \bar{E}_1 \bar{E}_2, \tag{4.3}$$

and

$$(\, r_j \quad r_{j+1} \quad r_{j+2} \,) \leftarrow (\, r_j \quad r_{j+1} \quad \hat{r}_{j+2} \,) \bar{E}_1 \bar{E}_2, \tag{4.4}$$

where $\bar{E}_1$ equals $E_1^{(0)}$ or $E_1^{(1)}$, and $\bar{E}_2$ equals $E_2^{(0)}$, $E_2^{(1)}$, $E_2^{(2)}$ or $E_2^{(3)}$. Two important observations are as follows. First, the resultant matrix L stays lower triangular, but the previously upper triangular R may have gained two nonzero subdiagonals. Second, our pivoting scheme increases the number of multiplications by only $O(n)$, i.e., the total number of multiplications is still $5n^2/2 + O(n)$.

5. General case

Recall that under the strong nonsingularity assumption, we knew that the $(j+1)$-st element of l_{j+1} must be nonzero. We now remove the assumption that the matrix H is strongly nonsingular. During the elimination process, we may get additional leading zero elements in l_{j+1}. For our discussion in this section, let us assume that both $(j+1)$-st and $(j+2)$-nd elements are zero but that the $(j+3)$-rd element is nonzero. Hence the procedure described in Section 3 would not work because there is a gap in the nonzero structure in (3.7). Now, let

$$r_{j+1} = \begin{pmatrix} \rho_1 \\ \vdots \\ \rho_j \\ \rho_{j+1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{and} \quad l_{j+1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 0 \\ \lambda_3 \\ \vdots \\ \lambda_{n-j} \end{pmatrix}.$$

Define some new vectors by

$$\hat{r}_{j+2} = Zr_{j+1}, \quad \hat{r}_{j+3} = Z\hat{r}_{j+2}, \quad \hat{r}_{j+4} = Z\hat{r}_{j+3}, \tag{5.1}$$

and

$$\hat{l}_{j+2} = H\hat{r}_{j+2}, \quad \hat{l}_{j+3} = H\hat{r}_{j+3}, \quad \hat{l}_{j+4} = H\hat{r}_{j+4}. \tag{5.2}$$

That is,

$$H\,(\,r_j \quad r_{j+1} \quad \hat{r}_{j+2} \quad \hat{r}_{j+3} \quad \hat{r}_{j+4}\,) = (\,l_j \quad l_{j+1} \quad \hat{l}_{j+2} \quad \hat{l}_{j+3} \quad \hat{l}_{j+4}\,).$$

From (3.1), the new vectors $\hat{l}_{j+2}$, $\hat{l}_{j+3}$ and $\hat{l}_{j+4}$ are calculated by

$$\begin{aligned} \hat{l}_{j+2} &= Z^T l_{j+1} + \xi_1 e_n, \\ \hat{l}_{j+3} &= Z^T \hat{l}_{j+2} + \xi_2 e_n, \\ \hat{l}_{j+4} &= Z^T \hat{l}_{j+3} + \xi_3 e_n, \end{aligned} \tag{5.3}$$

where

$$\begin{aligned} \xi_1 &= \eta_{n+1}\rho_1 + \eta_{n+2}\rho_2 + \cdots + \eta_{n+j+1}\rho_{j+1}, \\ \xi_2 &= \eta_{n+2}\rho_1 + \eta_{n+3}\rho_2 + \cdots + \eta_{n+j+2}\rho_{j+1}, \\ \xi_3 &= \eta_{n+3}\rho_1 + \eta_{n+4}\rho_2 + \cdots + \eta_{n+j+3}\rho_{j+1}. \end{aligned} \tag{5.4}$$

Indeed, the *million words* picture looks like:

$$(\,l_j \quad l_{j+1} \quad \hat{l}_{j+2} \quad \hat{l}_{j+3} \quad \hat{l}_{j+4}\,) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 \\ \times & 0 & 0 & 0 & \lambda_3 \\ \times & 0 & 0 & \lambda_3 & \lambda_4 \\ \times & 0 & \lambda_3 & \lambda_4 & \lambda_5 \\ \times & \lambda_3 & \lambda_4 & \lambda_5 & \lambda_6 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \times & \lambda_{n-j-2} & \lambda_{n-j-1} & \lambda_{n-j} & \xi_1 \\ \times & \lambda_{n-j-1} & \lambda_{n-j} & \xi_1 & \xi_2 \\ \times & \lambda_{n-j} & \xi_1 & \xi_2 & \xi_3 \end{pmatrix}. \tag{5.5}$$

As described in Section 4, we would like to pivot and eliminate the above matrix so that each row contains an unique pivot element. The elimination matrices in this case are four 5×5 matrices, each with all 1's on the diagonal and a multiplier in the (i, j) location. Same as before, the jth leading non-zero entry is to be eliminated by the ith column, and all the multipliers in the elimination are less than one in absolute value.

In practice, we work with finite-precision arithmetic, so an exact zero would hardly happen. In order to tell if we are getting any additional zeros in the column of l_{j+1}, we need to choose a threshold, such that any number smaller (in absolute value) than the threshold is regarded as a zero. If this is the case, we will then apply the technique in this section to deal with the situation.

6. Numerical Examples

We consider the Hankel matrix equation (1.2) and the corresponding Toeplitz matrix equation (1.3). We compare three procedures: the BM algorithm for (1.2), our new pivoted BM algorithm for (1.2), and the Levinson algorithm for (1.3). We construct two sets of examples, the first where BM would fail and the second where Levinson would fail. Specifically, we tinker with the 2×2 leading submatrices of the Hankel and the corresponding Toeplitz matrices:

$$H^{(2)} = \begin{pmatrix} \eta_1 & \eta_2 \\ \eta_2 & \eta_3 \end{pmatrix} \quad \text{and} \quad T^{(2)} = \begin{pmatrix} \eta_n & \eta_{n-1} \\ \eta_{n+1} & \eta_n \end{pmatrix}. \tag{6.1}$$

In Example 1, the submatrix $H^{(2)}$ is ill-conditioned but the submatrix $T^{(2)}$ is not, while in Example 2, the situation is reversed. In Example 3 both submatrices are ill-conditioned. Whereas the BM algorithm fails in Examples 1 and 3, and the Levinson algorithm fails in Examples 2 and 3, our new algorithm works well on all three sets of equations.

For all examples in this paper, we choose the left hand vector b such that the solution vector $x = (1 \quad 1 \quad \cdots \quad 1)^T$. To compare the algorithms, we calculate $\| x - \hat{x} \|_2$, where $\hat{x}$ denotes the computed solution. We ran our examples using MATLAB on a Sun Sparc station. In this section we choose as a threshold, $\epsilon \| H \|_2$, where ϵ ($\approx 2.22 \cdot 10^{-16}$) denotes the machine precision. We use $\kappa(M)$ to denote the condition number with respect to the 2-norm of a matrix M.

Example 1. This example shows why pivoting is necessary for the Berlekamp-Massey algorithm. Let

$$H_1 = \begin{pmatrix} 1-\delta & 2 & 4 & 8 \\ 2 & 4 & 8 & 4 \\ 4 & 8 & 4 & 2 \\ 8 & 4 & 2 & 1-\delta \end{pmatrix} \quad \text{and} \quad H_1^{(2)} = \begin{pmatrix} 1-\delta & 2 \\ 2 & 4 \end{pmatrix}.$$

The submatrix of $H_1^{(2)}$ is ill-conditioned when δ is small, and is singular when δ is zero. As expected, the BM algorithm delivers worse accuracy as we decrease the size of δ. The matrix H_1 is well conditioned, with $\kappa(H_1) = 5.6$. However, the BM algorithm determines an L that is ill-conditioned, contributing to the loss in accuracy when one solves (1.2) via (2.2). On the other hand, our new algorithm computes a very well-conditioned L.

Table 1. Error Behavior for Example 1

	BM		Pivoted BM		Levinson
δ	$\kappa(L)$	$\| x - \hat{x} \|_2$	$\kappa(L)$	$\| x - \hat{x} \|_2$	$\| x - \hat{x} \|_2$
10^{-2}	$2.0 \cdot 10^{5}$	$9.98 \cdot 10^{-14}$	23	$3.24 \cdot 10^{-15}$	$1.11 \cdot 10^{-16}$
10^{-4}	$2.0 \cdot 10^{9}$	$4.96 \cdot 10^{-12}$	23	$3.24 \cdot 10^{-15}$	$0.00 \cdot 10^{-16}$
10^{-6}	$2.0 \cdot 10^{13}$	$4.96 \cdot 10^{-10}$	23	$3.24 \cdot 10^{-15}$	$4.00 \cdot 10^{-16}$
10^{-8}	$2.0 \cdot 10^{17}$	$1.98 \cdot 10^{-7}$	23	$3.24 \cdot 10^{-15}$	$1.11 \cdot 10^{-16}$

Example 2. This is an example where the Levinson algorithm fails because the submatrix $T_2^{(2)}$ is ill-conditioned:

$$H_2 = \begin{pmatrix} 0 & 2 & 1-\delta & 1 \\ 2 & 1-\delta & 1 & 1-\delta \\ 1-\delta & 1 & 1-\delta & 2 \\ 1 & 1-\delta & 2 & 0 \end{pmatrix} \quad \text{and} \quad T_2^{(2)} = \begin{pmatrix} 1 & 1-\delta \\ 1-\delta & 1 \end{pmatrix}.$$

This example also portrays a unique property of the BM algorithm, that it still works even though the (1,1) element of the matrix equals zero. However, the algorithm may deliver a poor solution if the (1,1) element is non-zero but small in size. Again, here the Hankel matrix is well-conditioned, with $\kappa(H_2) = 7.3$. Both our new algorithm and the BM algorithm calculate well conditioned L.

Table 2. Error Behavior for Example 2

	BM		Pivoted BM		Levinson
δ	$\kappa(L)$	$\| x - \hat{x} \|_2$	$\kappa(L)$	$\| x - \hat{x} \|_2$	$\| x - \hat{x} \|_2$
10^{-2}	36	$1.87 \cdot 10^{-15}$	5.3	$3.14 \cdot 10^{-16}$	$6.08 \cdot 10^{-15}$
10^{-4}	34	$1.85 \cdot 10^{-15}$	5.3	$6.28 \cdot 10^{-16}$	$3.07 \cdot 10^{-13}$
10^{-6}	34	$4.15 \cdot 10^{-16}$	5.3	$5.87 \cdot 10^{-16}$	$9.32 \cdot 10^{-11}$
10^{-8}	34	$2.24 \cdot 10^{-15}$	5.3	$1.11 \cdot 10^{-16}$	$1.10 \cdot 10^{-8}$

Example 3. This is an example where both BM and Levinson algorithms fail because the submatrices $H_3^{(2)}$ and $T_3^{(2)}$ are ill-conditioned:

$$H_3 = \begin{pmatrix} 1-\delta & 2 & 4 & 1-\delta & 1 \\ 2 & 4 & 1-\delta & 1 & 1-\delta \\ 4 & 1-\delta & 1 & 1-\delta & 4 \\ 1-\delta & 1 & 1-\delta & 4 & 2 \\ 1 & 1-\delta & 4 & 2 & 1-\delta \end{pmatrix}, \quad H_3^{(2)} = \begin{pmatrix} 1-\delta & 2 \\ 2 & 4 \end{pmatrix} \quad \text{and} \quad T_3^{(2)} = \begin{pmatrix} 1 & 1-\delta \\ 1-\delta & 1 \end{pmatrix}.$$

Again, here the Hankel matrix is well-conditioned, with $\kappa(H_2) \approx 31$. Our new algorithm calculates a "good" L, but the BM algorithm determines an L that is ill-conditioned.

Table 3. Error Behavior for Example 3

	BM		Pivoted BM		Levinson
δ	$\kappa(L)$	$\| x - \hat{x} \|_2$	$\kappa(L)$	$\| x - \hat{x} \|_2$	$\| x - \hat{x} \|_2$
10^{-2}	$7.0 \cdot 10^{4}$	$9.45 \cdot 10^{-11}$	279	$2.75 \cdot 10^{-14}$	$1.12 \cdot 10^{-13}$
10^{-4}	$7.0 \cdot 10^{8}$	$4.63 \cdot 10^{-15}$	288	$4.63 \cdot 10^{-15}$	$2.52 \cdot 10^{-11}$
10^{-6}	$7.0 \cdot 10^{12}$	$1.15 \cdot 10^{-14}$	288	$1.15 \cdot 10^{-14}$	$3.85 \cdot 10^{-9}$
10^{-8}	$3.8 \cdot 10^{16}$	$8.67 \cdot 10^{-14}$	288	$8.67 \cdot 10^{-14}$	$7.21 \cdot 10^{-8}$

The three examples have shown how our pivoting scheme works better than the other two conventional algorithms. However, when the size of the matrix H increases, roundoff errors may accumulate so that it becomes hard to define a numerical zero. If we choose a small threshold, such as the one we have used, then the condition number of any principal submatrix may become as large as the inverse of the threshold and a significant loss in accuracy may occur. From our experiments, we observed that the accuracy of the solution is proportional to the largest condition number of any principal submatrices.

On the other hand, if the threshold is large, we effectively work in a lower precision, so the factorization will have limited numerical accuracy. Therefore, in the next section, we experiment with a compromised threshould value. To compensate for the loss in accuracy due to this choice of a larger threshold, we adopt iterative refinement at the end.

7. Further Examples

In this section we show how our pivoting scheme works when the size of the matrix H increases. We observe that with an increase in dimensions there is a danger of underflow. Hence some form of normalization is required. Our choice is to normalize L to make the diagonal elements all ones, so that underflow can be avoided.

We construct our matrices from the Toeplitz examples in Sweet's paper [7], and select an iterative refinement scheme for improving the accuracy of the initial solution $x^{(0)}$:

1. Compute $r^{(i)} = Hx^{(i)} - b$.
2. Solve $L^T y = R^T r^{(i)}$.
3. Update $x^{(i+1)} = x^{(i)} - y$.

The criterion for ending the iterative refinement is when

$$\| r^{(i)} \|_2 < 10 \cdot \epsilon \cdot \| H \|_2 \cdot \| x^{(i)} \|_2 .$$

The threshold for the examples in this section is chosen as $10 \cdot \sqrt{\epsilon} \cdot \| H \|_2$.

Example 4. We pick an example to show how iterative refinement improves our solution, and how the number of refinements is affected by the conditioning of principal submatrices. The order-6 Hankel matrix is

$$H_4 = \begin{pmatrix} 3 & 2 & 6 & 1 & 195/14+\delta & 8 \\ 2 & 6 & 1 & 195/14+\delta & 8 & 4 \\ 6 & 1 & 195/14+\delta & 8 & 4 & -34 \\ 1 & 195/14+\delta & 8 & 4 & -34 & 5 \\ 195/14+\delta & 8 & 4 & -34 & 5 & 3 \\ 8 & 4 & -34 & 5 & 3 & 1 \end{pmatrix}.$$

For δ smaller in the magnitude than 0.5, the matrix is well conditioned with $\kappa(H_4)$ less than 100. The threshold is approximately equal to $6 \cdot 10^{-6}$. For $\delta = 0$ the order-3 principal submatrix is singular. Starting with $\delta = 10^{-2}$ and then decreasing it , we can make this submatrix progressively worse conditioned without significantly changing the condition number of H_4.

Table 4.1. Error Behavior for Example 4

	BM	Pivoted BM	BM	Pivoted BM
δ	10^{-2}	10^{-2}	10^{-4}	10^{-4}
$\sigma_{min}(H_4^{(3)})$	$1.36 \cdot 10^{-3}$	$1.36 \cdot 10^{-3}$	$1.36 \cdot 10^{-5}$	$1.36 \cdot 10^{-5}$
$\kappa(L)$	$2.29 \cdot 10^{7}$	197	$2.29 \cdot 10^{11}$	197
$\kappa(R)$	$2.29 \cdot 10^{7}$	182	$2.29 \cdot 10^{11}$	182
$\| x - \hat{x}^{(0)} \|_2$	$1.06 \cdot 10^{-8}$	$1.33 \cdot 10^{-13}$	$6.59 \cdot 10^{-5}$	$2.94 \cdot 10^{-11}$
$\| x - \hat{x}^{(1)} \|_2$	$7.61 \cdot 10^{-16}$		$7.84 \cdot 10^{-10}$	$2.88 \cdot 10^{-15}$
$\| x - \hat{x}^{(2)} \|_2$			$5.27 \cdot 10^{-15}$	
#Refine.	1	0	2	1

Table 4.2. Error Behavior for Example 4 (Continued)

	BM	Pivoted BM	BM	Pivoted BM
δ	10^{-5}	10^{-5}	10^{-6}	10^{-6}
$\sigma_{min}(H_4^{(3)})$	$1.36 \cdot 10^{-6}$	$1.36 \cdot 10^{-6}$	$1.36 \cdot 10^{-7}$	$1.36 \cdot 10^{-7}$
$\kappa(L)$	$2.29 \cdot 10^{13}$	197	214	122
$\kappa(R)$	$2.29 \cdot 10^{13}$	182	202	53
$\| x - \hat{x}^{(0)} \|_2$	$1.23 \cdot 10^{-2}$	$1.12 \cdot 10^{-10}$	$1.47 \cdot 10^{-7}$	$2.11 \cdot 10^{-7}$
$\| x - \hat{x}^{(1)} \|_2$	$5.93 \cdot 10^{-6}$	$3.84 \cdot 10^{-15}$	$8.89 \cdot 10^{-14}$	$1.55 \cdot 10^{-13}$
$\| x - \hat{x}^{(2)} \|_2$	$2.71 \cdot 10^{-9}$		$1.37 \cdot 10^{-15}$	$4.29 \cdot 10^{-15}$
$\| x - \hat{x}^{(3)} \|_2$	$1.24 \cdot 10^{-12}$			
$\| x - \hat{x}^{(4)} \|_2$	$3.67 \cdot 10^{-15}$			
$\#Refine.$	4	1	2	2

Table 4.3. Error Behavior for Example 4 (Continued)

	BM	Pivoted BM	BM	Pivoted BM
δ	10^{-8}	10^{-8}	10^{-10}	10^{-10}
$\sigma_{min}(H_4^{(3)})$	$1.36 \cdot 10^{-9}$	$1.36 \cdot 10^{-9}$	$1.36 \cdot 10^{-11}$	$1.36 \cdot 10^{-11}$
$\kappa(L)$	214	122	214	122
$\kappa(R)$	202	53	202	53
$\| x - \hat{x}^{(0)} \|_2$	$1.47 \cdot 10^{-9}$	$2.11 \cdot 10^{-9}$	$1.47 \cdot 10^{-11}$	$2.11 \cdot 10^{-11}$
$\| x - \hat{x}^{(1)} \|_2$	$2.51 \cdot 10^{-15}$	$3.73 \cdot 10^{-15}$	$5.87 \cdot 10^{-16}$	$3.92 \cdot 10^{-15}$
$\#Refine.$	1	1	1	1

Note that the pivoted BM algorithm always produces factors R and L that are better conditioned than those produced by the BM algorithm without pivoting. As a consequence, the first approximation to the solution computed by the pivoted BM algorithm is more accurate than that computed by the BM algorithm. When the the smallest singular value of the order 3 principal submatrix becomes smaller than the threshold, both algorithms behave in a similar way.

Example 5. We construct a 13×13 Hankel matrix H_5 whose first row is given by

$$\eta_{1-13} = (-15, 10, 1, -7, -2, -5, -14.2766, -25.5087, -48.8789, -96.8384, -188.8878, -1, 5),$$

and the last column by

$$\eta_{13-25} = (5, 1, -3, 12.755, -19.656, 28.361, -7, -1, 2, 1, -6, 1, -0.5)^T,$$

the threshold is $5.35 \cdot 10^{-5}$.

The matrix is well conditioned in that $\kappa(H_5) = 89.0$, but it contains five consecutive ill-conditioned submatrices $H_5^{(4)}$ to $H_5^{(8)}$, i.e., orders 4 through 8. The smallest singular values of these five principal submatrices are $2.36 \cdot 10^{-5}$, $5.23 \cdot 10^{-5}$, $5.33 \cdot 10^{-5}$, $5.23 \cdot 10^{-5}$ and $2.36 \cdot 10^{-5}$. The effect of encountering a sequence of ill-conditioned submatrices is felt later in the elimination process and is manifested by a severe loss in accuracy in the subsequent columns of R and L. Hence, some form of restoration is required for high accuracy. A way to bring back the lost information is to recompute the most recent columns of R by solving a Yule-Walker problem as in (1.4), utilizing the decomposition we already have at hand. In this example, columns 10 and 11 of R are recomputed, so that again the process restarts from a new and accurate point. The results are presented in Table 5. Notice that the two factor matrices produced by BM algorithm are nearly singular.

Table 5. Error Behavior for Example 5

	BM	Pivoted BM
$\kappa(L)$	$3.93 \cdot 10^{+13}$	$1.33 \cdot 10^{+3}$
$\kappa(R)$	$3.93 \cdot 10^{+13}$	$1.24 \cdot 10^{+3}$
$\| x - \hat{x}^{(0)} \|_2$	$6.58 \cdot 10^{-2}$	$3.29 \cdot 10^{-4}$
$\| x - \hat{x}^{(1)} \|_2$	$2.93 \cdot 10^{-5}$	$2.11 \cdot 10^{-9}$
$\| x - \hat{x}^{(2)} \|_2$	$8.82 \cdot 10^{-8}$	$4.18 \cdot 10^{-14}$
$\| x - \hat{x}^{(4)} \|_2$	$1.58 \cdot 10^{-13}$	
#Refine.	4	2

Example 6. We extend the previous example to size 50×50 by appending random numbers. For this example, the Hankel matrix is moderately ill-conditioned with $\kappa(H_6) = 1.97 \cdot 10^{+3}$. The results are shown in Table 6.

Table 6. Error Behavior for Example 6

	BM	Pivoted BM
$\kappa(L)$	$4.01 \cdot 10^{+13}$	$1.33 \cdot 10^{+3}$
$\kappa(R)$	$4.01 \cdot 10^{+13}$	$8.71 \cdot 10^{+4}$
$\| x - \hat{x}^{(0)} \|_2$	$5.86 \cdot 10^{-1}$	$2.44 \cdot 10^{-3}$
$\| x - \hat{x}^{(1)} \|_2$	$2.39 \cdot 10^{-3}$	$4.54 \cdot 10^{-8}$
$\| x - \hat{x}^{(2)} \|_2$	$7.46 \cdot 10^{-6}$	$9.86 \cdot 10^{-13}$
$\| x - \hat{x}^{(3)} \|_2$	$6.87 \cdot 10^{-8}$	$5.58 \cdot 10^{-14}$
$\| x - \hat{x}^{(5)} \|_2$	$1.17 \cdot 10^{-12}$	
#Refine.	5	3

8. Conclusion

We believe that the Berlekamp-Massey algorithm works well when the Hankel matrix is positive definite and well-conditioned, so that none of its principal submatrices is ill-conditioned, and no pivoting is necessary. In general, consider the Hankel matrix as a moments matrix with respect to certain weights. These weights are not necessarily all positive, and thus we may need to deal with Hankel matrices that do not have positive definite property. Strategies such as pivoting, normalization and gap-jumping are required in this case.

For the previous examples, we adopt a scheme that combines both pivoting and normalization. To generate a new column of the triangular factor, say column i, we combine it with columns $i-1$ and $i-2$. Since normalization is performed after each new column is generated, and column i is a shifted version of column $i-1$, so all the three columns have a 1 as the leading nonzero element. Therefore, no pivoting is performed in the first phase of column combination. After columns i and $i-2$ are combined, pivoting takes place in the second phase, in which columns i and $i-1$ are combined. Both steps are crucial to the stability of the procedure. Pivoting prevents multipliers from being too large, while normalization keeps the norm of the columns from underflowing.

Since we remove the constraint of positive-definiteness, a well-conditioned Hankel matrix may have several ill-conditioned submatrices. The choice of the threshold is a subtle issue, and from the previous examples, we see that a compromised threshould value, such as $10 \cdot \sqrt{\epsilon} \cdot \| H \|_2$, may be a good choice.

Whenever there is a sequence of ill-conditioned principal submatrices, i.e., a gap, we simply shift up the previous column of L until the next well-conditioned submatrix is encountered. Thus, we avoid the computation within the gap by "jumping over" it. Two columns of L are recomputed right after the gap, so that the errors in the factorization caused by the jumps are confined within the gap and do not propagate to the succeeding columns. Therefore, after a few steps of iterative refinement at the end, all the errors in the solution (not the decomposition) will be corrected and the solution will be accurate to machine precision.

Another possibility to deal with the gap is to perform a LU decomposition with partial pivoting instead of jumping over it. But the worst case for this approach requires $O(p^2n)$ operations to decompose the part of the matrix corresponding to the gap, in contrast to $O(pn)$ operations for the gap-jumping approach, where p denotes the size of the gap.

Acknowledgements

A. W. Bojanczyk was supported in part by the Army Research Office under contract DAAL03-90-G-0092. T. J. Lee and F. T. Luk were supported in part by the Army Research Office under contract DAAL03-90-G-0104, and by the Joint Services Electronics Program under contract F49620-90-C0039.

References

[1] E. R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill, New York, NY, 1968.

[2] T. F. Chan and P. C. Hansen, "A stable Levinson algorithm for general Toeplitz systems," CAM Report 90-11, Computational and Applied Mathematics, University of California, Los Angeles, CA, 1990.

[3] E. Jonckheere and C. Ma, "A simple Hankel interpretation of the Berlekamp-Massey algorithm," *Linear Alg. Applics.*, vol. 125 (1989), pp. 65-76.

[4] S.-Y. Kung, "Multivariable and Multidimensional Systems: Analysis and Design," Ph.D. Dissertation, Department of Electrical Engineering, Stanford University, Stanford, CA, 1977.

[5] N. Levinson, "The Wiener RMS (root-mean-square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25 (1946), pp. 261-278.

[6] J. L. Massey, "Shift register synthesis and BCH decoding," *IEEE Trans. Inform. Theory*, vol. IT-15 (1967), pp. 122-127.

[7] D. R. Sweet, "Numerical methods for Toeplitz matrices," Ph.D. Dissertation, Department of Computing Science, University of Adelaide, Australia, 1982.

The Hyperbolic Transformations in Signal Processing and Control

Adam Bojanczyk and Allan O. Steinhardt
Department of Electrical Engineering
Phillips Hall
Cornell University
Ithaca, NY 14853-3801

October 8, 1991

1 Introduction

The difference $X = A_1{}^\dagger A_1 - A_2{}^\dagger A_2$ of two matrix outer products $A_1{}^\dagger A_1$ and $A_2{}^\dagger A_2$ arises in regression problems, in signal processing in the context of bearing estimation, and other applications. It is of practical interest to consider two problems related to the matrix X. The first problem is to find the triangular decomposition of X, the second problem is to find the eigendecomposition of X. For numerical reasons it is desirable not to form explicitly the products $A_1{}^\dagger A_1$ and $A_2{}^\dagger A_2$. In this paper we describe how these two problems can be solved with the help of hyperbolic type transformations.

2 Triangular Decomposition

In considering the difference $A_1{}^\dagger A_1 - A_2^\dagger A_2$ it is helpful to introduce an indefinite inner product $[\cdot,\cdot]_\Phi$ on C^n induced by a weighting matrix Φ, $\Phi = diag(\pm 1)$. This product is defined as follows

$$x, y \in C^n \ , \ [x,y]_\Phi = x^\dagger \Phi y$$

where $\dagger$ denotes conjugate transpose. (A broad treatment of indefinite inner products and their applications can be found in [3].) The indefinite inner product $[.,.]_\Phi$ defines the indefinite norm $||\cdot||_\Phi$,

$$||x||_\Phi = sign([x,x]_\Phi)\sqrt{|[x,x]_\Phi|}.$$

Note that despite the notation $\|v\|_\Phi$ is not a norm, because norms are always non-negative. The weighting matrix Φ, often referred to as the signature matrix, defines also hypernormal (with respect to Φ) matrices. A matrix V is hypernormal if

$$V^\dagger \Phi V = \Phi \ .$$

The indefinite norm $||\cdot||_\Phi$ is preserved under hypernormal transformations. Hypernormal matrices can be used in the computation of the triangular factorization of the difference $A_1{}^\dagger A_1 - A_2^\dagger A_2$

(without forming the products $A_1{}^\dagger A_1$ and $A_2{}^\dagger A_2$). If X is positive definite and the Cholesky factor of $A_1^\dagger A_1$ is known then the problem of computing the Cholesky factor of $A_1{}^\dagger A_1 - A_2^\dagger A_2$ is known as the downdating of the Cholesky factor. For a discussion of algorithms based on hypernormal transformations that can be applied to downdating problem see [1] and references therein.

Here we want to show how to compute a triangular decomposition of an indefinite strongly nonsingular $X = A_1{}^\dagger A_1 - A_2^\dagger A_2$. A square matrix is called strongly nonsingular iff its all principal minor are nonsingular. Strongly nonsingular matrices admit a triangular decomposition of the type $R^\dagger \Phi R$ where R is upper triangular and Φ is diagonal [4]. The tool that we propose to use is the hyperbolic Householder transformation [8].

The original Householder transformation [4] of a (column) vector v involves finding an orthonormal matrix Q so that

$$Qv = \pm \| v \| e_1, \tag{1}$$

where $\| v \| = \sqrt{v^\dagger v}$ and e_1 is the unit vector with the first element one and all the rest zeros. This can be viewed as *compressing* all the vector's energy into the first entry. It is easily verified that a matrix Q given by: $Q = I - 2bb^\dagger / b^\dagger b$ where $b = v \mp \| v \| e_1$ satisfies (1).

The hyperbolic Householder transform will take on a similar form, and a signature matrix Φ has to be specified as well as the vector v. The natural thing is to let

$$b = v \mp \| v \|_\Phi e_1 \ , \ H = \Phi - 2bb^\dagger / b^\dagger \Phi b . \tag{2}$$

If v's and Φ's are such that $v^\dagger \Phi v > 0$ then H is always well defined, see [8], [7]. We would like to be able to obtain H for any pair Φ and v for which $v^\dagger \Phi v \neq 0$.

If v denotes the original vector and $\hat{v}$ denotes the transformed vector, then we expect the following:

$$\hat{v}^\dagger \Phi \hat{v} = v^\dagger \Phi v \ , \tag{3}$$

$$\hat{v} = \pm \| v \|_\Phi e_1 \ , \tag{4}$$

$$H^\dagger \Phi H = \Phi \ . \tag{5}$$

The relation (4) can be viewed as compressing all *hyperbolic energy* of v into its first component. It turns out that the two conditions (3) and (4) cannot generally be met simultaneously. This is because from (3) we would expect that

$$\text{sign}(v^\dagger \Phi v) = \text{sign}(\hat{v}^\dagger \Phi \hat{v}) \tag{6}$$

From (4), the sign of the right hand side of (6) is determined by $\text{sign}(\Phi(1,1))$ and is independent of the sign of $v^\dagger \Phi v$. Hence (3) and (4) may contradict each other. Note however that if $\| v \|_\Phi \neq 0$ then there exists k, $1 \leq k \leq n$ such that

$$\text{sign}(\Phi(k,k)) = \text{sign}(\| v \|_\Phi) \ . \tag{7}$$

Now, by permuting entries 1 and k in v, and entries $(1,1)$ and (k,k) in Φ we obtain $\tilde{v}$ and $\tilde{\Phi}$,

$$\tilde{v} = Pv \ , \ \tilde{\Phi} = P \Phi P^T \ ,$$

where P is the permutation matrix, for which (6) will be satisfied. We will now show that (3) and (4) will hold for the permuted quantities.

Let v and Φ be such that $v^\dagger \Phi v \neq 0$ and without loss of generality we can assume that (6) is satisfied. Define

$$b = \Phi v + \theta \mathrm{abs}(\| v \|_\Phi) e_1 \ . \tag{8}$$

Note that

$$b^\dagger \Phi b = \mathrm{sign}(v^\dagger \Phi v) \| v \|_\Phi^2 + \theta \bar{v}_1 \mathrm{abs}(\| v \|_\Phi) + \bar{\theta} v_1 \mathrm{abs}(\| v \|_\Phi) + |\theta|^2 \| v \|_\Phi^2 \mathrm{sign}(e_1^T \Phi e_1) \tag{9}$$

If we pick

$$\theta = \begin{cases} \mathrm{sign}(e_1^T \Phi e_1) \frac{v_1}{|v_1|} & \text{if } v_1 \neq 0 \\ \mathrm{sign}(e_1^T \Phi e_1) & \text{otherwise} \end{cases} \tag{10}$$

then (9) becomes

$$b^\dagger \Phi b = 2 \, \mathrm{sign}(e_1^T \Phi e_1)(\| v \|_\Phi^2 + \mathrm{abs}(\| v \|_\Phi)|v_1|) \ . \tag{11}$$

Now it is easy to check that for H defined by (2) and (8)

$$Hv = -\theta \mathrm{abs}(||v||_\Phi) e_1 \tag{12}$$

and H is hypernormal with respect to Φ [1]. The relation (12) states that any vector v with nonzero hyperbolic norm can be "reflected" by a hyperbolic Householder transformation onto the first coordinate e_1. It is easy to see that e_1 can be replaced by any direction d for which $d^\dagger \Phi d \neq 0$.

At this point a problem that should be addressed: what happens when $\| v \|_\Phi = 0$? The answer is that both procedures per se fail (see [2] for some implications of this problem). What we rely upon in recovering from a situation of $\| v \|_\Phi = 0$ is that the hyperbolic Householder is applied to whole matrices, not merely to isolated column vectors. When the matrix under consideration is strongly nonsingular then for a suitable permutation of columns we will always be able to assure that $\| v \|_\Phi \neq 0$.

Recall that we seek a decomposition of the following form

$$A^\dagger \Phi A = A_1^\dagger A_1 - A_2^\dagger A_2 = R^\dagger \tilde{\Phi} R$$

We can construct a sequence of hyperbolic Householder transformations $H_1, H_2, \cdots, H_k$, where H_i is hyperbolic with respect to a signature matrix Φ_i, such that

$$H_k P_k \cdots H_2 P_2 H_1 P_1 \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} = \begin{bmatrix} R \\ 0 \end{bmatrix} \tag{13}$$

where P_i is a suitable permutation for which the relation (7) is satisfied. The permutations P_i, $i = 1, 2, \ldots, k$, and the signature matrices Φ_i, $i = 1, 2, \ldots, k$, are related via

$$P_i^T \Phi_i P_i = \Phi_{i-1}$$

with $\Phi_0 \equiv \Phi$.

[1] This extension of the hyperbolic Householder transform for nonpositive normed vectors was also developed (independently) by Cybenko [2] in a different context.

From (13) we obtain Φ_k and R such that

$$A^\dagger \Phi A = R^\dagger \Phi_k R$$

If in step i the working column v_i and the signature matrix Φ_i are such that

$$v_i^\dagger \Phi_i v_i = 0$$

then a suitable permutation S_i of the remaining columns of A has to chosen so for the new working column v_i

$$v_i^\dagger \Phi_i v_i \neq 0 \, .$$

This is possible as $A^\dagger \Phi A$ is assumed to be strongly nonsingular [4]. On completion we get the desired triangular factorization of $A = [A_1^\dagger \, , \, A_2^\dagger]^\dagger$,

$$P^T A^\dagger \Phi A P = R^\dagger \Phi_k R \, .$$

3 Eigendecomposition

Recall that the SVD of an $n \times m$ matrix A is given by :

$$A = USV^\dagger,$$

where S is an $n \times m$ diagonal with non-negative diagonal, U is an $n \times n$ unitary, V is an $m \times m$ unitary. Note that for $A = [A_1^\dagger \, , \, A_2^\dagger]^\dagger$,

$$A^\dagger A = A_1^\dagger A_1 + A_2^\dagger A_2 = VS^2V^\dagger \, .$$

Thus the SVD provides the eigenvalues and eigenvectors of $A_1^\dagger A_1 + A_2^\dagger A_2$. For numerical reasons it proves more numerically accurate to operate on the data matrix A directly and the SVD is the tool makes this possible.

Consider now an analogous problem of finding the eigenvectors and eigenvalues of $A_1^\dagger A_1 - A_2^\dagger A_2$ where A_1, A_2 are two $n \times m$ matrices. Or more generally, given a matrix A and a matrix Φ that is diagonal with ± 1 on the diagonal, find the eigenvalues and the eigenvectors of $A\Phi A^\dagger$. This by setting:

$$A = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \, , \; \Phi = \begin{bmatrix} I & \mathbf{0} \\ \mathbf{0} & -I \end{bmatrix}$$

is equivalent to that of finding the eigenvectors and eigenvalues of $A_1^\dagger A_1 - A_2^\dagger A_2$. Such a problem comes up in at least three distinct physical scenarios. One is the downdating problem, another is the so-called covariance differencing problem, and a third is array calibration. For a description of these problems and how thet arise in applications see [6].

In order to find the eigenvectors and eigenvalues of $A_1^\dagger A_1 - A_2^\dagger A_2$ without forming the outer products explicitly, a new decomposition called the Hyperbolic Singular Value Decomposition, the HSVD in short, was proposed in [6]. The HSVD is described in the following theorem.

Theorem: Let Φ be an $m \times m$ diagonal matrix, with entries ± 1 and let A be an $m \times n$ matrix, $m \geq n$, such that $A\Phi A^H$ is full rank. Then there exists an $n \times n$ unitary matrix U, and an $m \times m$ matrix V with

$$V^\dagger \Phi V = \hat{\Phi} \tag{14}$$

where $\hat{\Phi}$ is a diagonal matrix with entries ± 1 (possibly different from Φ), and an $n \times m$ diagonal matrix D with positive real diagonal entries, such that

$$A = VDU^\dagger. \tag{15}$$

□

From the HSVD of A we obtain that

$$A^\dagger \Phi A = UD\hat{\Phi}D^\dagger U^\dagger .$$

Hence the matrix U is the matrix of eigenvectors and the diagonal of $D\hat{\Phi}D^\dagger$ are eigenvalues of $A^\dagger \Phi A$.

One way of finding the HSVD in the case when $A\Phi A^H$ is full rank is via Hestenes method. Recall that the Hestenes technique [5] was originally designed for finding a unitary matrix U such that $W = AU$ has orthogonal columns. We outline this technique in some detail and then modify it to find the HSVD of A.

The Hestenes process of finding W and U is iterative and proceeds by constructing a sequence of matrices W_k , $k = 0, 1, \ldots$,

$$W_0 \equiv A,\ W_{k+1} = W_k G_k\ , \tag{16}$$

and a sequence of matrices U_k

$$U_0 \equiv I,\ U_{k+1} = U_k G_k \tag{17}$$

where G_k is a plane rotation matrix operating on columns $i = i(k)$ and $j = j(k)$ of W_k,

$$G_k = \begin{bmatrix} 1 & . & . & . & . \\ . & \cos\phi_k & . & -\sin\phi_k & . \\ . & . & 1 & . & . \\ . & \sin\phi_k & . & \cos\phi_k & . \\ . & . & . & . & 1 \end{bmatrix} .$$

The angles of rotations are chosen in such a way that the resulting columns become orthogonal. Equivalently, this is to say that the similarity transformation G_k on the symmetric matrix $W_k^\dagger W_k$ zeros its off-diagonal elements (i, j) and (j, i). The angle ϕ_k , $0 < |\phi| \leq \frac{\pi}{4}$, can be determined from the relation

$$\cot 2\phi_k = \frac{a_{i,i}^{(k)} - a_{j,j}^{(k)}}{2a_{i,j}^{(k)}}\ , \tag{18}$$

where $a_{i,j}^{(k)} = e_i^T (W_k^\dagger W_k) e_j$. Thus, the the Hestenes method for computing SVD is an implicit realization of the two-sided Jacobi method for computing the eigendecomposition of $A^\dagger A$.

By orthogonalizing all pairs of columns of W_k in a predescribed order called a sweep, and by iterating sweeps, the columns in the limit become orthogonal. In practice, the process terminates when columns of W_k are consider to be numerically orthogonal. Then, on one hand we have that

$$W_k = AU_k, \tag{19}$$

and on the other hand

$$W_k = V_k\Sigma_k, \tag{20}$$

where $\Sigma_k = diag\left(\sigma_i^{(k)}\right)_{i=1}^{n}$ and V has orthonormal columns. Thus, numerically, the factorization

$$A = V_k\Sigma_k U_k^{\dagger} \tag{21}$$

is an approximate SVD of A.

Now, if we insist that AU be hypernormal with respect to the matrix Φ, then $(AU)^{\dagger}\Phi(AU) = U^{\dagger}\Sigma_k^{\dagger}\hat{\Phi}\Sigma_k U$ will give the eigendecomposition of $A^{\dagger}\Phi A$, the precise decomposition that we were sought in the first place. The only difference in implementation is that the angles of rotations are chosen in such a way that, for a single rotation, the resulting rows become hypernormal. More precisely,

$$cot\phi_k = \frac{(w_i^{(k)})^H\Phi w_i(k) - (w_j^{(k)})^H\Phi w_j(k)}{(w_i^{(k)})^H\Phi w_j(k)}$$

where $w_i^{(k)} = W_k e_i$. Again, by applying rotations to all different pairs of rows in a sweep, and iterating the sweeps, the limit matrix itself becomes hypernormal.

As we can see, the Hestenes technique for computing the HSVD has essentially the same structure as the Hestenes technique for computing the SVD. The numerical properties of the Hestenes technique for computing the HSVD are the subject of the ongoing investigation.

4 Numerical Examples

In order to illustrate the usefulness of hyperbolic transformations in factoring differences of outer products we have conducted two numerical experiments. In the first experiment we compared the numerical accuracy of the eigenvalues of the difference of two matrix outer products computed by the two-sided Jacobi method applied explicitly to the difference $A_1^{\dagger}A_1 - A_2^{\dagger}A_2$ with the Hestenes method for computing the HSVD applied to the original data $A = [A_1^{\dagger}, A_2^{\dagger}]^{\dagger}$.

In the second experiment we compared the accuracy of finding the inverse of the difference of two matrix outer products $A_1^{\dagger}A_1 - A_2^{\dagger}A_2$ directly from the difference, indirectly via the triangular decomposition of $A = [A_1^{\dagger}, A_2^{\dagger}]^{\dagger}$, and via the HSVD of the data $A = [A_1^{\dagger}, A_2^{\dagger}]^{\dagger}$.

For the first experiment we formed an n by m matrix $\Psi \equiv [\mathrm{diag}(\lambda_1, ..., \lambda_n)|0]$ and defined the signature matrix Φ via $\Phi \equiv \mathrm{diag}(\{(-1)^i,\ i = 0, ..., n-1\})$. The eigenvalues of $\Psi\Phi\Psi^H$ are quite clearly $\lambda_1^2, -\lambda_2^2, ..., \lambda_n^2$. By picking a random $n \times n$ unitary U and a random $m \times m$ hypernormal (w.r.t Φ) V we can form

$$A = V\Psi U^{\dagger}$$

for which

$$A^\dagger \Phi A = U\Psi^\dagger V^\dagger \Phi V \Psi U = U\Psi^\dagger \Phi \Psi U^\dagger$$

has the same eigenvalues as $\Psi^\dagger \Phi \Psi$, but is now a full matrix.

We computed the eigenvalues of $A^\dagger \Phi A$ via hyperbolic Hestenes method which operated on the original data matrix A and next via the two-sided Jacobi method which operated on $A^\dagger \Phi A$. Simulations were conducted using MATLAB for which relative precision ϵ is 2^{-48}. For a given data matrix $A \equiv V_k \Psi U^\dagger$ we constructed the corresponding covariance matrix $A^\dagger \Phi A$. We chose $\Psi \equiv$ diag $(10^8, 10^4, 1)$, and generated the hypernormal matrix V_k as a product of k, $k = 1, 2, 3, 4, 6$, random hyperbolic Householder matrices. Note that the condition number of $A^\dagger \Phi A$ is 10^{16} which is comparable to the reciprocal of the relative precision used in the computations.

Let us denote the exact eigenvalues of $A^\dagger \Phi A$ as λ_i^E, the computed eigenvalues by Hestenes method as λ_i^H, and by Jacobi method as λ_i^J. In addition, let $\gamma_i^J \equiv \frac{\lambda_i^E - \lambda_i^J}{\lambda_i^E}$ and $\gamma_i^H \equiv \frac{\lambda_i^E - \lambda_i^H}{\lambda_i^E}$.

k	λ	γ^J	γ^H
1	λ_1	10^{-4}	10^{-9}
2		10^{-4}	10^{-8}
3		10^{-3}	10^{-8}
4		10^{-3}	10^{-8}
6		10^{0}	10^{-4}
1	λ_3	10^{-14}	10^{-13}
6		10^{-12}	10^{-12}

Table I.

The hyperbolic Hestenes method always gave better approximation of the eigenvalues than the Jacobi method, see Table I. However, the accuracy of the hyperbolic Hestenes was influenced by the number of terms in the product V_k and varied from simulation to simulation.

For the second experiments we generated a $k \times k$ random unitary U, random $n \times k$ V_1 and $m \times k$ V_2 such that $V_1^\dagger V_1 = I_k = V_2^\dagger V_2$. Next we picked diagonal matrices $\Sigma_1 = diag(\sigma_i^{(1)})$ and $\Sigma_2 = diag(\sigma_i^{(2)})$. Now by letting

$$\Sigma = \Sigma_1 - \Sigma_2 \,, \quad X = U\Sigma U^\dagger \,,$$

$$A_1 = V_1 \Sigma_1 U^\dagger \,, \quad A_2 = V_2 \Sigma_2 U^\dagger \,,$$

we got the test matrix

$$X = A_1^\dagger A_1 - A_2^\dagger A_2 \,.$$

Two tests were performed. In one the inverse of X was computed, in the other the eigenvalues of X were computed.

The inverese of X was computed in four different ways:

- $Xinv = U\Sigma^{-1}U^\dagger$ was considered to be the "true" inverse.

- The inverse $covXinv$ was computed directly from the difference of the covariances, $covXinv = (A_1^\dagger A_1 - A_2^\dagger A_2)^{-1}$.
- The triangular decomposition $X = R^\dagger DR$ was computed using hyperbolic Householder transformations operating on $Y = [A_1^\dagger \; A_2^\dagger]^\dagger$. Next the inverse $HHXinv$ was calculated as $HHXinv = R^{-1}DR^{-\dagger}$.
- The HSVD $[A_1^\dagger \; A_2^\dagger]^\dagger K = H$, where K is orthogonal, H is Φ-orthogonal, was computed via Hestenes method. The inverse $HSVDXinv$ was calculated as $HSVDXinv = K(H^\dagger \Phi H)^{-1} K^\dagger$.

By picking $(\sigma_i^{(1)})_{i=1}^k$ and $(\sigma_i^{(1)})_{i=1}^k$ different test matrices were generated. The test matrices were divided into five categories as illustrated by Table II.

class	cond$(A_1^\dagger A_1)$	cond$(A_2^\dagger A_2)$	cond$(A^\dagger \Phi A)$
I	O(1)	O(1)	O(1)
II	O(1)	O(1)	$O(10^{14})$
III	$O(10^{14})$	$O(10^{14})$	O(1)
IV	$O(10^{14})$	$O(10^{14})$	$O(10^{14})$
V	$O(10^{14})$	O(1)	$O(10^{14})$

Table II.

For each method the relative errors with respect to the norm of $Xinv$ were recorded. Typical results of this test are summarized in Table III below.

class	$e_{covXinv}$	e_{HHXinv}	$e_{HSVDXinv}$
I	6.0e-16	2.0e-16	8.0e-16
II	1.3e-2	8.0e-3	1.1e-2
III	1.1e-2	7.0e-9	1.2e-9
IV	6.7e-3	1.7e-2	1.1e-2
V	4.5e-2	1.3e-9	1.2e-9

Table III.

In the second experiment the eigenvalues were computed in three different ways:

- $trueeig = diag(\Sigma)$ were considered to be the true eigenvalues
- $coveig$ were the eigenvalues computed via two-sided Jacobi method directly from $A_1^\dagger A_1 - A_2^\dagger A_2$
- $HSVDeig = diag((H^\dagger \Phi H)$ were the eigenvalues computed form the HSVD of Y

The magnitude of the relative errors was analogous to that for the inverse of X.

The numerical results indicate that for class III and class V the methods that operated on the original data via hyperbolic type transformations produced better numerical results than the

methods that operated on the explicit difference of the outer products. Theoretical backing for this improved accuracy remains a topic for future investigation.

We feel confident that there are many more applications within and beyond digital signal processing or control where the hyperbolic transformations will be useful for its numerical stability, fast computational characteristics, and as a theoretical structure.

References

[1] A.W. Bojanczyk and A. Steinhardt, "Matrix Downdating Techniques for Signal Processing", *Proceedings of the SPIE Conference on Advanced Algorithms and Architectures for Signal Processing III*, vol 975, pp 68-75, 1988.

[2] G. Cybenko and M. Berry, "Hyperbolic Householder Algorithm for Factoring Structured Matrices", *SIAM, J. Matrix Anal. Appl.*, vol 11, pp 499-520, 1990.

[3] I. Gohberg, P. Lancaster, and L. Rodman, *Matrices and Indefinite Scalar Products*, Birkhauser Verlag, Basel, Switzerland, 1983.

[4] G.H. Golub and C. Van Loan, *Matrix Computations*, John Hopkins Press, Baltimore, MD, 1983.

[5] M.R. Hestenes, "Inversion of matrices by biorthogonalization and related results", *J. Soc. Indust. Appl. Math.*, vol 6, pp 51-90, 1958.

[6] R. Onn, A. Steinhardt, and A. Bojanczyk, "Hyperbolic Singular Value Decomposition and Its Applications", *IEEE Trans. on SP*, pp 1575-88, July 1991.

[7] C. Rader and A. Steinhardt, "Hyperbolic Householder Transformations", *IEEE Trans. Acoust., Speech, Signal Proc.*, Dec. 1986.

[8] A. Steinhardt, "Householder Transformations in Signal Processing", *IEEE ASSP Magazine*, July, 1988.

Iterative Algorithms for Integral Equations of the First Kind

Mark G. Vangel
U. S. Army Materials Technology Laboratory
SLCMT-MRS-MM, Arsenal St., Watertown, MA 02172-0001

Abstract

Integral equations of the first kind are usually *ill-posed*, that is, they have solutions which do not depend continuously on the right hand side. When solving these equations numerically, roundoff error is introduced in the right hand side, and even this small change can cause very large changes in the solution of the numerical problem. This problem is made even worse when the right hand side is observed with error, i.e. for ill-posed *inverse problems*.

It is the purpose of this paper to do two things. First, we point out that, for a certain class of problems, simple Richardson iteration can provide a numerically stable means of approximately solving an integral equation of the first kind numerically. However, Richardson's algortihm can converge very slowly. We therefore also discuss a *preconditioned* Richardson algorithm, which can greatly accelerate convergence and which has a natural probabilistic interpretation when applied to equations with positive, bounded kernels.

1 Introduction

Consider the following integral equation:

$$\int_0^1 k(x,y)f(y)dy = g(x). \tag{1}$$

We will illustrate the main ideas of this paper by means of two examples.

For the first example, we take $k(x,y)$ to equal

$$k_1(x,y) = \begin{cases} y(1-x) & \text{for } 0 \le y < x \\ x(1-y) & \text{for } x \le y \le 1 \end{cases}, \tag{2}$$

and the right hand side to be

$$g_1(x) = x^2(1-x)^2. \tag{3}$$

This equation is classified as a *Fredholm* integral equation of the *first kind.* It is an equation of the first kind because the unknown function, f, appears only in the integrand. It is a Fredholm equation because the limits of integration are constant. The function $k(x,y)$ is called the *kernel* of the equation. Some important features of the kernel chosen for this example are that it is continuous, bounded, and peaked along the line $x = y$.

For the second example, we take $k(x,y)$ to equal

$$k_2(x,y) = \begin{cases} 1 & \text{for } 0 \le y < x \\ 0 & \text{for } x \le y \le 1 \end{cases}, \tag{4}$$

and the right hand side $g_2(x)$ to be an arbitrary bounded, differentiable function. With the kernel (4), the equation (1) has upper limit of integrtion x. An equation of this form is called a *Volterra* integral equation of the first kind.

When treating the Volterra equation with kernel (4) numerically, we can work with the equivalent equation

$$\int_0^1 x f(xy)dy = g(x). \tag{5}$$

The limits of integration for (5) do not depend on x, so a single set of fixed quadrature points can be used.

The solutions of the Fredholm equation with kernel (2) is

$$f_1(x) = -\frac{d^2g(x)}{dx}, \tag{6}$$

and the solution of the Volterra equation with kernel (4) is

$$f_2(x) = \frac{dg(x)}{dx}. \tag{7}$$

Note that both of these solutions involve differentiation of the given function, and that numerical differentiation is notoriously difficult.

2 The Ill-Posed Nature of Integral Equations of the First Kind

We might first approach these problem by, naively, approximating equation (1) as a matrix equation and solving this equation directly. For example, let

$$y_1 < y_2 < \ldots < y_n$$

be Gauss-Legendre quadrature points with corresponding weights $\{w_i\}_1^n$, and choose a mesh of values $\{x_i\}_1^n$ with $x_i = y_i$ for each i. For the present discussion, we will take n to equal 50.

Consider the *matrix* equation

$$Kf = g, \tag{8}$$

where the typical element of K is

$$k_{ij} \equiv k(x_i, y_j)w_j, \tag{9}$$

and the ith element of g is

$$g_i \equiv g(x_i). \tag{10}$$

Let the singular values corresponding to K be denoted $\{\sigma_i^K\}_{i=1}^n$, where we can omit the superscript when the matrix to which the singular values correspond is clear from the context. We denote the matrices corresponding to discretizations with kernel (2) and (4) as K_1 and K_2, respectively.

The condition number of K_1 is

$$\kappa(K_1) \equiv \frac{\sigma_1^{K_1}}{\sigma_n^{K_1}} = \frac{.1019671}{6.525424\text{x}10^{-7}} = 1.562612\text{x}10^5. \tag{11}$$

This matrix equation is ill-conditioned, and noise in the computer representation of g can result in a noisy 'solution' to the matrix equation which is *very* different from the discretized solution to the continuous problem (1).

Actually, the direct solution of

$$K_1 f = g \tag{12}$$

is feasible, using double precision arithmetic and a good Gaussian elimation or singular value decomposition algorithm. But in general this is *not* the approach to take for integral equations of the first kind.

These matrix equations are *ill-conditioned* because the corresponding integral equations are *ill-posed*: small changes in g can cause large changes in the solution f. This is intuitively reasonable because the process of integration, with respect to a reasonably smooth kernel, will tend to produce a result which is 'smoother' than the integrand. In solving the equation, we are inverting this smoothing process, and so we encounter the difficulties associated with numerical differentiation.

3 Regularization Methods

One approach to solving the equation (1) is the *method of regularization* of Tikhonov (1962) and Phillips (1963) (see also Tikhonov and Arsenin, 1977, and Groetsch 1984). The basic idea is very simple. Because the integral equation (1) is ill-posed, we do not want to solve any discretized version of this equation exactly. Rather, we would like to find a *smooth* function which *nearly* satisfies the equation. So, instead of solving the matrix equation (8), we minimize the quadratic form

$$U(z) \equiv (z - f)^{\mathrm{T}}(z - f) + \lambda z^{\mathrm{T}} L z, \tag{13}$$

where L is positive semi-definite, and is chosen so that $z^{\mathrm{T}}Lz$ will tend to be large when z is not smooth. A positive constant, λ, determines the relative importance of the first (*least-squares*) and second (*penalty*) terms of the functional $U(z)$. When λ is small, then the minimum will occur near an exact solution f. As λ is increased, increasing weight is put on the smoothness of the solution, and less on 'fidelity' to the equation (8).

4 Richardson's Algorithm and Implicit Regularization

Another approach to solving ill-conditioned linear matrix equations is by iteration. For the discussion in this section, we will consider (8), where K is positive definite and $\kappa(K)$ is large enough for direct

solution without regularization to not be a viable approach. We choose an arbitrary first approximation f^0, and define the iteration

$$\delta^k = f^{k+1} - f^k = \theta B^{-1}(g - Kf^k), \tag{14}$$

where θ is a positive constant and B is a *peconditioning* matrix chosen to accelerate convergence. When $B = I$, (14) is the well known *Richardson algorithm*, first proposed in Richardson (1910) for the solution of sparse linear systems of equations. The kth approximation to the solution can be written as

$$f^k = \textstyle\sum_{i=0}^{k-1} \delta^i \quad \text{for } k > 0. \tag{15}$$

It is easy to show that f^k converges to a solution f for arbitrary right hand side g if, and only if, all of the eigenvalues of $I - \theta B^{-1}K$ are within the unit circle.

Another feature which is clear from the form of (14) is that if K is acting as a *linear smoother*, then the iteration should be numerically stable, at least for the initial iterates. If a matrix is obtained from an integral equation, and if the kernel of this equation is bounded and not highly oscillatory, then this matrix will act as a smoother. Both (2) and (4) meet these criteria.

Let f be the solution to (8), and define the difference between the kth approximation and this solution as

$$u^k \equiv f - f^k, \tag{16}$$

so that

$$\delta^k = \theta B^{-1}(g - Kf^k) = \theta B^{-1}K(f - f^k) = \theta B^{-1}Ku^k. \tag{17}$$

Let K^{-1} be the inverse of K, and define the quadratic form

$$Q(z) \equiv Q_{\mathrm{LS}}(z) + Q_{\mathrm{P}}(z) \equiv (u^k - z)^{\mathrm{T}}(u^k - z) + z^{\mathrm{T}}(K^{-1}B/\theta - I)z. \tag{18}$$

Differentiating $Q(z)$ with respect to z, and using the fact that K is positive definite, we observe that

$$\min_z Q(z) = Q(\delta^k). \tag{19}$$

Note the similarity between (13) and (18). We have shown that each step (14) corresponds to solving a *penalized least squares* problem,

where the penalty term is determined by the kernel K. Further discussion of the relationship between linear smoothers and penalized least squares can be found in Buja, et. al. (1989).

Although (14) does not make explicit use of regularization, at each iteration regularization is *implicit* in this algorithm and the character of this regularization is determined by the kernel itself. To see how the second term in (18) can penalize 'rough' iterates, assume that K is symmetric with (positive) eigenvalues λ_i and corresponding eigenvectors t_i, that is

$$K = \sum_{i=1}^{n} \lambda_i t_i t_i^{\mathrm{T}}, \tag{20}$$

where

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n > 0.$$

Let the expansions of δ^k in terms of these eigenvectors be

$$\delta^k = \sum_{i=1}^{n} \beta_i^k t_i. \tag{21}$$

We will assume further that (8) has been scaled so that $\lambda_1 \leq 1$, and we take $\theta = 1$ and $B = I$. In terms of the spectral decomposition (20) of K, the penalty term (at the minimum) becomes

$$Q_{\mathrm{P}}(\delta^k) = \delta^{k\mathrm{T}}(K^{-1} - I)\delta^k = \sum_{i=1}^{n} (\beta_i^k)^2(\lambda_i^{-1} - 1). \tag{22}$$

Since the matrix K is a discretization of a smooth function, the more oscillatory eigenvectors will correspond to small eigenvalues. Components of δ^k in the directions of these highly oscillatory eigenvectors will have a large contribution in the penalty term, hence the minimum of Q will tend to occur at a vector δ^k which has small components in the direction of the 'rougher' eigenvectors – that is, δ^k will tend to be smooth if K is smooth.

If the algorithm (14) is convergent, and if the matrix equation (8) is sufficiently ill-conditioned, then as the iterates approach the solution they will eventually become noisy and meaningless. However, the rate of convergence of Richardson's algorithm for this problem can be easily shown to be governed by the powers $(1 - \lambda_i)^k$. Once the (smoother) components in the direction of the the largest eigenvalues have been

nearly determined, the convergence rate will become very slow. A practical implication of this is that the iteration eventually becomes useless, often before instability in the solution becomes noticable.

Iteration, therefore, is equivalent to regularization with the reciprocal of the number of terms taken in the iteration corresponding to the smoothing parameter. This observation was apparently first made by Bakushinskii (1967).

5 A Preconditioned Richardson Algorithm

Consider the integral equation (1), where we assume that the kernel, $k(x,y)$, is positive and bounded. We transform the equation (1) into a new equation, having the same solution, as follows:

$$\int_0^1 \tilde{k}(x,y)f(y)dy = \tilde{g}(x), \tag{23}$$

where

$$\tilde{k}(x,y) \equiv \frac{k(x,y)}{\int_0^1 k(x,y)dy}, \tag{24}$$

and

$$\tilde{g}(x) \equiv \frac{g(x)}{\int_0^1 k(x,y)dy}. \tag{25}$$

We now discretize (23) as discussed above, and apply the Richardson iteration (14) with $\theta = 1$.

If we let K denote the matrix in the discretization of (1), and $\tilde{K}$ denote the corresponding matrix from (23), we have

$$\tilde{K} = B^{-1}K, \tag{26}$$

where B is a diagonal matrix with ith element equal to the sum of the elements in the ith row of K. The preconditioned matrix $\tilde{K}$ is thus a *stochastic* matrix, and by the Perron-Frobenius theorem (e.g., Horn and Johnson, 1989), $\tilde{K}$ has largest eigenvalue equal to one, and all other eigenvalues on or within the unit circle.

There are several ways of motivating this particular choice of a preconditioning matrix. From the point of view of numerical analysis,

scaling a matrix in this way tends to make the matrix better conditioned. The following is a special case of a theorem proved by Van der Sluis (1969, p.18):

Theorem 5.1 *Let K be a nonsingular matrix, and let $\|\cdot\|_*$ be any Hölder norm, or the Frobenius norm. Let D be a diagonal matrix. Then the following measures of the condition of DK are minimized when the rows of DK each sum to one:*

- $\chi_1(DK) \equiv \|DK\|_\infty \|(DK)^{-1}\|_*$, *and*
- $\chi_2(DK) \equiv \|DA\|_\infty / \|DA\|_*$.

Although χ_1 and χ_2 each differs from the usual condition number κ, all three quantities are reasonable measures of the condition of a matrix. A preconditioning which minimizes χ_1 and χ_2 can be expected to usually reduce κ as well.

A simple probabilistic argument provides another motivation for scaling the rows to sum to one. Since k is bounded and positive, it is propotional to the joint density of two random variables, say X and Y. We write this as

$$\pi_{X,Y}(x,y) \equiv ck(x,y), \tag{27}$$

where the constant c is

$$c = \left[\int_0^1 \int_0^1 k(x,y)dxdy\right]^{-1}. \tag{28}$$

The normalized kernel (24) is exactly the *conditional density* of the random variable Y given the random variable X:

$$\pi_{Y|X}(y|x) = \frac{\pi_{X,Y}(x,y)}{\int_0^1 \pi_{X,Y}(x,y)dy} = \tilde{k}(x,y). \tag{29}$$

Richardson's algorithm applied to (23) with $\theta = 1$ is

$$f^{k+1}(x) = f^k(x) + \int_0^1 \tilde{k}(x,y)(f(y) - f^k(y))dy. \tag{30}$$

Since the integral on the right hand side of (30) can be interpreted as the conditional expectation of the difference $f - f^k$, we can rewrite (30) (in terms of the *random variables X and Y*) as

$$f^{k+1}(X) - f^k(X) = E\left[f(Y) - f^k(Y)\,|X\right]. \tag{31}$$

In words, the kth step in this preconditioned Richardson algorithm (with $\theta = 1$) is the conditional expectation of the difference between the solution and the approximation f^k.

This probabilistic interpretation suggests that the preconditioned Richardson algorithm will converge rapidly when the conditional expectation, with respect to the density (29), of $f - f^k$ is nearly equal to $f - f^k$. For this to occur, X and Y must be *correlated* random variables – the more highly X and Y are correlated, the closer $f^{k+1} - f^k$ will be to $f - f^k$. For these random variables to be correlated, the original kernel $k(x, y)$ must be peaked about the line $x = y$. The more highly the kernel is peaked, the more rapidly convergent the preconditioned Richardson algorithm will be. The limiting case of perfect correlation (i.e. $X = Y$) is achieved by the δ-function kernel

$$k(x, y) = \delta(x - y). \tag{32}$$

6 A Fredholm Example

We now illustrate the above discussion with two examples. First we consider the Fredholm integral equation of the first kind with kernel (2) and right hand side (3). We discretize the problem using 50 point Gauss-Legendre quadrature as discussed in Section 2. The largest eigenvalue for the matrix equation (8) is .1013913, which is approximately equal to π^{-2}, the largest eigenvalue of the corresponding integral equation. For the Richardson iteration *without* preconditioning ($B = I$), we take θ to equal the reciprocal of the largest eigenvalue, i.e. $\theta \approx 9.863$, so that the largest eigenvalue of θK is equal to 1. For the *preconditioned* Richardson algorithm, the largest eigenvalue is approximately 1, so we let $\theta = 1$. Fifty iterations of both methods are compared in Figure 1. The preconditioned method gives an approximation very near the solution

$$f(x) = -12x^2 + 12x - 2 \tag{33}$$

before the convergence rate begins to decrease dramatically. The method without preconditioning is still far from the solution at the 50th iteration, and, since by the 50th iteration the steps taken at each iteration are very small, it will take many iterations to get appreciably closer to the solution.

Another way of seeing the dramatic effect preconditioning has had on the convergence rate is to examine the distance, in L_2 norm, to the solution as a function of the iteration index. This comparison is made in Figure 2.

Both of the Richardson algorithms are numerically stable, which we would expect given the discussion in Section 3. We would expect that eventually the approximations will become less smooth, as the components in the directions of eigenvectors corresponding to smaller eigenvalues begin to have an effect. Since the right hand side for this example is smooth, and since preconditioning has reduced the condition number substantially (to 810.34), it would require an unreasonable number of iterations to observe the approximations depart from the true solution, and even then the deviation would be slight. In order to see an effect in a reasonable number of iterations, we added a component, with coefficient .01, in the direction of the 25th singular vector of the matrix K_1 to the right hand side (3). The Fourier coefficients of the perturbed right hand side are presented in Figure 3, and a plot of this perturbed function is given in Figure 4. In Figure 5, we display 50 iterations of the preconditioned algorithm with the perturbed right hand side, and in Figure 6 we give the L_2 distance to the solution (33) as a function of the number of iterations. Notice that the approximations are closest in norm to this solution at the 10th itertation. From that point on, the iterations move further away from the solution which corresponds to the *unperturbed* right hand side as they approach the exact soltuion, which corresponds to the *perturbed* right hand side.

7 A Volterra Example

As an example of a Volerra equation, we take the numerical differentiation problem with kernel (4). This example is useful because it is easy to examine the nature of the 'implicit regularization' analytically.

To precondition the kernel, we divide by

$$\int_0^1 k(x,y)dy = \int_0^x dy = x. \tag{34}$$

A little algebra shows that, if $g(x) = x^{s+1}/(s+1)$, then f^k is given by

$$f^k(x) = [1-(1-1/(s+1))^k]x^s. \tag{35}$$

Without this preconditioning, it is easy to show that the Richardson iteration does not converge for *any* θ.

Assume that the right hand side of this Volterra equation has a convergent Taylor series expansion:

$$g(x) = \sum_{s=0}^{\infty} a_s x^s \tag{36}$$

From the linearity of the Volterra integral operator and (35) we see that

$$f^k(x) = \sum_{s=1}^{\infty} s a_s [1 - (1 - 1/s)^k] x^{s-1}. \tag{37}$$

If g is a smooth function plus noise, then f^k will reflect the smooth components initially, since these will correspond to fairly small values of s. Eventually, the solution will become rougher, but only when $(1 - 1/s)^k$ becomes small for fairly large s.

Numerical experimentation suggests that, for reasonably smooth right hand sides, the iterative algorithm outlined in this section can be useful for numerical differentiation.

Acknowledgements

The author is grateful to Donald M. Neal of the Army Materials Technology Laboratory and to Professors Herman Chernoff and Donald G. M. Anderson of Harvard University for many helpful discussions.

References

[1] Buja, A; Hastie, T; and Tibshirani, R (1989), "Linear Smoothers and Additive Models", *Annals of Statistics*, 17, 453-555.

[2] Bakushinskii, A. B. (1967), "A General Merhod of Constructing Regularizing Algorithms for a Linear Ill-Posed Equation in Hilbert Space", *U. S. S. R. Computational Mathematics and Mathematical Physics*, 7, 3, 279-286.

[3] Groetsch, C. W. (1984), *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Marshfield, Massachusetts.

[4] Horn, R. A. and Johnson, C. R. (1985), *Matrix Analysis*, Cambridge University Press, Cambridge.

[5] Phillips, D. L. (1962), "A Technique for the Numerical Solution of Certain Integral Equations of the First Kind", *J. of the Association of Computing Machinery*, 9, 84-97.

[6] Richardson, L. F. (1910), "The Approximate Arithmetical Solution by Finite Differences of Physical Problems Involving Differential Equations With an Application to Stresses in a Masonry Dam", *Phil. Trans. Roy. Soc. Lond.*, A, 210, 307-357.

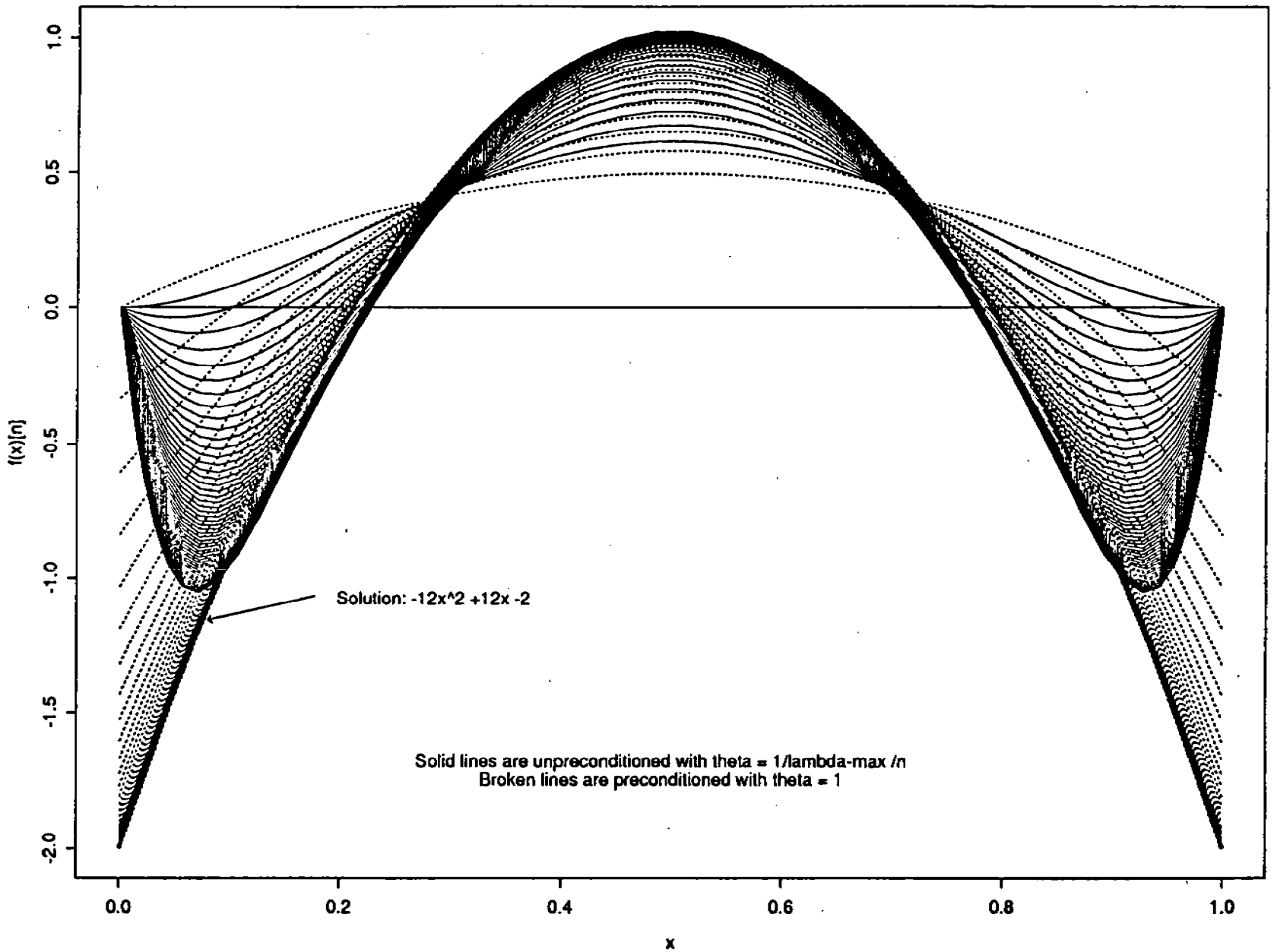
[7] Tikhonov, A. N. (1963), "Regularization of Incorrectly Posed Problems", *Soviet Math. Doklady*, 4, 1624-1627.

[8] Tikhonov, A. N. and Arsenin, V. Y. (1977), *Solutions of Ill-Posed Problems*, Wiley, New York.

[9] Van der Sluis, A (1969), "Condition Numbers and Equilibration of Matrices", *Numer. Math.*, 14, 14-23.

Fifty iterations of Richarson with and without preconditioning

Figure 1



Comparison of convergence rates for Richarson with and without preconditioning

Figure 2

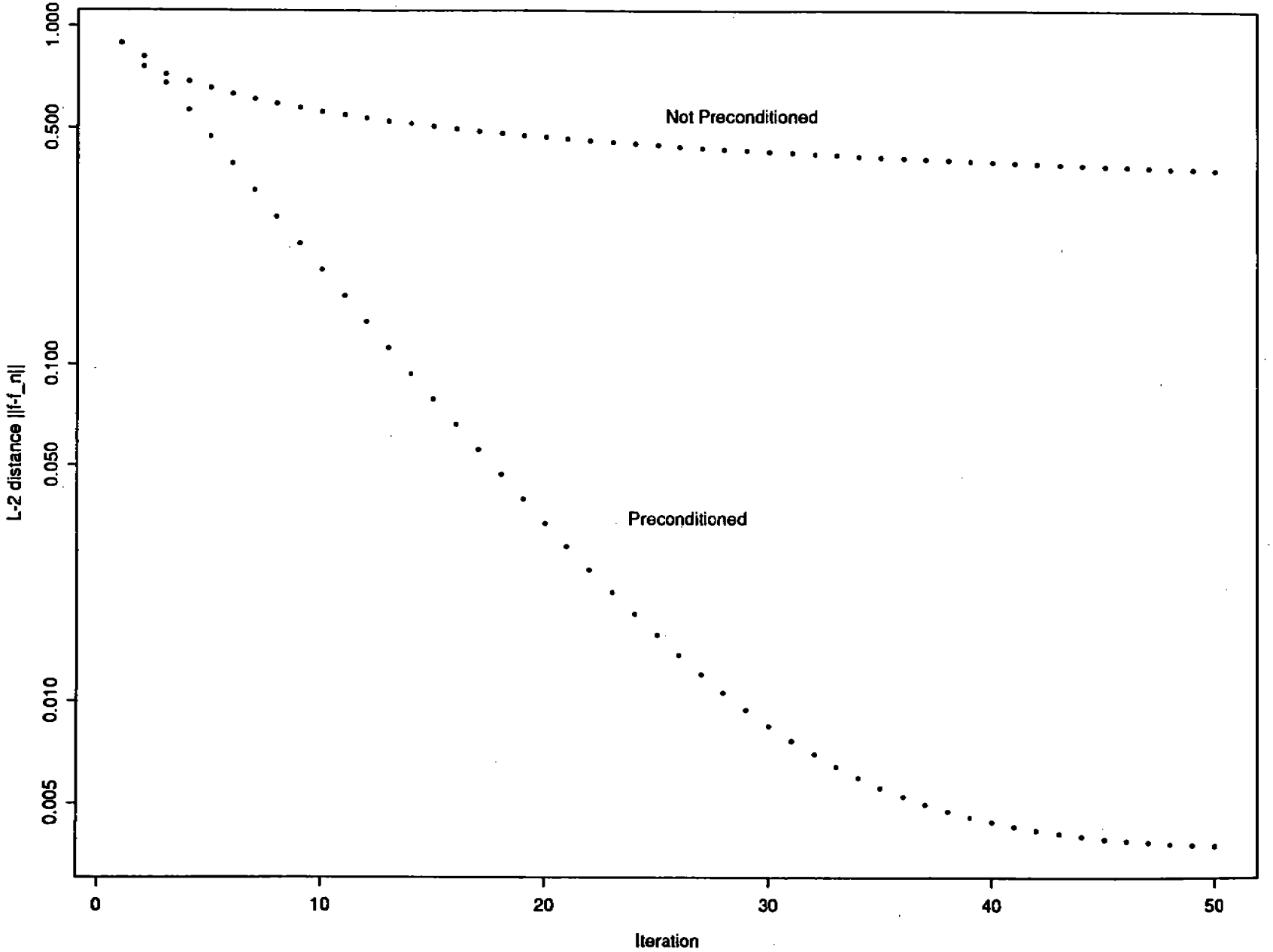


Figure 3

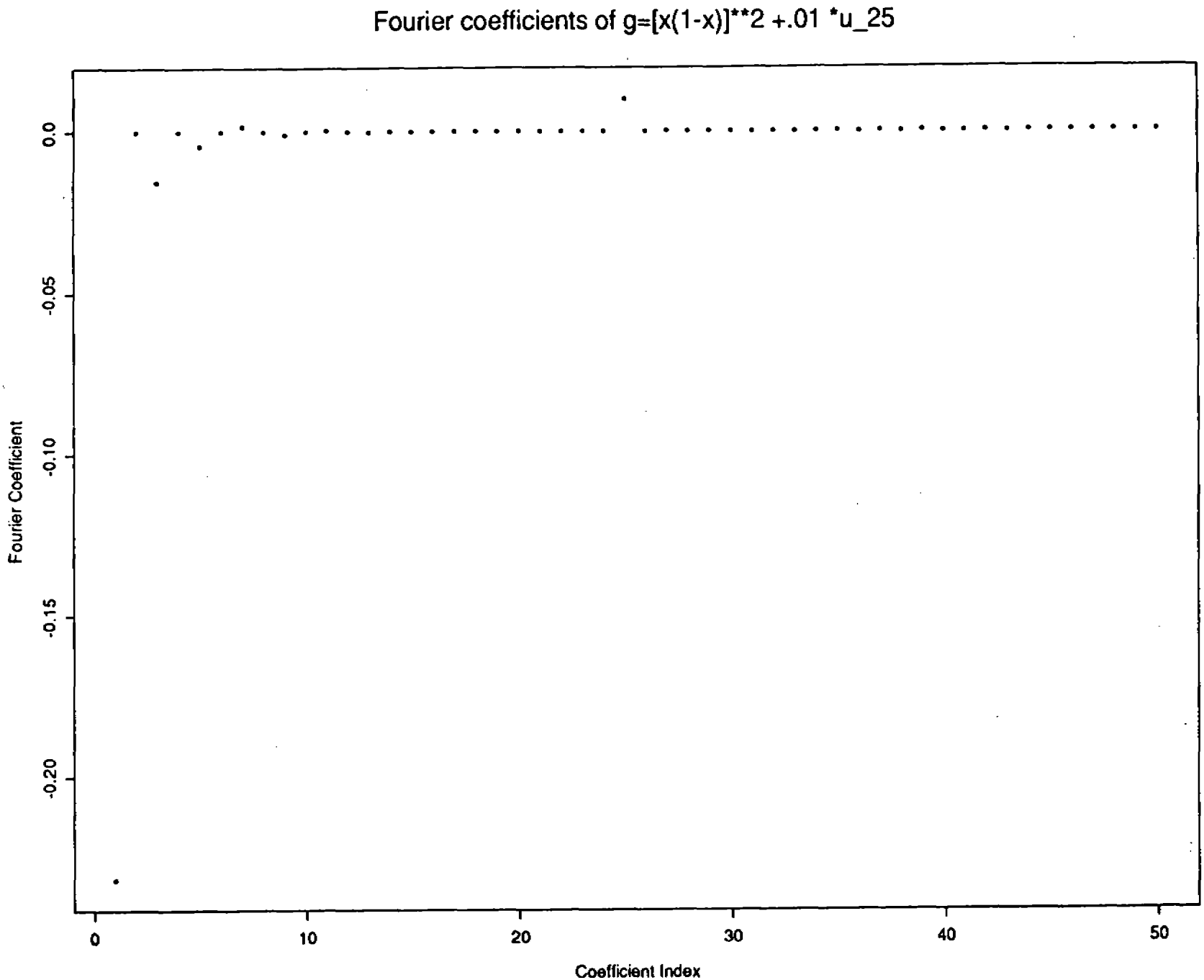

Fourier coefficients of g=[x(1-x)]**2 +.01 *u_25
Fourier Coefficient
0.0
-0.05
-0.10
-0.15
-0.20
0
10
20
30
40
50
Coefficient Index

Figure 4

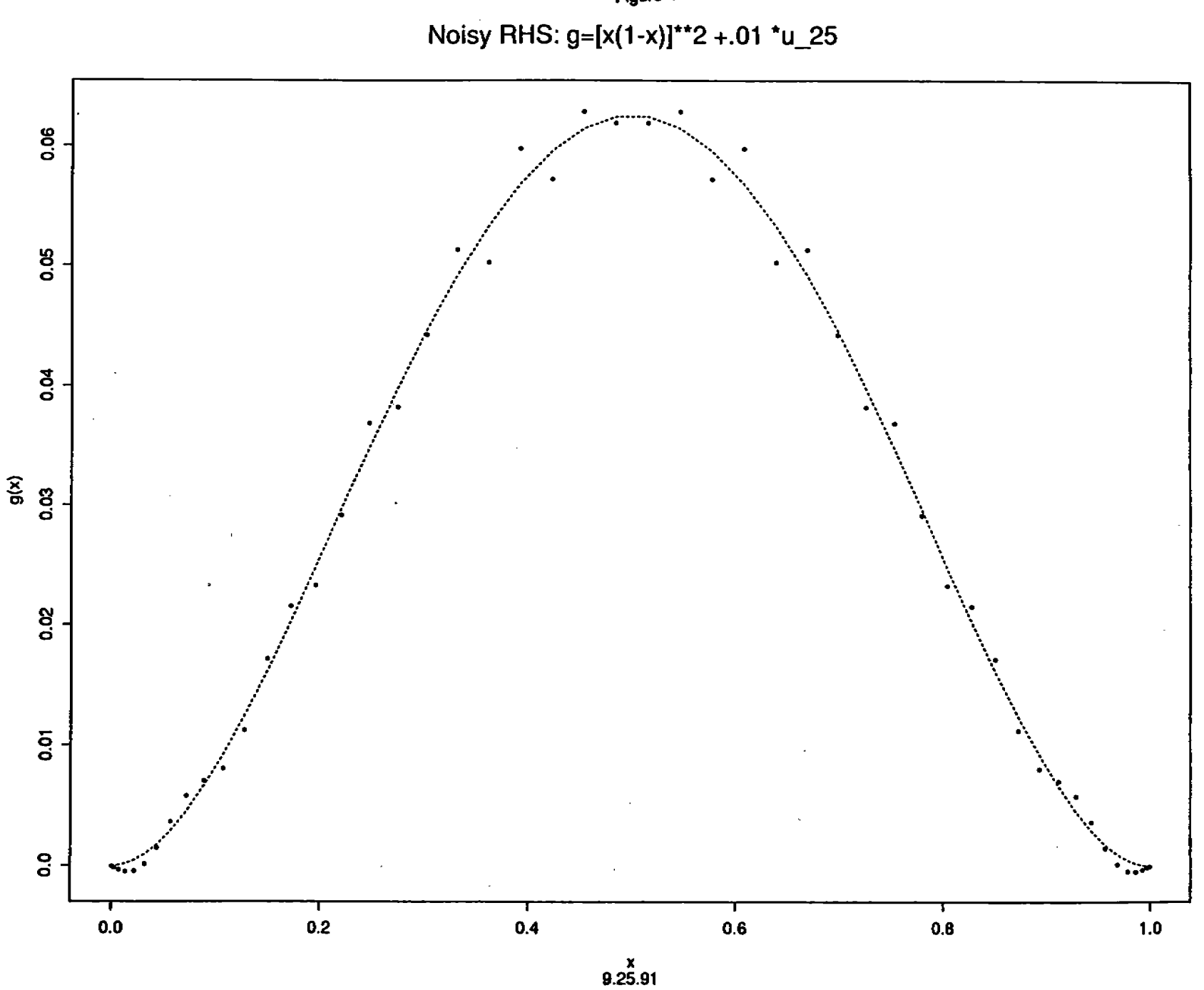

Noisy RHS: g=[x(1-x)]**2 +.01 *u_25
g(x)
0.0
0.01
0.02
0.03
0.04
0.05
0.06
0.0
0.2
0.4
0.6
0.8
1.0
x
9.25.91

Figure 5

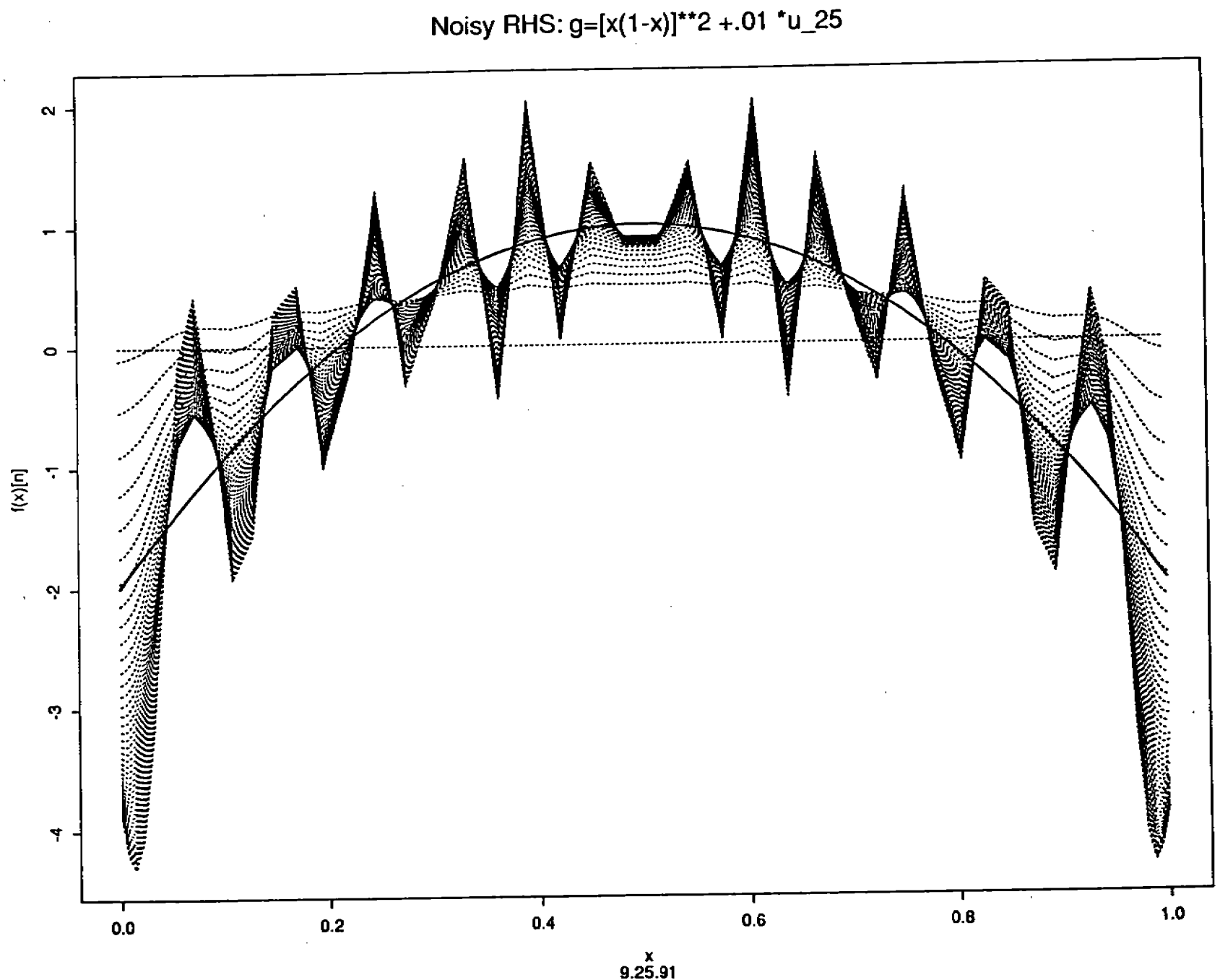
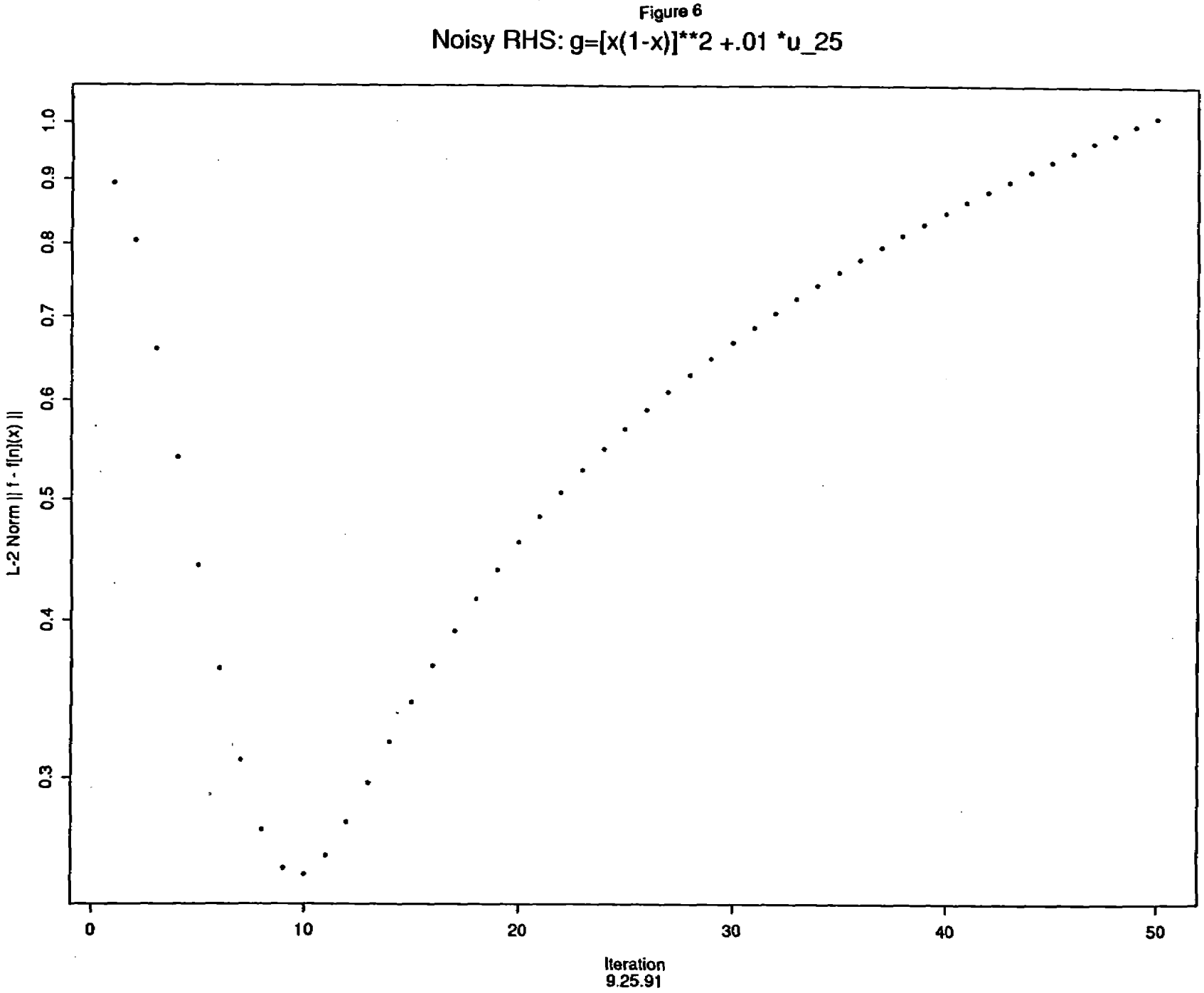


Figure 6


Noisy RHS: g=[x(1-x)]**2 +.01 *u_25
L-2 Norm || f - f[n](x) ||
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
0
10
20
30
40
50
Iteration
9.25.91

ON THE ANALYSIS OF SUPERHARMONIC OSCILLATIONS[1]

J. J. Wu
US Army Research Office
Research Triangle Park, NC 27709

ABSTRACT

This paper presents an analysis for the superharmonics of a forced nonlinear vibration problem involving small parameters, using a generalized harmonic balance method. A nonlinear ordinary differential equation with several nonlinear terms and a periodic forcing function is considered. For the case of superharmonic oscillations of order 2, the key equations for the obtaining the information on the superharmonics will be derived, including a new, nonlinear ordinary differential equation of a slow varying function compared with the original dependent variable. Using these equations, the steady state solution and its stability behavior can be calculated. Results for a special set of parameters are obtained, including a stable node for the steady state solution and the associated van del Pol plane.

[1] The original version of this paper appeared in the Proceedings of the 13th World Congress on Computation and Applied Mathematics (IMACS '91, July 22-26, 1991, Trinity College, Dublin, Ireland), pp. 918-920 (Vol. 2), Criterion Press, Ireland, 1991. Many typographical errors in the original paper have been corrected here.

1. INTRODUCTION

It is well known that nonlinearities can cause sub- and super-harmonic excitations in vibratory systems. The analytical understanding of such phenomena is often difficult to obtain. It has been shown that the method of multiple scales can be used to solve such problems as demonstrated in several papers by Nayheh [1,2]. However, the procedures involved are quite complicated and requires recursive solution of differential equations, the elimination of secular terms and reconstitution, all of which are nontrivial procedures. More recently, in a paper by Noble and Hussain [3], an expansion method was introduced together with suggestions of several other approaches which may be used as alternatives to obtain pertinent information. One of these is the genralized harmonic balance method (GHB) [4,5,6]. This variant of the harmonic balance method consists of two parts: first, to derive the form of solution using only the basic steps of multiple scales, and then, solve for the coefficients of various harmonics. In this approach, the elimination of the secular terms is accomplished implicitly, thus avoiding the trouble of solving recursive differential equations. This paper begins with a general nonlinear ordinary differential equation with several nonlinear terms and a periodic forcing function, a specific case of superharmonic oscillations of order 2 will be investigated. Next, the key equations are derived, from them the essential information on the superharmonics can be obtained. Finally Numerical results are presented on the steady solution and the stability behavior for a special sets of parameters.

2. DERIVATION OF THE KEY EQUATIONS

We shall consider the following rather general differential equation:

$$d^2u/dt^2+u+2\varepsilon\mu(du/dt)+\varepsilon\alpha_2u^2+\varepsilon^2\alpha_3u^3$$

$$+\varepsilon\alpha_4(du/dt)^2+\varepsilon^2\alpha_5u(du/dt)^2=2f\cos(\Omega t) \qquad (1)$$

where u(t) is the unknown function μ and α_k, k=2,3,4,5 and 6, are given constants, ε is the small perturbation parameter; f and Ω pertain to the magnitude and frequency of the forcing function. For superharmonics of order 2, one has

$$2\Omega=\omega=\omega_0+\varepsilon\sigma=1+\varepsilon\sigma \qquad (2)$$

where ω is the "fundamental" frequency of the nonlinear vibration, which is a perturbation from that of the linearalized system ω_0, taking to be unity in (2) without a loss of generality. We shall derive a two-term approximate solution $u=u_0+\varepsilon u_1$ for equation (1). Using a procedure described previously in [4,5], it can be shown easily that that the final form of the solution u, which is good to the order of ε must have the following form:

$$u=\varepsilon U_0+[(U_1A+U_2A^2)+\varepsilon(U_3A^3+U_4A^4)+cc] \qquad (3)$$

where cc stands for the complex conjugate. The following symbols are introduced:

$$A=\exp(it/2), \quad S=\exp(i\varepsilon\sigma t/2) \tag{4}$$

Eq. (1) can then be written as

$$d^2u/dt^2+u+2\varepsilon\mu(du/dt)+\varepsilon\alpha_2u^2+\varepsilon^2\alpha_3u^3$$

$$+\varepsilon\alpha_4(du/dt)^2+\varepsilon^2\alpha_5u(du/dt)^2=fSA^2+cc \tag{1'}$$

Here we note that S is a slow varying function compared with A in the sense that while dA/dt is of O(1), dS/dt is of O(ε). Since we are using the small parameter ε as a means to identify quantities with different order of magnitudes, it is assumed that all the symbols other than ε are of O(1) unless stsated otherwise. We shall also use the fact that

$$\bar{A}=\varepsilon^{-i\sigma/2}, \text{ and } \quad \bar{A}A=1 \tag{5}$$

where an overbar denotes the complex conjugate. The procedure here is to substitute (3) in (1') and set to zero the coefficients of A_k, k=0,1 and 2, since any higher harmonics will be of $O(\varepsilon^2)$ or higher according to (3). We first obtain the following approximate expressions (in other words, the right hand side should have added "+ terms of $O(\varepsilon^3)$ and higher" in each of these equations):

$$du/dt=(dU_11/dt+iU_1)A$$

$$+\varepsilon[dU_0/dt+(dU_2/dt+2iU_2)A^2]+cc \tag{6}$$

$$d^2u/dt^2=\varepsilon d^2U_00/dt^2+(d^2U_1/dt^2+2idU_1/dt-U_1)A$$

$$+\varepsilon(d^2U_22/dt^2+4idU_22/dt-4U_22)A^2+cc \tag{7}$$

$$u^2=2U_1\ \bar{U}_1+U_1^2A^2+2\varepsilon(\ \bar{U}_1U_2+U_0U_1)A+cc \tag{8}$$

Since u^3 appears with a coefficient of ε^2 in (1), one only needs to keep terms of O(1) in the expansion:

$$u^3=3U_1^2\ \bar{U}_1A+cc \tag{9}$$

Similarly, one keeps O(ε) terms in $(du/dt)^2$, but only O(1) terms in $u(du/dt)^2$:

$$(du/dt)^2=2U_1\ \bar{U}_1-(U_1^2A^2+cc) \tag{10}$$

$$u(du/dt)^2=U_1^2\ \bar{U}_1A+cc \tag{11}$$

We now substitute (3) and (6)-(11) in (1'), collect terms of like power of A_k, k=0,1 and 2, and

then set the coefficients to zero. The resulting equations, for the coefficients of A_0, A_1 and A_2 respectively, are:

$$\varepsilon[U_0+(1/2)(4\alpha_2+\alpha_4)U_1\ \tilde{U}_1+2(\alpha_2+\alpha_4)U_2\ \bar{U}_2]=0 \tag{12}$$

$$3U_1/4-fS+idU_1/dt+i\varepsilon\mu U_1+\varepsilon(2\alpha_2+\alpha_4)\ \bar{U}_1U_2=0 \tag{13}$$

$$2i(dU_2/dt+\varepsilon\mu U_2)+\varepsilon(4\alpha_2-\alpha_4)U_1^2/4+d^2U_1/dt^2$$

$$+\varepsilon(2\mu dU_2/dt+i\alpha_4U_1dU_1/dt)$$

$$+\varepsilon^2[2\alpha_2U_0U_2+(2\alpha_2+3\alpha_4/2)U_1U_3+2(\alpha_2+2\alpha_4)U_2U_4$$

$$+(3\alpha_3+\alpha_5)U_2^2\ \bar{U}_2+(6\alpha_3+\alpha_5/2)U_1\ \bar{U}_1U^2)]=0 \tag{14}$$

$$-5\varepsilon U_3/4+\varepsilon(2\alpha_2-\alpha_4)U_1U_2=0 \tag{15}$$

$$-3\varepsilon U_4+\varepsilon(\alpha_2-\alpha_4)U_2^2=0 \tag{16}$$

From (12), (15) and (16), U_0, U_3 and U_4 can be solved directly in terms of U_1 and U_2:

$$U_0=-2(\alpha_2+\alpha_4)U_1\ \bar{U}_1-(1/2)(4\alpha_2+\alpha_4)U_2\ \bar{U}_2 \tag{17}$$

$$U_3=(4/5)(2\alpha_2-\alpha_4)U_1U_2 \tag{18}$$

$$U_4=(\alpha_2-\alpha_4)U_2^2/3 \tag{19}$$

In equation (13) and (14), however, it is observed that some terms are of one order of ε greater than the others. The terms of higher order in ε can thus be less accurate than others and still yield the same order of approximation in these equations. One then can solve these equation first using only the dominant terms. Then, substitute the results back into the terms of higher order in ε, solve the full equations and obtain improved results. The immediate purpose here is to reduce (16) into a first order differential equation in U_2 and express all the other U_ks in terms of U_2.

Using the dominant terms in (13) and (14), one has

$$U_1 = 4fS/3 \tag{20}$$

$$2i(dU_2/dt+\varepsilon\mu U_2)+\varepsilon(4\alpha_2-\alpha_4)U_1^2/4 = 0 \qquad (21)$$

Equation (20) is used in the terms of order ε in (13) to yield the improved U_1:

$$U_1=4fS/3+(1/9)\varepsilon[8(\sigma-2i\mu)fS-16(2\alpha 2+\alpha 4)fSU_2]=0 \qquad (22)$$

Now, the terms in (14), which are of higher order in ε, contain such quantities as d^2U_2/dt^2, dU_2/dt, dU_1/dt, U_1, U_0, U_3, U_4. These expressions can be obtained by using (20), (21), their differentiations (for d^2U_2/dt^2 and dU_1/dt), by using (17), (18) and (19). The final form of (14) can be written as the following:

$$2idU_2/dt+\varepsilon(2i\mu U_2+c_1f^2S^2U_2)$$

$$+\varepsilon^2[c_2U_2^2\ \bar{U}_2+c_{34}f^2S^2+(c_5f^2-\mu^2)U_2)]=0 \qquad (23)$$

where

$$c_1=4(4\alpha_2-\alpha_4)/9$$

$$c_2=(9\alpha_3+3\alpha_5-10\alpha_2^2-10\alpha_2\alpha_4-4\alpha_4^2)/3$$

$$c_{34}=c_3+ic_4$$

with (24)

$$c_3=2\sigma(20\alpha_2-17\alpha_4)/27$$

$$c_4=-2\mu(52\alpha_2-13\alpha_4)/27$$

and

$$c_5=(1440\alpha_3+120\alpha_5-1472\alpha_2^2-368\alpha_2\alpha_4-128\alpha_4^2)/135$$

The key equations (3), (23), (17), (18), (19) and (22) can be further simplified by the following change of variables. Let

$$U_k=V_kS_k,\ \ V_k=U_kS^{-k},\ \ k=0,1,..4 \qquad (25)$$

where S was defined in (4). One also has

$$dU_k/dt=dV_k/dt+ik\varepsilon\sigma V_k/2 \qquad (26)$$

In terms of V_k, equations (4), (23), (17), (18), (19) and (22) become respectively

$$u=\varepsilon V_0+[V_1B+V_2B^2+\varepsilon(V_3B^3+V_4B^4)+cc] \qquad (27)$$

with

$$V_0=-(32/9)(\alpha_2+\alpha_4)f^2-2(\alpha_2+\alpha_4)V_2\ \bar{V}_2 \qquad (28)$$

$$V_1=4f/3+(1/9)\varepsilon[8(\sigma-2i\mu)f-16(2\alpha_2+\alpha_4)fV_2 \qquad (29)$$

$$2i(dU_2/dt+\varepsilon\mu U_2)+\varepsilon(4\alpha_2-\alpha_4)U_1^2/4 = 0 \qquad (21)$$

Equation (20) is used in the terms of order ε in (13) to yield the improved U_1:

$$U1=4fS/3+(1/9)\varepsilon[8(\sigma-2i\mu)fS-16(2\alpha2+\alpha4)fSU2]=0 \qquad (22)$$

Now, the terms in (14), which are of higher order in ε, contain such quantities as d^2U_2/dt^2, dU_2/dt, dU_1/dt, U_1, U_0, U_3, U_4. These expressions can be obtained by using (20), (21), their differentiations (for d^2U_2/dt^2 and dU_1/dt), by using (17), (18) and (19). The final form of (14) can be written as the following:

$$2idU_2/dt+\varepsilon(2i\mu U_2+c_1f^2S^2U_2)$$

$$+\varepsilon^2[c_2U_2^2\ \bar{U}_2+c_{34}f^2S^2+(c_5f^2-\mu^2)U_2)]=0 \qquad (23)$$

where

$$c_1=4(4\alpha_2-\alpha_4)/9$$

$$c_2=(9\alpha_3+3\alpha_5-10\alpha_2^2-10\alpha_2\alpha_4-4\alpha_4^2)/3$$

$$c_{34}=c_3+ic_4$$

with (24)

$$c_3=2\sigma(20\alpha_2-17\alpha_4)/27$$

$$c_4=-2\mu(52\alpha_2-13\alpha_4)/27$$

and

$$c_5=(1440\alpha_3+120\alpha_5-1472\alpha_2^2-368\alpha_2\alpha_4-128\alpha_4^2)/135$$

The key equations (3), (23), (17), (18), (19) and (22) can be further simplified by the following change of variables. Let

$$U_k=V_kS_k,\quad V_k=U_kS^{-k},\quad k=0,1,..4 \qquad (25)$$

where S was defined in (4). One also has

$$dU_k/dt=dV_k/dt+ik\varepsilon\sigma V_k/2 \qquad (26)$$

In terms of V_k, equations (4), (23), (17), (18), (19) and (22) become respectively

$$u=\varepsilon V_0+[V_1B+V_2B^2+\varepsilon(V_3B^3+V_4B^4)+cc] \qquad (27)$$

with

$$V_0=-(32/9)(\alpha_2+\alpha_4)f^2-2(\alpha_2+\alpha_4)V_2\ \bar{V}_2 \qquad (28)$$

$$V_1=4f/3+(1/9)\varepsilon[8(\sigma-2i\mu)f-16(2\alpha_2+\alpha_4)fV_2 \qquad (29)$$

$$V_3=(4/5)(2\alpha_2-\alpha_4)V_1V_2 \tag{30}$$

$$V_4=(\alpha_2-\alpha_4)V_2^2/3 \tag{31}$$

and,

$$2idV_2/dt+\varepsilon(-2\sigma+2i\mu+c_1f^2)V_2$$

$$+\varepsilon^2[c_2V_2^2V_2+c_{34}f^2+(c_5f^2-\mu^2)V_2)]=0 \tag{32}$$

where, in (29),

$$B=SA=\exp[(1+\varepsilon\sigma/2)t]=e^{i\Omega t} \tag{33}$$

Hence the original differential equation (1) has been reduced to (32), where V_2 is the unknown function. Once V_2 is solved, other V_ks can be obtained from (28) through (31). Then u(t) is given by (27).

To illustrate what kind of information one can extract from the equations derived so far, we shall obtain the magnitude for a superharmonic in the steady state solution and determine the stability of such a solution. First, we shall write the needed equations in terms of real variables. To this end, let

$$V_2=V_{2R}+iV_{2I}=\rho^2\exp(i\gamma_2)$$

$$V_2=(x-iy)/2 \tag{34}$$

where now ρ_2, γ_2, $V_{2R}=x/2$ and $V_{2I}=-y/2$ are all real functions of t. One also has

$$dV_2/dt=(dx/dt-idy/dt)/2 \tag{35}$$

Note that we have introduced two new variables x and y such that

$$x=2V_{2R},\quad y=-2V_{2I} \tag{36}$$

to save some writing. Substitute (34) and (35) in (32) and separate the real and imaginary part, one has two equations for two real variables x and y:

$$dx/dt+\varepsilon[\mu x+\sigma y]+\varepsilon 2[c_4f^2-c_2(x^2+y^2)y/8+(c_5f^2-\mu^2)y/2] = 0 \tag{37a}$$

$$dy/dt+\varepsilon[\mu y-\sigma x]+\varepsilon 2[c_3f^2+c_2(x^2+y^2)y/8+(c_5f^2-\mu^2)x/2] = 0 \tag{37b}$$

For steady state solutions, we require that the amplitudes and phase angles of various harmonic components to be constant with respect to time t,

$$d\rho_k/dt=0,\ d\gamma_k/dt=0,\ k=0,1..,4 \tag{38}$$

In particular,

$$d\rho_2/dt=0,\quad d\gamma_2/dt=0 \tag{39a}$$

and, what is equivalent:

$$dx/dt=0,\quad dy/dt=0 \tag{39b}$$

It should be noted that (39a) actually also quarantee the validity of (38) for k other than 2. This fact can be easily observed from the relations of (28)-(31), which relate V_k, k=0,1,3 and 4, to V_2.

Now, substitute (39b) in (37), one has

$$\mu x+\sigma y+\varepsilon[c_4f^2-c_2(x^2+y^2)y/8+(c_5f^2-\mu^2)y/2]=0 \tag{40a}$$

$$\mu y-\sigma x+c_1f^2+\varepsilon[c_3f^2+c_2(x^2+y^2)x/8+(c_5f^2-\mu^2)x/2]=0 \tag{40b}$$

Some numerical results will be presented in determining the presence of superharmonic oscillations for the following given set of parameters:

$$\begin{gathered}\alpha_2=0.3,\ \alpha_3=0.1,\ \alpha_4=0.,\ \alpha_5=0.,\\ \varepsilon=0.1,\ \mu=2.0,\ \sigma=3.0,\ f=2.0\end{gathered} \tag{41}$$

This is a very simple case due to the fact that c_2 vanishes as can be seen from (24). Thus (40) become linear and the solution can be easily obtained as

$$x=0.1824,\ y=-0.0418 \tag{42}$$

Hence, from (34), the magnitude of the super-harmonic oscillation of order 2, ρ_2 is

$$\rho_2=0.5(x^2+y^2)=0.0936 \tag{43}$$

Next, equations (37) are integrated numerically. The result is the so called van del Pol plane [7] as show in Figure 1. As indicated in this plot, solutions converge to the steady state solution obtained above as the time increases. Hence the steady state solution is stable and the point "A" of (42) is known as a stable node. Results for more general cases will be reported in the future.

REFERENCES

[1] A. H. Nayfeh, The response of single degree of freedom systems with quadratic and cubic non-linearities to a subharmonic excitation, Journal of Sound and Vibration (1983), Vol. 89(4), pp.457-470.

[2] A. H. Nayfeh, Perturbation Methods in Nonlinear Dynamics, Lecture Notes in Physics: Nonlinear Dynamics Aspects of Particle Accelerators - Proceedings of the Joint US-CERN School on Particle Accelerators, Editors: J. M. Jowett, M. Month and S. Turner, Spring-Verlag, 1985, pp.238-314.

[3] B. Noble and M. A. Hussain, Multiple Scaling and a Related Expansion Method, with Applications, Lasers, Molecules and Methods (J. O. Hirschfelder, R. E. Wyatt and R. D. Coalson, Eds.), John Wiley & Sons,1989, pp.83-136.

[4] M. A. Hussain, B. Noble and J. J. Wu, Using Macsyma in a Generalized Harmonic Balance Method for a Problem od Forced Nonlinear Oscillation, Proc. Sixth Army Conference on Applied Mathematics and Computing (held 31 May - 3 June 1988, Univ. of Colarado, Boulder, Colarado), 1989, pp.713-732.

[5] B. Noble, M. A. Hussain and J. J. Wu, A Generalized Harmonic Balance Method for a Forced Nonlinear Oscillation - Numerical Solution Formulation and Results, Proc. Seventh Army Conference on Applied Mathematics and Computing (held 6-9 June 1989, U.S. Military Academy, West Point, New York), 1990, pp.837-861.

[6] J. J. Wu, On the Analysis of Subharmonic Oscillations, Submitted for publication.

[7] D. W. Jordan and P. Smith, Nonlinear Differential Equations, Second Edition, Oxford University Press, 1986, p.183.

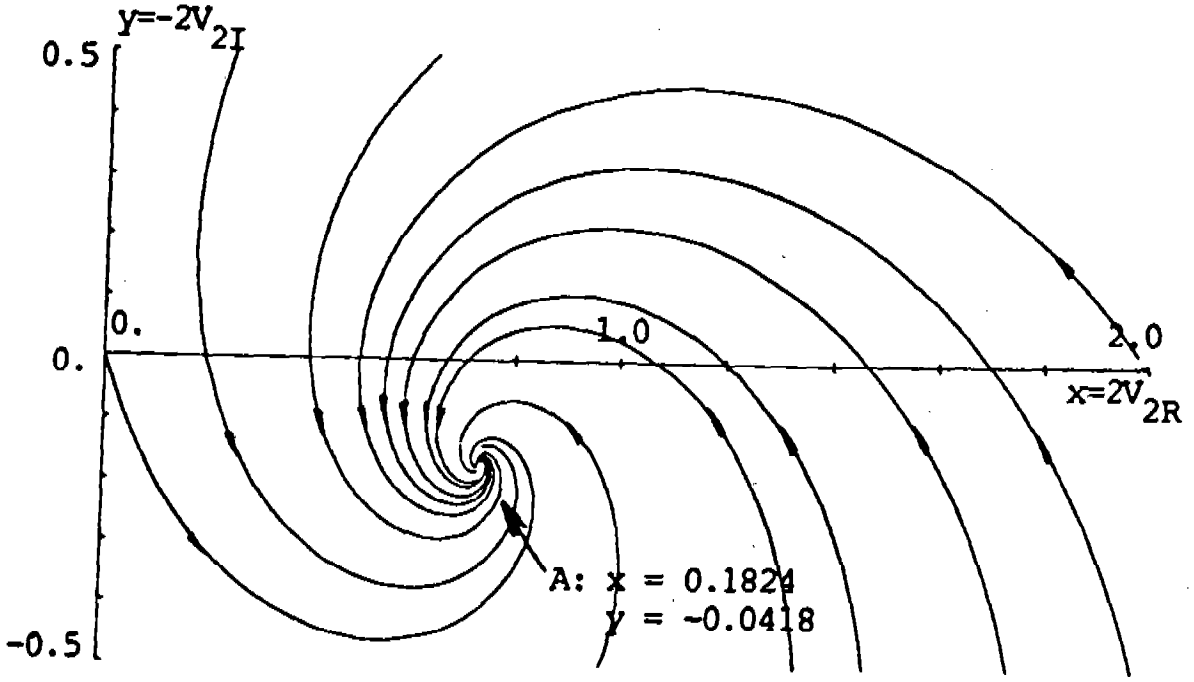


FIGURE 1. The van del Pol plane for the superharmonics of order 2 for the set of parameters given in equations (41). Point "A" shown is a stable node.

Constitutive Coefficients for Viscohyperelastic Materials

A. R. Johnson and C. J. Quigley
Army Materials Technology Laboratory
Watertown, MA 02172-0001

D. L. Cox*, L. C. Bissonnette**, and W. C. Maciejewski*
Naval Underwater Systems Center

Introduction

Elastic and viscous stresses in rubberlike materials can be modeled using strain energy density functions. The large strain elastic (hyperelastic) deformations are often modeled with the Rivlin strain invariant power series[1]. Similarly, large strain viscous deformations of rubberlike materials (viscohyperelastic) can be modeled using an internal solid theory with hyperelastic solids[2,3,4,5]. The energy function's material coefficients are found by least square fitting to the classical tension, shear, and equibiaxial stress-stretch tests[6]. These least squares fits typically produce energy functions which are not stable for deformations other than those covered by the test data. That is, when strain states not included in the test data are considered the models often suffer from the flaw that (for isothermal deformations) they predict a decrease in the solid's internal strain energy for an increment of applied stress which does positive work on the solid. This conservation of energy statement is known as Drucker's postulate on stability. Such a flaw cannot be accepted since computations for complex deformations will include strain states which are not the same as those used to determine the energy density function. Energy conservation will then be violated somewhere in the solid (or the computational algorithm will fail, etc.)

In this effort we derive formulas for the constraints on the coefficients of a hyperelastic Rivlin (third order invariant expansion) energy density function which enforce Drucker stability. Then, an example is presented in which uniaxial and equibiaxial stress-stretch data is least squares fit to both an unconstrained and a constrained third order invariant

* New London, CT

** Newport, RI

Rivlin energy density function. The stability of these functions is then addressed. It is shown that the simple constraint of requiring the Rivlin series coefficients to be positive is a practical way to determine the energy density function. We note, however, that the constraint of positive coefficients is not sufficient for stability (that is, the constraint equations must still be checked).

Least Squares Fit to Rivlin Energy Function

The stress-strain response of rubber, without consideration of viscoelastic effects, is modeled with strain energy density functions. There are numerous algebraic forms available for the energy function. Typically, these functions are represented by expansions in powers of the strain invariants or stretch ratios. In this effort we use the following Rivlin energy function.

$$W = \sum_{\ell+m \geq 1}^{3} \sum^{3} C_{\ell m}(I_1 - 3)^{\ell}(I_2 - 3)^{m} \qquad (1)$$

where $I_1 = \lambda_1^2 + \lambda_2^2 + \lambda_3^2$

and $I_2 = 1/\lambda_1^2 + 1/\lambda_2^2 + 1/\lambda_3^2$

The quantities I_1 and I_2 are invariants of the deformation and λ_1, λ_2, λ_3 are the principal stretch ratios. The coefficients $C_{\ell m}$ are typically computed by the following test and least squares fitting procedure[1]. The engineering stresses for uniaxial tension and shear, and for equibiaxial tension computed using equation 1 (σ^T, σ^S and σ^B respectively) are

$$\sigma^{*} = \sum_{\ell+m \geq 1}^{3} \sum^{3} C_{\ell m} A^{*}_{\ell m} \qquad (2)$$

where * = T, S, or B,

$$A^{T}_{\ell m} = 2[\lambda - \frac{1}{\lambda^2}][\ell(I_1 - 3)^{\ell-1}(I_2-3)^{m} + \frac{m}{\lambda}(I_1-3)^{\ell}(I_2-3)^{m-1}] \qquad (3)$$

$$A^{S}_{\ell m} = 2[\lambda - \frac{1}{\lambda^3}][\ell(I_1 - 3)^{\ell-1}(I_2-3)^{m} + \frac{m}{\lambda}(I_1-3)^{\ell}(I_2-3)^{m-1}] \qquad (4)$$

$$A^{B}_{\ell m} = 2[\lambda - \frac{1}{\lambda^5}][\ell(I_1 - 3)^{\ell-1}(I_2-3)^m + \frac{m}{\lambda}(I_1-3)^{\ell}(I_2-3)^{m-1}] \quad (5)$$

and λ = the stretch ratio measured in the direction of loading (i.e., the extensional stretch, see reference 1). The invariants for tension are given by

$$I_1 = \lambda^2 + 2/\lambda^2 \text{ and } I_2 = 2\lambda + 1/\lambda^2 \quad (6)$$

for shear by

$$I_1 = I_2 = 1 + \lambda^2 + 1/\lambda^2 \quad (7)$$

and for equibiaxial tension by

$$I_1 = 2\lambda^2 + 1/\lambda^4 \text{ and } I_2 = 2/\lambda_2 + \lambda^4 \quad (8)$$

An error function Π is constructed from the experimental data as follows (* = T, S, B),

$$\Pi = \sum_{*} \sum_{e} (\sigma^{*}_{e} - \sigma^{*}(\lambda_e))^2 \quad (9)$$

where "e" implies the measured data, σ^{*}_{e} implies the measured engineering stresses and, $\sigma^{*}(\lambda_e)$ implies the engineering stresses computed using the measured stretch, λ_e, in equations 3, 4, and 5. The constants $C_{\ell m}$ are then selected to minimize the least squares error given by equation 9. They are computed as follows. Let

$$\{A^{*}_{e}\}^T = \{A^{*}_{10}(\lambda_e), A^{*}_{01}(\lambda_e), \ldots, A^{*}_{01}(\lambda_e)\} \quad (10)$$

and

$$\{C\}^T = \{C_{10}, C_{01}, \ldots, C_{03}\} \quad (11)$$

Then, equation 9 becomes

$$\Pi = \text{constant} - 2 \sum_{*} \sum_{e} \sigma_e^* \{A^*(\lambda_e)\}^T \{C\} + \{C\}^T \left(\sum_{*} \sum_{e} \{A^*(\lambda_e)\}\{A^*(\lambda_e)\}^T \right) \{C\} \tag{12}$$

Let

$$\{b\} = \sum_{*} \sum_{e} \sigma_e^* \{A^*(\lambda_e)\}^T \tag{13}$$

and

$$[A] = \sum_{*} \sum_{e} \{A^*(\lambda_e)\{A^*(\lambda_e)\}^T \tag{14}$$

Then, the least squares error function Π becomes

$$\Pi = \text{constant} - 2\{b\}^T\{C\} + \{C\}^T[A]\{C\} \tag{15}$$

and the minimum error occurs when the first variation of Π is zero. That is, when

$$\{C\} = [A]^{-1}\{b\} \tag{16}$$

Stability Requirement

It is common practice to perform at least two of the stress-stretch tests mentioned above and then to find the constants {C} using equation 16. Recently, the general purpose nonlinear finite element code ABAQUS[10] has added a routine to check the user's energy function for Drucker stability under several specified deformations (tension, compression, shear, and equibiaxial tension). We outline this stability check here for isotropic materials. Let $d\tau_i$ = an increment in the i'th principal Cauchy stress and $d\varepsilon_i$ = an increment in the corresponding strain at any point in the solid. Then Drucker's stability postulate states

$$\sum_{i} d\tau_i \, d\varepsilon_i > 0 \tag{17}$$

The Cauchy stresses are given by

$$\tau_i = \lambda_i \frac{\partial W}{\partial \lambda_i} + p \tag{18}$$

where p = the hydrostatic pressure. For the case of plane stress we have $\tau_3 = 0$ and find

$$\tau_1 = \lambda_1 \frac{\partial W}{\partial \lambda_1} - \lambda_3 \frac{\partial W}{\partial \lambda_3}$$

$$\tau_2 = \lambda_2 \frac{\partial W}{\partial \lambda_2} - \lambda_3 \frac{\partial W}{\partial \lambda_3} \tag{19}$$

Using the chain rule it can be shown[10] that

$$\begin{pmatrix} d\tau_1 \\ d\tau_2 \end{pmatrix} = \begin{pmatrix} D_{11} & D_{12} \\ D_{12} & D_{22} \end{pmatrix} \begin{pmatrix} d\varepsilon_1 \\ d\varepsilon_2 \end{pmatrix} \tag{20}$$

where

$$D_{11} = 4(\lambda_1^2 + \lambda_3^2)(W_1 + \lambda_2^2 W_2) + 4(\lambda_1^2 - \lambda_3)^2[W_{11} + 2\lambda_2^2 W_{12} + \lambda_2^4 W_{22}]$$

$$D_{22} = 4(\lambda_2^2 + \lambda_3^2)(W_1 + \lambda_2^2 W_2) + 4(\lambda_2^2 - \lambda_3)^2[W_{11} + 2\lambda_1^2 W_{12} + \lambda_1^4 W_{22}] \tag{21}$$

$$D_{12} = 4\lambda_3^2 W_1 + 4\lambda_3^{-2} W_2 + 4(\lambda_1^2 - \lambda_3^2)(\lambda_2^2 - \lambda_3^2)[W_{11} + (\lambda_1^2 + \lambda_2^2)W_{12} + \lambda_1^2\lambda_2^2 W_{22}]$$

$$W_i = \frac{\partial W}{\partial I_i}$$

and $$W_{ij} = \frac{\partial^2 W}{\partial I_i \partial I_j}$$

The material is then stable (equation 17 satisfied) when the matrix [D] in equation 20 is positive definite. This is true when

$$D_{11} + D_{22} > 0 \tag{22}$$

and

$$D_{11}D_{22} - D_{12}^2 > 0 \tag{23}$$

Given an energy function and a strain state one can compute the matrix [D] in equation 20 and use equations 22 and 23 to check for stability.

Constraints for Stable Energy Function

Checking for and verifying stability with equations 22 and 23 above for specific strain states (tension, compression, shear, equibiaxial, etc.) does not assure stability elsewhere in strain, and checking for stability at each element's integration point in a nonlinear finite element analysis is computationally expensive. In this section we derive the constraints on the constants {C} in equation 15 so that the constrained least squares fit will satisfy equation 22 and 23 for all possible strain states. We note that applying these constraints will increase the least squares error (the data fit will not be as good) but the resulting energy functional will be stable everywhere. Let

$$x = \lambda_1^2,\ y = \lambda_2^2,\ z = \lambda_3^2,$$

$$A_{ij} = \frac{D_{ij}}{4}$$

$$\{W\} = \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} \tag{24}$$

$$[W] = \begin{pmatrix} W_{11} & W_{12} \\ W_{12} & W_{22} \end{pmatrix}$$

and $$\{X\} = \begin{pmatrix} 1 \\ x \end{pmatrix},\ \{Y\} = \begin{pmatrix} 1 \\ y \end{pmatrix},\ \{Z\} = \begin{pmatrix} z \\ xy \end{pmatrix}$$

We then find that

$$A_{11} = (x + z)\{W\}^T\{Y\} + (x - z)^2\{Y\}^T[W]\{Y\}$$

$$A_{22} = (y + z)\{W\}^T\{X\} + (y - z)^2\{X\}^T[W]\{X\} \tag{25}$$

and $$A_{12} = \{W\}^T\{Z\} + (x - z)(y - z)\{X\}^T[W]\{Y\}$$

so that a sufficient condition for the stability requirement of equation 22 is

$$C_{\ell m} \geq 0 \tag{26}$$

We now consider equation 23 and determine additional constraints (beyond

equation 26) to assure stability (sufficient constraints). Using equation 25 we find

$$\begin{aligned} A_{11}A_{22} - A_{12}^2 = {} & (W_{11}W_{22} - W_{12}^2)[(x - z)^2(y - z)^2(x - z)^2] \\ & + (W_1W_{22})[z^3(x - y)^2 + y^3(z - x)^2 + x^3(z - y)^2] \\ & + (W_2W_{11})[xz(z - x)^2 + yz(z - y)^2 + xy(y - x)^2] \\ & + 4(W_1W_{12})[x^2(y - z)^2 + y^2(x - z)^2 + z^2(x - y)^2] \\ & + ((W_2W_{22}) + (W_1W_{11}))[y(x - z)^2 + x(y - z)^2 + z(x - y)^2] \\ & + 2(W_2W_{12})[(x - y)^2 + (x - z)^2 + (y - z)^2] \\ & + (W_1W_2)[xz(x + z) + xy(x + y) + yz(z + y)] \\ & + W_1^2[xz + xy + yz] \\ & + W_2^2[x + y + z] \end{aligned} \tag{27}$$

Since x, y, z > 0, $W_i \geq 0$ and $W_{ij} \geq 0$, we have the stability requirement of equation 23 as

$$G = W_{11}W_{22} - W_{12}^2 \geq 0 \tag{28}$$

We note that equation 28 is not a necessary condition for stability. It is only a sufficient condition (it is not even a sufficient condition unless equation 26 is true).

Let $\xi = I_1 - 3$ and $\eta = I_2 - 3$ then

$$W = \sum_{\ell+m \geq 1}^{N} \sum^{N} C_{\ell m}\, \xi^{\ell} \eta^{m} \tag{29}$$

where N = the highest order to which the invariants are raised in the energy function. The constraints are computed as follows.

N=1, First Order Invariant

We have

$$W = C_{10}\xi + C_{01}\eta \tag{30}$$

and G = 0 always.

N=2, Second Order Invariant

We have

$$W = C_{10}\xi + C_{01}\eta + C_{11}\xi\eta + C_{20}\xi^2 + C_{02}\eta^2 \tag{31}$$

and $G \geq 0$ (by direct computation) when

$$A = 4C_{20}C_{02} - C_{11}^2 \geq 0 \tag{32}$$

N=3, Third Order Invariant

We have

$$W = \sum_{\substack{\ell+m \geq 1}}^{3}\sum^{3} C_{\ell m}\,\xi^{\ell}\eta^{m} \tag{33}$$

and $G \geq 0$ becomes

$$G = A + B\xi + C\eta + D\xi^2 + E\xi\eta + F\eta^2 \geq 0 \tag{34}$$

where

$$\begin{aligned} B &= 12\,C_{30}C_{02} + 4C_{20}C_{12} - 4C_{11}C_{21} \\ C &= 12\,C_{03}C_{20} + 4C_{02}C_{21} - 4C_{11}C_{12} \\ D &= 12\,C_{30}C_{12} - 4C_{21}^2 \\ E &= 36\,C_{30}C_{03} - 4C_{12}C_{21} \end{aligned} \tag{35}$$

and $F = 12\,C_{03}C_{21} - 4C_{12}^2$.

By direct calculation it can be shown that $G \geq 0$ for all $\xi, \eta \geq 0$ when

$$A, B, C, D, E, F \geq 0 \tag{36}$$

and $4FD - E^2 \geq 0$

Equations 36 represent the constraints which when combined with the constraint $C_{\ell m} \geq 0$ assures stability of an incompressible hyperelastic material in plane stress modeled with a third order invariant Rivlin energy functional.

Unconstrained and constrained models for a filled butadiene - styrene copolymer

Experimental data in uniaxial and equibiaxial tension[11] for a filled butadiene-styrene copolymer was fit to the third order Rivlin energy

function of equation 1. The material tested had been conditioned by repeated stretching beyond the levels reported below and was allowed to recover for at least fifteen minutes prior to testing. The uniaxial tension data was obtained by pulling at a slow strain rate (0.02 in/in/min) and the equibiaxial data was obtained in a flat disk inflation experiment in which the material was allowed to creep for about five minutes at constant pressure prior to recording the inflated shape.

The error function of equation 15 was minimized both with and without the constraint $C_{\ell m} \geq 0$. The stability constraints of equation 36 were checked after the constrained minimum was found. Figures 1 and 2 show plots of the unconstrained and constrained least squares fits. Also, the classical uniaxial shear test response was computed and shown for each case. The constrained least square model satisfied the stability equations 36. The unconstrained model was obviously unstable and also gave what appears to be a poor approximation in shear.

Summary

A set of constraints were derived for the coefficients of a third order invariant Rivlin energy function which assure Drucker stability in plane stress. Data for a filled butadiene-styrene copolymer was fit to the Rivlin function with and without the constraint $C_{\ell m} \geq 0$. The constrained model was stable and the unconstrained model was not.

References

1. James, A. G., Green, A. and Simpson, G.M., Strain energy functions of rubber I. Characterization of gum vulcanizates, Journal of Applied Polymer Science, 19, 1975, 2033-2058.

2. Johnson, A. R., Quigley, C. J., Cavallaro, C. and Weight, K. D., A large deformation viscoelastic finite element model for elastomers, in The Mathematics of Finite Elements and Applications VII, Ed. J. R. Whiteman, Academic Press, New York, 1991, 287-299.

3. Johnson, A. R., Quigley, C. J., Weight, K. D., Cavallaro, C. and Cox, D. L., The inflation and deflation of a thick walled visco-hyperelastic

sphere, The Transactions of the Eighth Army Conference on Applied Mathematics and Computing, U. S. Army Research Office Report No. 91-1, 1991, 847-857.

4. Johnson, A. R. and Quigley, C. J., A viscohyperelastic Maxwell model for rubber viscoelasticity, presented at the 139th Meeting of the Rubber Division, American Chemical Society, Toronto, Ontario, Canada, 21-24 May 1991 (accepted for publication in Rubber Chemistry and Technology).

5. Johnson, A. R., Quigley, C. J., Young, D. G. and Danik, J. A., Viscohyperelastic modeling of rubber vulcanizates, submitted to Tire Science and Technology.

6. Treloar, L. R. G., Stress-strain data for vulcanized rubber under various types of deformation, Transactions of the Faraday Society, 40, 1944, 59-70.

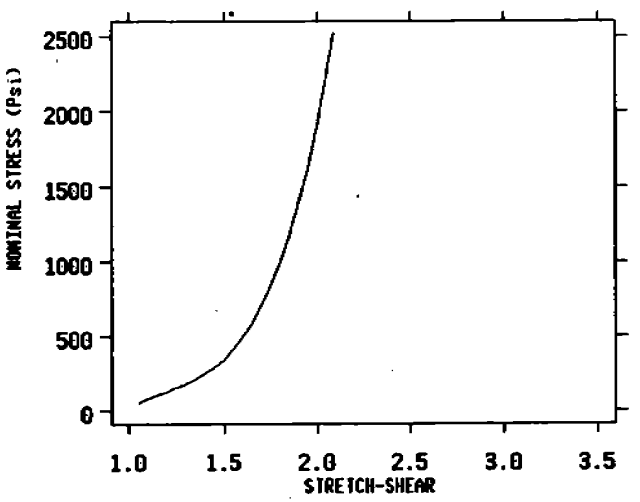
7. Treloar, L. R. G., The mechanics of rubber elasticity, The Proceedings of the Royal Society of London A, 351, 1976, 301-330.

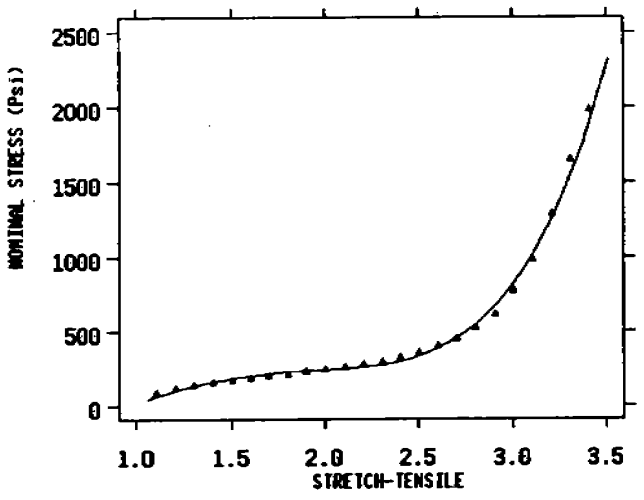
8. Glucklich, J. and Landel, R. F., Strain energy function of styrene butadiene rubber, Journal of Polymer Science, Polymer Physics Edition, 15, 1977, 2185-2199.

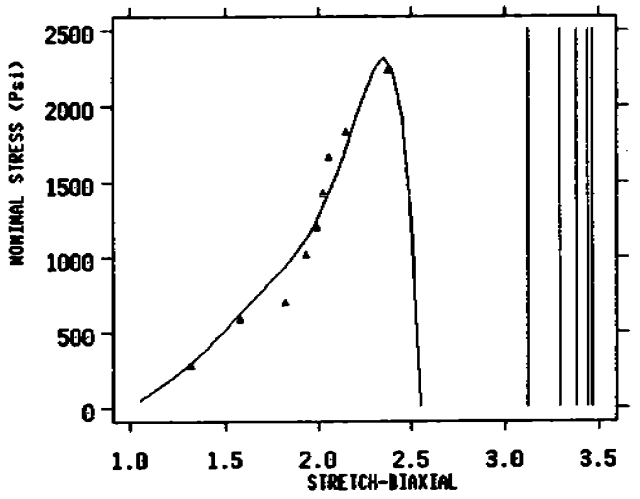
9. Ogden, R. W., Nearly isochoric elastic deformations: application to rubberlike solids, Journal of the Mechanics and Physics of Solids, 26, 1978, 2185-2199.

10. ABAQUS User's Manual Version 4.8, Hibbit, Karlsson, and Sorensen, Inc., Providence, R. I., 1989.

11. Bamberg, R. P., Aghababian, R. R., Cavallaro, C., and Johnson, A. R., Equibiaxial testing of TP-14AX carbon black rubber sheets, Army Materials Laboratory, Technical Report (under review).







$\xi = I_1 - 3$ $\eta = I_2 - 3$

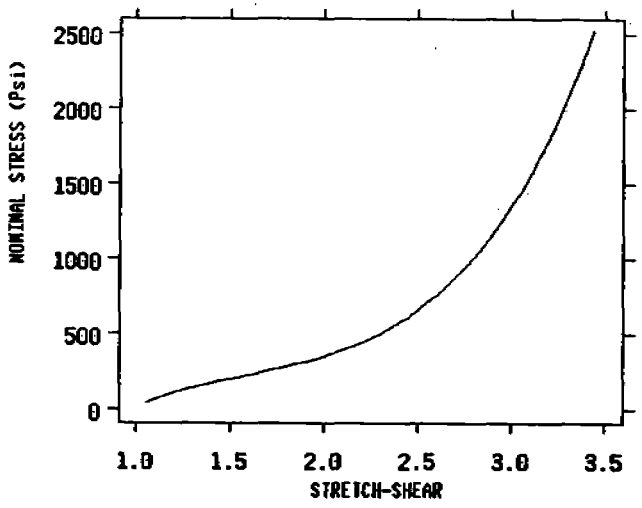
$$W = 65.6\,\xi + 20.1\,\eta - 213.0\,\xi^2 + 413.0\,\xi\eta - 189.0\,\eta^2 - 0.103\,\xi^3 + 15.7\,\xi^2\eta + 10.2\,\xi\eta^2 - 1.08\,\eta^3$$

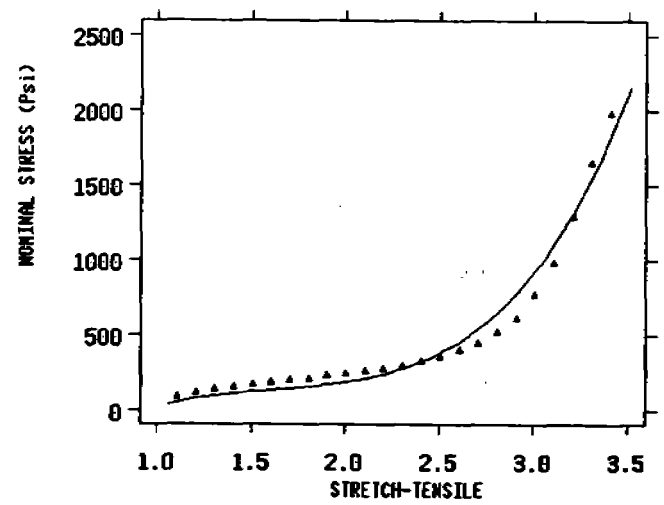
W is unstable.

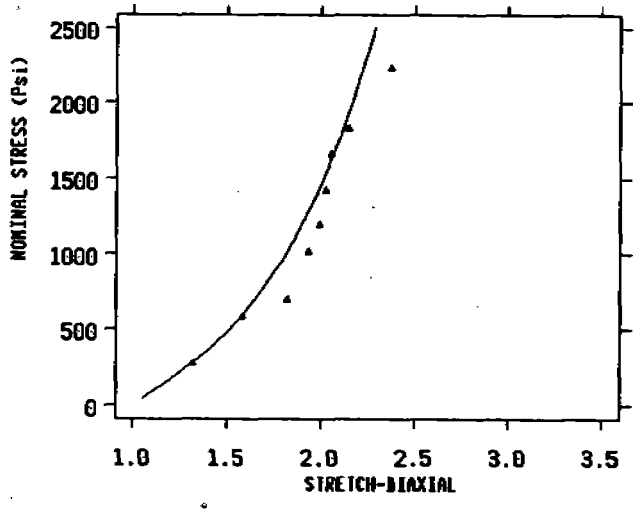
▲ = Test data.

--- = Least squares fit.

Figure 1. Nominal stress vs stretch for the unconstrained least squares fit.







$\xi = I_1 - 3$ $\eta = I_2 - 3$

$W = 73.8\ \eta + 1.00\ \xi^3$

W is stable.

▲ = Test data.

--- = Least squares fit.

Figure 2. Nominal stress vs stretch for the constrained ($C_{\ell m} \geq 0$) least squares fit.

HIGH-T_c SUPERCONDUCTIVITY AND THE PHOTOELECTRIC EFFECT

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. This paper interprets the phenomenon of high-T_c superconductivity in the oxide, heavy fermion and organic salt superconductors as a coherent spacetime state of electrons in a two-dimensional system of Cooper electron pairs. At a transition temperature the electrical resistance of a substance whose electrons are in a coherent spacetime state can go to zero in two ways, the first is the ordinary BCS case of superconductivity that is associated with the vanishing of the magnitude of the measured resistance, and the second is the case of coherent spacetime superconductivity that is associated with a value of $\pi/2$ for the internal phase angle of the resistance and a corresponding zero value for the measured resistance. The superconducting state ($T < T_c$) of a high-T_c superconductor is described by a completely coherent spacetime state, while the normal state ($T > T_c$) of a high-T_c superconductor is described by a partially coherent spacetime state. The normalized superconductivity energy gap for high-T_c substances is found to have the value $(6/\pi)(3.52)/(1 - 4/3\ \theta_a)$ where θ_a = relative internal phase angle of the electron-electron acceleration (force) within a Cooper pair of electrons. A comparison of this formula with the experimental values of the superconductivity energy gaps of high-T_c compounds gives the values of θ_a for particular lattice structures and their associated phonon interactions with the electrons. Small values of θ_a suggest that electron pairing is weak. The large values of the normalized superconductivity energy gaps for the high-T_c superconducting compounds is due to the factor $6/\pi$ which arises from the complete spacetime coherence of the superconducting state. Thermodynamic processes in high-T_c substances are examined. One of the experimental techniques for determining the characteristic parameters of high-T_c superconductors utilizes the photoelectric effect. Because the electron-electron interaction is weak in the Cooper pairs of high-T_c superconductors, the Fowler theory of the photoelectric effect for ordinary metals, that is based on a noninteracting electron gas, is extended to the cases of total and partial coherence of the spacetime states that describe the superconducting and normal states respectively of high-T_c superconductors.

1. INTRODUCTION. A spectacular discovery of physics in recent years was the observation of high-T_c superconductivity in a class of planar copper oxide compounds with $T_c \sim 30K$.[1-11] Already superconducting materials with T_c above liquid nitrogen have been created, and the possibility of room temperature and higher superconductors is now considered likely. The highest transition temperature to this date is $T_c \sim 125K$.[1-11] In addition to the high-T_c planar copper oxides (such as the famous 1-2-3 yttrium-barium-copper oxide) there are two other groups of superconductors with unusual properties. These are the organic salt superconductors (such as the Bechgaard salts) and the heavy fermion superconductors (such as UPt_3).[1-11] These discoveries have produced an intense research effort to correlate high-T_c superconductivity with the atomic structure of these materials in order to obtain a theoretical picture of the physical processes that cause high-T_c superconductivity.[12]

A. Basic Concepts.

High-T_c superconductivity may not be described by the Bardeen-Cooper-Schrieffer (BCS) theory that describes ordinary metallic superconductivity in terms of the phonon-mediated weakly coupled electron-electron attractive interaction and the formation of Cooper electron pairs in a relative s-state (the BCS singlet pairing with $\ell = 0$).[1-11] The BCS metallic superconductors and some of the high-T_c superconductors exhibit bulk superconductivity in the sense that the resistivity goes to zero in all three crystallographic directions at a common transition temperature. In general, however, both the superconducting state ($T < T_c$) and the normal state ($T > T_c$) of high-T_c superconductors are highly anisotropic.[1-11] For the normal state of the copper oxides, the resistivity in the Cu-O planes ρ_{ab} is essentially metallic while the resistivity in the out-of-plane direction, ρ_{ac} or ρ_{bc}, is like that of a semiconductor having the ordinary conductivity for oxides. This anisotropy of the Cu-O materials is also exhibited by the heavy fermion and organic salt superconductors.[1-11] Conflicting evidence exists that shows that the superconducting state of a high-T_c superconductor may not be described by the BCS theory of Cooper electron pairs.[1-11] For instance, there appears to be a close connection between non-BCS superconductivity and antiferromagnetism which is not yet explained.[1-11] Also, it should be pointed out that the normal state ($T > T_c$) of a high-T_c compound may not be an ordinary Fermi liquid because it exhibits peculiar properties, and therefore a description of the $T > T_c$ state would be of value.[13-15] However, some evidence suggests that an ordinary Fermi liquid description of the normal state is correct.[11,16-18]

Several experimental methods have been used to determine the mechanism of high-T_c superconductivity. These include Raman scattering, infrared reflectivity, nuclear magnetic resonance, Knight shift, nuclear spin-lattice relaxation lines, neutron scattering intensities, ultrasound, circular dichroism, specific heat, electrical resistivity, magnetic properties, positron emission, cold emission, and photoemission.[1-11] There are many other experimental techniques that are not listed here. Only the photoelectric effect is considered in this paper.

Superconductors are often described as being weakly or strongly coupled according to the strength of the electron-electron attractive interaction in the Cooper electron pairs.[1-11] It is presently thought that the strength of the electron-electron interaction is related to the value of the dimensionless relative superconductivity energy gap given by[1-11]

$$\Delta' = 2\Delta/(kT_c) \tag{1}$$

where Δ' = relative superconductivity energy gap, 2Δ = full superconductivity energy gap, k = Boltzmann constant, and T_c = superconductivity transition temperature. For a BCS superconductor[1-11]

$$\Delta'_{it} = 3.52 \tag{2}$$

where in the notation of the present paper the subscript it = incoherent time that is associated with the BCS theory. The BCS theory is based on weak coupling. When a superconductor has $\Delta' > 3.52$ it is generally referred to in the literature as being a strongly coupled superconductor. Measured values of Δ'

for high-T_c superconductors are as high as 10.[1-11,14,19-22] Many theories have been developed that attempt to describe the basic mechanism of strongly coupled superconductivity and the peculiar normal state properties of the non-BCS superconductors.[1-11] Later in this paper it will be shown that the large experimental values obtained for Δ' do not necessarily imply strong coupling, but are in fact due to a coherent spacetime state that exists in the electrons that constitute the Cooper pairs. It is suggested that, in fact, high-T_c superconductors are weakly coupled systems whose electrons exist in a coherent time state.

B. Spacetime with Broken Symmetry.

For space and time with broken internal symmetries the coordinates are written as [23-25]

$$\bar{\chi} = \chi e^{j\theta\chi} \qquad \bar{t} = te^{j\theta t} \tag{3}$$

where $\chi = x, y, z$ and where θ_χ = internal phase angles of the cartesian coordinates and θ_t = internal phase angle of the time coordinate. The complex number volume can be written as

$$\bar{V} = Ve^{j\theta V} = \bar{x}\bar{y}\bar{z} \tag{4}$$

so that

$$V = xyz \qquad \theta_V = \theta_x + \theta_y + \theta_z \tag{5}$$

and for isotropy $\theta_V = 3\theta_x$. The differential changes in the complex number space and time coordinates are given by

$$d\bar{t} = \sec \beta_{tt}\, dt\, e^{j(\theta_t+\beta_{tt})} = \csc \beta_{tt}\, td\theta_t\, e^{j(\theta_t+\beta_{tt})} \tag{6}$$

$$d\bar{\chi} = \sec \beta_{\chi\chi}\, d\chi\, e^{j(\theta_\chi+\beta_{\chi\chi})} = \csc \beta_{\chi\chi}\, \chi d\theta_\chi\, e^{j(\theta_\chi+\beta_{\chi\chi})} \tag{7}$$

where

$$\tan \beta_{tt} = t\partial\theta_t/\partial t \qquad \tan \beta_{\chi\chi} = \chi\partial\theta_\chi/\partial\chi \tag{8}$$

where $\chi = x, y, z$. The lengths of the time and space coordinates are given for partially coherent spacetime as[23]

$$t' = \int |d\bar{t}| = \int \sec \beta_{tt}\, dt = \int \csc \beta_{tt}\, td\theta_t \tag{9}$$

$$\chi' = \int |d\bar{\chi}| = \int \sec \beta_{\chi\chi}\, d\chi = \int \csc \beta_{\chi\chi}\, \chi d\theta_\chi \tag{10}$$

while the volume in partially coherent space is given by

$$V' = \int |d\bar{V}| = \int \sec \beta_{VV}\, dV = \int \csc \beta_{VV}\, Vd\theta_V \tag{11}$$

where

$$\tan \beta_{VV} = V\partial\theta_V/\partial V \tag{12}$$

For coherent spacetime $\beta_{tt} = \beta_{xx} = \beta_{yy} = \beta_{zz} = \pi/2$ and $\beta_{VV} = \pi/2$ and from equations (9) and (10) it follows that

$$t' = t\theta_t \qquad x' = x\theta_x \qquad y' = y\theta_y \qquad z' = z\theta_z \qquad V' = V\theta_V \tag{13}$$

where t , x , y , z and V = constants.

C. Coherent Spacetime Theory of Non-BCS Superconductivity.

Recently a new theory of high-T_c superconductivity was developed in which the characteristic properties of the superconducting state are attributed to the fact that for $T < T_c$ both time and space for the electrons in a Cooper pair become coherent, and physical processes occur while time and space rotate in an internal space as in equation (13).[23] For a spacetime with broken internal symmetries, the complex number potential difference across the battery terminals situated in the $\chi = x , y , z$ direction is written as[23]

$$\bar{W}_\chi = W_\chi e^{j\theta W\chi} \tag{14}$$

where W_χ = magnitude of the battery potential difference and $\theta_{W\chi}$ = internal phase angle of the potential difference given by[23]

$$\theta_{W\chi} = 2(\theta_\chi - \theta_t) \tag{15}$$

The measured potential difference across the battery in the χ direction (without current flowing) is given by

$$W_{\chi m} = W_\chi \cos \theta_{W\chi} \tag{16}$$

The complex number current in a conductor situated in the χ direction is written as[23]

$$\bar{I}_\chi = I_\chi e^{j\theta I\chi} \tag{17}$$

where I_χ = magnitude of current and $\theta_{I\chi}$ = internal phase angle of the current which is given by[23]

$$\theta_{I\chi} = -\theta_t \tag{18}$$

assuming that $\beta_{tt} = 0$ which means θ_t = constant. The measured current in the χ direction is then given by

$$I_{\chi m} = I_\chi \cos \theta_{I\chi} = I_\chi \cos \theta_t \tag{19}$$

The complex number resistance in the χ direction is given by[23]

$$\bar{R}_\chi = R_\chi e^{j\theta R\chi} = \bar{W}_\chi/\bar{I}_\chi \tag{20}$$

where the magnitude and internal phase angle of the resistance are given by[23]

$$R_\chi = W_\chi / I_\chi \qquad \theta_{R\chi} = \theta_{W\chi} - \theta_{I\chi} = \theta_{W\chi} + \theta_t \qquad (21)$$

$$= 2\theta_\chi - \theta_t$$

where R_χ = magnitude of resistance and $\theta_{R\chi}$ = internal phase angle of the resistance. The measured resistance of a conductor is given by[23]

$$R_{\chi m} = R_\chi \cos \theta_{R\chi} = R_\chi \cos(2\theta_\chi - \theta_t) \qquad (22)$$

where $R_{\chi m}$ = measured resistance in the χ direction. Combining equations (16), (19), (21) and (22) allows the measured resistance to be written as

$$R_{\chi m} = W^{eff}_{\chi m} / I_{\chi m} \qquad (23)$$

where $W^{eff}_{\chi m}$ = effective measured potential difference in the χ direction of a conductor with current flowing, which is given by

$$W^{eff}_{\chi m} = W_{\chi m}(\cos \theta_{R\chi} \cos \theta_{I\chi})/\cos \theta_{W\chi} \qquad (24)$$

$$= W_{\chi m}\{1 - \tan \theta_t \tan[2(\theta_\chi - \theta_t)]\}\cos^2 \theta_t$$

The measured resistance is then obtained from equations (23) and (24) as[23]

$$R_{\chi m} = R_{\chi c}\{1 - \tan \theta_t \tan[2(\theta_\chi - \theta_t)]\}\cos^2 \theta_t \qquad (25)$$

where the conventionally measured resistance is given by

$$R_{\chi c} = W_{\chi m} / I_{\chi m} \qquad (26)$$

where $R_{\chi c}$ = conventionally measured resistance in the χ direction.

According to equations (22) or (25) there are two ways that the measured resistance can be zero.[23] The first way corresponds to ordinary BCS superconductivity and occurs when $W_{\chi m} = 0$ or equivalently when[23]

$$R_{\chi c} = 0 \qquad \text{(BCS superconductivity)} \qquad (27)$$

and the second way occurs when[23]

$$\theta_{R\chi} = \theta_{W\chi} + \theta_t = \pi/2 \qquad 2\theta_\chi - \theta_t = \pi/2 \qquad (28)$$

which is the condition for high-T_c superconductivity according to the coherent spacetime theory of high-T_c superconductivity.[23] Combining the result in equation (28) with the condition of free electrons (weak pairing) in Cooper pairs $\theta_\chi = 2\theta_t$ gives the following condition for structurally induced superconductivity[23]

$$\theta_\chi = \pi/3 \qquad \theta_t = \pi/6 \tag{29}$$

The Heisenberg uncertainty principle can then be invoked to deduce the relationship between the normalized superconductivity energy gap for the coherent time theory and the corresponding normalized superconductivity energy gap for the BCS theory as follows[23]

$$\Delta'_{ct} = 6/\pi \; \Delta'_{it} = 6/\pi(3.52) = 1.91(3.52) = 6.72 \tag{30}$$

where ct = coherent time and it = incoherent time. The result in equation (30) is similar to values of the normalized superconductivity energy gaps that are measured by the various experimental methods mentioned earlier. However, the measured values of Δ'_{ct} are material dependent and can vary in the range of from 4 through 10, so that equation (30) must be replaced by a substance dependent method of calculating Δ'_{ct} and this is one of the calculations presented in this paper. Present day thought suggests that the large values of Δ'_{ct} suggest strongly coupled electrons, but equation (30) and the subsequent analysis in this paper shows that the large values of the normalized superconductivity energy gap are due to the coherent spacetime factor $6/\pi$ and that high-T_c superconductors are in fact weakly coupled systems. This paper will generalize the result in equation (30) to account for the weak electron pairing force.

D. Thermodynamic Gauge Functions.

The theory of high-T_c superconductivity is related to a gauge theory of time and energy in bulk matter. This theory is based on the following gauge and conformal invariant renormalization group equations for energy and time[24]

$$\bar{E}' + \bar{\beta}'_{\bar{E}} - 3\bar{\beta}'_{\bar{P}} = \bar{E}^{a'} + \bar{\beta}^{a'}_{\bar{E}} \tag{31}$$

$$\bar{t}' - \bar{\beta}'_{\bar{E}} \, \partial\bar{t}'/\partial\bar{E}' + 3\bar{\beta}'_{\bar{P}} \, \partial\bar{t}'/\partial\bar{P}' = \bar{t}^{a'} - \bar{\beta}^{a'}_{\bar{E}} \, \partial\bar{t}^{a'}/\partial\bar{E}^{a'} \tag{32}$$

where $\bar{E}'$ and $\bar{E}^{a'}$ = renormalized and unrenormalized average energy densities respectively, $\bar{t}'$ and $\bar{t}^{a'}$ = renormalized and unrenormalized time intervals respectively, and where the gauge functions $\bar{\beta}'_{\bar{E}}$ and $\bar{\beta}'_{\bar{P}}$ are given by[25,26]

$$\bar{\beta}'_{\bar{E}} = T/V'(d\bar{U}'/dT)_{\bar{P}'V'} \qquad \bar{\beta}'_{\bar{P}} = d/dV'(\bar{P}'V')_{\bar{U}'} \tag{33}$$

where $\bar{P}'$ = renormalized pressure and V' is given by equation (11). The corresponding unrenormalized gauge functions are[25,26]

$$\bar{\beta}^{a'}_{\bar{E}} = T/V'(d\bar{U}^{a'}/dT)_{\bar{P}^{a'}V'} \qquad \bar{\beta}^{a'}_{\bar{P}} = d/dV'(\bar{P}^{a'}V')_{\bar{U}^{a'}} \tag{34}$$

The average energy densities that appear in equations (31) and (32) are defined by

$$\bar{E}' = \bar{U}'/V' \qquad \bar{E}^{a'} = \bar{U}^{a'}/V' \tag{35}$$

and therefore equation (31) yields the renormalized internal energy for partially broken symmetry space.

For the special case of coherent space with $\beta_{VV} = \pi/2$ it follows from equation (11) that $V' = V\theta_V$ where V = constant and the average energy densities in equation (35) become the coherent space average energy densities defined by

$$\bar{E}^{cs} = \bar{U}^{cs}/(V\theta_V) \qquad \bar{E}^{csa} = \bar{U}^{csa}/(V\theta_V) \tag{36}$$

where cs = coherent space. For this case the gauge functions become

$$\bar{\beta}_{\bar{E}}^{cs} = T/(V\theta_V)(d\bar{U}^{cs}/dT)_{\bar{P}^{cs}V\theta_V} \qquad \bar{\beta}_P^{cs} = d/d\theta_V(\bar{P}^{cs}\theta_V)_{\bar{U}^{cs}} \tag{37}$$

The internal energy $\bar{U}^{cs}$ is then calculated from a trace equation of the form in equation (31) but now using the energy densities and gauge functions of equations (36) and (37) respectively.

For totally coherent matter with coherent energy and coherent spacetime the average energy densities are defined as

$$\bar{E}^{tc} = \bar{U}^{tc}/(V\theta_V) \qquad \bar{E}^{tca} = \bar{U}^{tca}/(V\theta_V) \tag{37A}$$

where tc = total coherence, while the gauge functions are written as

$$\bar{\beta}_{\bar{E}}^{tc} = T/(V\theta_V)(d\bar{U}^{tc}/dT)_{\bar{P}^{tc}V\theta_V} \qquad \bar{\beta}_P^{tc} = d/d\theta_V(\bar{P}^{tc}\theta_V)_{\bar{U}^{tc}} \tag{38}$$

$$= j\bar{E}^{tc}(Td\theta_U/dT)_{\bar{P}^{tc}V\theta_V}$$

The internal energy $\bar{U}^{tc}$ is obtained as a solution to an equation analogous to equation (31) but with the energy densities and gauge functions given by equations (37A) and (38) respectively. In this case $\beta_{VV} = \pi/2$, $\beta_{UU} = \pi/2$ and V and U are constants.

For the special case of incoherent spacetime with $\beta_{VV} = 0$ equation (11) gives $V' = V$ where V is now a variable, and the average energy densities in equation (35) become

$$\bar{E} = \bar{U}/V \qquad \bar{E}^a = \bar{U}^a/V \tag{39}$$

and the gauge functions become the more standard expressions[25,26]

$$\bar{\beta}_{\bar{E}} = T/V(d\bar{U}/dT)_{\bar{P}V} \qquad \bar{\beta}_P = d/dV(\bar{P}V)_{\bar{U}} \tag{40}$$

which are valid for incoherent spacetime.

This paper generalizes the coherent spacetime theory of high-T_c superconductivity by introducing a lattice structure dependent internal phase angle of the coherent relative acceleration of the electrons in a Cooper pair, and thereby develops a structure dependent expression for the normalized superconductivity energy gap for high-T_c superconductors. Also considered are thermodynamic processes that occur in the partially coherent energy conditions and partially coherent spacetime state that are associated with the normal state of a high-T_c

superconductor. Application is then made to the photoemission from the superconducting and normal states of high-T_c materials. It is shown that the leading term of the photoelectric current is linear in T for the superconducting state with $T < T_c$, while for the normal state with $T > T_c$ the leading term of the photoelectric current is quadratic in T in agreement with the quadratic dependence on temperature that is predicted by the conventional Fowler theory of the photoemission from ordinary metals. The paper is arranged as follows: Section 2 deals with the coherent spacetime theory of the normalized superconductivity energy gap, Section 3 considers thermodynamic processes in the normal state of high-T_c materials, and Section 4 develops the theory of photoemission from the normal and superconducting states of high-T_c substances.

2. COHERENT SPACETIME THEORY OF HIGH-T_c SUPERCONDUCTIVITY. This section presents a coherent spacetime theory of the motion of electrons in a Cooper pair that can be used to determine the effects of the electron-phonon interaction (lattice structure effects) on the value of the normalized superconductivity energy gap for high-T_c superconductors.

A. Cooper Pairs in Coherent Spacetime.

In the BCS theory, superconductivity is associated with a broken gauge symmetry that is related to the phonon-mediated formation of Cooper electron pairs.[1-11] The analysis presented in this section does not depend on a particular type of mechanism that mediates the electron pairing. The only requirement for the calculations in this section is that the electrons in the Cooper pairs are in a coherent spacetime state. The electrons within a Cooper pair experience an attractive interaction due to the electron-phonon coupling. The electron pairs themselves can interact with each other with an inter-pair force $\bar{F}_{p\chi}$ in the χ direction. Then the equation of motion of an electron pair is

$$\bar{F}_{p\chi} = m_e d^2\bar{\chi}_1/d\bar{t}_1^2 + m_e d^2\bar{\chi}_2/d\bar{t}_2^2 \tag{41}$$

where $\chi = x, y, z$, m_e = electron mass, and where the subscripts 1 and 2 designate each of the two electrons of the pair. Note that the inter-electron force cancels out of equation (41). For the simple case of zero inter-pair forces, $\bar{F}_{p\chi} = 0$ and equation (41) becomes for a free pair of interacting electrons (or holes)

$$d^2\bar{\chi}_1/d\bar{t}_1^2 + d^2\bar{\chi}_2/d\bar{t}_2^2 = 0 \tag{42}$$

Note that both the space and time coordinates are local to each electron.

Consider now the case of coherent spacetime in which the space and time coordinates have the following differentials[25]

$$\begin{aligned} d\bar{t}_1 &= j\bar{t}_1 d\theta_{t1} & d\bar{\chi}_1 &= j\bar{\chi}_1 d\theta_{\chi 1} \\ d\bar{t}_2 &= j\bar{t}_2 d\theta_{t2} & d\bar{\chi}_2 &= j\bar{\chi}_2 d\theta_{\chi 2} \end{aligned} \tag{43A}$$

so that the electron speeds in coherent spacetime are given as[25]

$$\bar{v}^{cs}_{\chi 1} = d\bar{\chi}_1/d\bar{t}_1 = \bar{\chi}_1/\bar{t}_1 \; d\theta_{\chi 1}/d\theta_{t1} = \chi_1/t_1 \; d\theta_{\chi 1}/d\theta_{t1} \; e^{j(\theta_{\chi 1}-\theta_{t1})}$$
$$\bar{v}^{cs}_{\chi 2} = d\bar{\chi}_2/d\bar{t}_2 = \bar{\chi}_2/\bar{t}_2 \; d\theta_{\chi 2}/d\theta_{t2} = \chi_2/t_2 \; d\theta_{\chi 2}/d\theta_{t2} \; e^{j(\theta_{\chi 2}-\theta_{t2})} \tag{43B}$$

where χ_1, χ_2, t_1 and t_2 are all constants. Then a simple calculation shows that the coherent spacetime acceleration of the particles is given by

$$\bar{a}^{cs}_{\chi 1} = d^2\bar{\chi}_1/d\bar{t}_1^2 = \bar{\chi}_1/\bar{t}_1^2(C_{\chi 1} - jD_{\chi 1}) \tag{44}$$

$$\bar{a}^{cs}_{\chi 2} = d^2\bar{\chi}_2/d\bar{t}_2^2 = \bar{\chi}_2/\bar{t}_2^2(C_{\chi 2} - jD_{\chi 2}) \tag{45}$$

where χ_1, χ_2, t_1 and t_2 are constants and where

$$C_{\chi 1} = d\theta_{\chi 1}/d\theta_{t1}(d\theta_{\chi 1}/d\theta_{t1} - 1) \qquad C_{\chi 2} = d\theta_{\chi 2}/d\theta_{t2}(d\theta_{\chi 2}/d\theta_{t2} - 1) \tag{46}$$

$$D_{\chi 1} = d^2\theta_{\chi 1}/d\theta_{t1}^2 \qquad D_{\chi 2} = d^2\theta_{\chi 2}/d\theta_{t2}^2 \tag{47}$$

Then combining equations (41) through (47) gives

$$\bar{F}_{p\chi}/m_e = \bar{\chi}_1/\bar{t}_1^2(C_{\chi 1} - jD_{\chi 1}) + \bar{\chi}_2/\bar{t}_2^2(C_{\chi 2} - jD_{\chi 2}) \tag{48}$$

where $\chi = x, y, z$. If the electron pairs are themselves noninteracting then $\bar{F}_{p\chi} = 0$ in equation (48).

Note that $\bar{t}_1 \neq \bar{t}_2$ (or $t_1 \neq t_2$ and $\theta_{t1} \neq \theta_{t2}$) because the two electrons of the pair are situated in different locations within the solid lattice and are therefore located in regions of different energy density and pressure of the lattice so that by the fundamental time equation (32) it follows that $\bar{t}_1 \neq \bar{t}_2$. In general $\bar{t}_1 \neq \bar{t}_2$ because the crystal lattice is anisotropic and inhomogeneous at the atomic scale. According to the gauge theory of time as represented by equation (32) time is a function of local energy density and pressure which varies throughout the solid lattice on an atomic scale.[24]

For the case of interacting pairs of interacting electrons equation (48) can be written as

$$\bar{F}_{p\chi}/m_e = \chi_1/t_1^2 A_{\chi 1} \; e^{j(\theta_{\chi 1}-2\theta_{t1}-\theta_{a\chi 1})} + \chi_2/t_2^2 A_{\chi 2} \; e^{j(\theta_{\chi 2}-2\theta_{t2}-\theta_{a\chi 2})} \tag{49}$$

where $\chi = x, y, z$ and where

$$A_{\chi 1} = (C_{\chi 1}^2 + D_{\chi 1}^2)^{1/2} \tag{50}$$

$$A_{\chi 2} = (C_{\chi 2}^2 + D_{\chi 2}^2)^{1/2} \tag{51}$$

$$\tan \theta_{a\chi 1} = D_{\chi 1}/C_{\chi 1} \tag{52}$$

$$\tan \theta_{a\chi 2} = D_{\chi 2}/C_{\chi 2} \tag{53}$$

The standard way of solving equation (49) is to take the real and imaginary parts of equation (49) and relate the force and acceleration terms. But this leads to complicated expressions which are difficult to use. A simpler way is to use the approximation that the internal phase angles of each of the three component terms of equation (49) are equal. This gives the following approximate solution

$$F_{p\chi}/m_e \sim A_{\chi 1}\chi_1/t_1^2 - A_{\chi 2}\chi_2/t_2^2 \tag{54}$$

$$\theta_{Fp\chi} \sim \theta_{\chi 1} - 2\theta_{t1} - \theta_{a\chi 1} \tag{55}$$

$$\sim \theta_{\chi 2} - 2\theta_{t2} - \theta_{a\chi 2} - \pi$$

Define the following phase angle differences

$$\theta_\chi = \theta_{\chi 2} - \theta_{\chi 1} - \pi \tag{56}$$

$$\theta_t = \theta_{t2} - \theta_{t1} \tag{57}$$

$$\theta_{a\chi} = \theta_{a\chi 2} - \theta_{a\chi 1} \tag{58}$$

then equation (55) can be written as

$$\theta_\chi - 2\theta_t \sim \theta_{a\chi} \tag{59}$$

For the case when the inter-pair interaction force is zero the following exact equations are valid

$$A_{\chi 1}\chi_1/t_1^2 = A_{\chi 2}\chi_2/t_2^2 \tag{54A}$$

$$\theta_{\chi 1} - 2\theta_{t1} - \theta_{a\chi 1} = \theta_{\chi 2} - 2\theta_{t2} - \theta_{a\chi 2} - \pi \tag{55A}$$

$$\theta_\chi - 2\theta_t = \theta_{a\chi} \tag{59A}$$

Note that if $\theta_{\chi 1}$ is in the first quadrant then $\theta_{\chi 2}$ is in the third quadrant. The phase angle $\theta_{a\chi}$, defined by equation (58), is the relative phase angle between the accelerations of the two electrons in a Cooper pair, and can be a positive or negative number. Equations (43A) through (59A) describe the motion of coherent spacetime electrons in a Cooper pair of a high-T_c superconductor.

B. Normalized Superconductivity Energy Gap.

Combining the coherent spacetime superconductivity condition given in equation (28) with the coherent spacetime condition for Cooper electron pair acceleration given in equation (59) yields

$$\theta_t = \pi/6(1 - 4/\pi\ \theta_{a\chi}) \tag{61}$$

$$\theta_\chi = \pi/3(1 - 1/\pi\ \theta_{a\chi}) \tag{62}$$

where $\chi = x, y, z$. These equations reduce to the previously obtained values of θ_t and θ_χ given in equation (29) if the relative internal phase angle of the electron accelerations given in equation (58) is set equal to zero as $\theta_{a\chi} = 0$. For $\theta_{a\chi} = 0$ equations (61) and (62) reduce to the results obtained in Reference 23. For an isotropic system $\theta_{a\chi} = \theta_a$ for $\chi = x, y, z$. The fact that $\theta_{ax} \neq \theta_{ay} \neq \theta_{az}$ is an indication that each electron of a Cooper pair is located in a different region of local energy density and pressure of the crystal lattice. The departure of the values of the internal phase angles of time and space from the values given in equation (29) is a measure of the degree of anisotropy of the electron-phonon interaction due to the anisotropy of the atomic structure of a high-T_c material.

The coherent time and incoherent time normalized superconductivity energy gaps are given by[23,24]

$$\Delta'_{it}\, t = \tau \qquad \Delta'_{ct\chi}\, t\theta_t = \tau \qquad \tau = \hbar^3/(m_e e^4) \tag{63}$$

where Δ'_{it} = incoherent time normalized (BCS) superconductivity energy gap given by equation (2), $\Delta'_{ct\chi}$ = coherent time normalized superconductivity energy gap in the χ direction = measured normalized superconductivity energy gap in the χ direction, and where τ = Bohr time or the characteristic time of an electron in a Bohr orbit about the other electron in a Cooper pair. Combining equations (61) and (63) gives

$$\Delta'_{ct\chi}/\Delta'_{it} = 1/\theta_t = 6/\pi(1 - 4/\pi\ \theta_{a\chi})^{-1} \tag{64}$$

The value of the coherent time normalized superconductivity energy gap is then obtained from equations (2) and (64) to be

$$\begin{aligned} \Delta'_{ct\chi} &= (6/\pi)(3.52)(1 - 4/\pi\ \theta_{a\chi})^{-1} \\ &= 6.72(1 - 4/\pi\ \theta_{a\chi})^{-1} \\ &\sim 6.72(1 + 4/\pi\ \theta_{a\chi} + \cdots) \end{aligned} \tag{65}$$

where the approximation in equation (65) holds only for small values of $\theta_{a\chi}$. The Bohr time τ does not enter the expression for the normalized superconductivity energy gap given by equation (65).

The value of $\theta_{a\chi}$ can be obtained from equations (64) and (65) to be

$$\begin{aligned} \theta_{a\chi} &= \pi/4(1 - 6/\pi\ \Delta'_{it}/\Delta'_{ct\chi}) \\ &= \pi/4(1 - 6.72/\Delta'_{ct\chi}) \end{aligned} \tag{66}$$

Equation (66) can be used to determine $\theta_{a\chi}$ from the measured values of the normalized superconductivity energy gap in the χ direction $\Delta'_{ct\chi}$. The range of the measured values of $\Delta'_{ct\chi}$ goes from 2 through 10 depending on material type.[1-11,14,19-22] Values of $\theta_{a\chi}$ for selected values of $\Delta'_{ct\chi}$ are evaluated from equation

(66) as follows

$\Delta'_{ct\chi}$	$\theta_{a\chi}$, rad
2	- 1.85
3.52	- 0.71
4	- 0.53
6.72	0
8	+ 0.13
10	+ 0.26

The values of $\theta_{a\chi}$ can be positive or negative. The factor $6/\pi$ that occurs in equations (64) through (66) is responsible for the large measured values of $\Delta'_{ct\chi}$ relative to the BCS normalized superconductivity energy gap value of 3.52 . Therefore the large values of $\Delta'_{ct\chi}$ are due to the coherent time state associated with high-T_c superconductivity, and are not associated with strong couplings of the electrons in a Cooper pair. Also a measured value of $\Delta'_{ct\chi} \sim 3.52$ for a high-T_c superconductor does not imply a BCS superconductivity mechanism but only that the relative phase angle of the electron acceleration in the Cooper pair has a value $\theta_{a\chi} \sim - 0.71$. The small values of $\theta_{a\chi}$ that occur for relatively large values of $\Delta'_{ct\chi}$ suggest that the electron-electron pairing interaction is weak. The fact that the electron-electron coupling in Cooper pairs is weak is utilized in Section 4 where the theory of the photoelectric effect in high-T_c superconductors is considered.

3. THERMODYNAMIC PROCESSES IN THE NORMAL STATE OF HIGH-T_c SUPERCONDUCTORS. This section considers the possible thermodynamic processes and spacetime states for the normal state ($T > T_c$) of a high-T_c superconductor. The normal states of organic, heavy fermion, and copper oxide high-T_c superconductors have peculiar experimental properties and may not be describable as an ordinary Fermi gas ground state because of the presence of antiferromagnetism and the extreme lack of isotropy of the electrical properties such as resistivity, magnetic penetration depth and correlation length.[1-11,13-15] For instance, parallel to the CuO planes the conductivity is metallic while in the perpendicular direction it is like a semiconductor.[1-11] Thus the normal state of a high-T_c superconductor needs a scientific investigation along with the superconducting state. This section considers the general case of partially coherent thermodynamic states associated with partially coherent spacetime states. This describes thermodynamic processes occurring in the normal state of a high-T_c superconductor because the normal state with $T > T_c$ is assumed to be in a partially coherent spacetime state. The superconducting state with $T < T_c$ is associated with complete spacetime coherence and is a special case of the calculations done in this section. The BCS state is taken to be an incoherent spacetime state.

A. Energy Density and Entropy Density.

This subsection calculates the energy densities and entropy densities of a partially coherent thermodynamic state of the normal, partially coherent spacetime, state of a high-T_c superconductor. The combined first and second

laws of thermodynamics are written for this case as[25]

$$T d\bar{\bar{S}} = d\bar{U} + \bar{\bar{P}} d\bar{V} + \bar{\bar{M}} d\bar{\alpha} \tag{67}$$

$$= d\bar{U} + \bar{P}|d\bar{V}| + \bar{M}|d\bar{\alpha}|$$

where

$$\bar{S} = Se^{j\theta S} \qquad \bar{U} = Ue^{j\theta U} \qquad \bar{\bar{P}} = Pe^{j\theta'_P} \tag{68}$$

$$\bar{\bar{M}} = Me^{j\theta'_M} \qquad \bar{M} = Me^{j\theta M} \qquad \bar{P} = Pe^{j\theta P} \tag{69}$$

$$\theta'_P = \theta_P - \theta_V - \beta_{VV} \tag{70}$$

$$\theta'_M = \theta_M - \theta_\alpha - \beta_{\alpha\alpha} \tag{71}$$

where β_{VV} is defined in equation (12) and $\beta_{\alpha\alpha}$ is given by

$$\tan \beta_{\alpha\alpha} = \alpha \partial\theta_\alpha / \partial\alpha \tag{72}$$

The differential entropy density for broken symmetry thermodynamics and broken symmetry spacetime is written as

$$\bar{\bar{s}}_{bs} = d\bar{S}/d\bar{V} = e^{j(\theta S - \theta V)} (dS + jSd\theta_S)/(dV + jVd\theta_V) \tag{73}$$

The broken symmetry differential entropy density can be written as

$$\bar{\bar{s}}_{bs} = s_{bs} e^{j\Phi SV} \qquad s_{bs} = |d\bar{S}|/|d\bar{V}| \tag{74}$$

where

$$s_{bs} = \sec \beta_{SS} \cos \beta_{VV} s_{inc} \tag{75A}$$

$$= \csc \beta_{SS} \sin \beta_{VV} s_{tc} \tag{75B}$$

$$= \sec \beta_{SS} \sin \beta_{VV} s_{cs} \tag{75C}$$

$$= \csc \beta_{SS} \cos \beta_{VV} s_{cth} \tag{75D}$$

where the following entropy densities are defined

$$s_{inc} = dS/dV \tag{76}$$

$$s_{tc} = S_{tc}/V \; \partial\theta_S/\partial\theta_V = \mathcal{S}_{tc} \; \theta_V \partial\theta_S/\partial\theta_V \tag{77}$$

$$s_{cs} = 1/V \; \partial S_{cs}/\partial\theta_V = \partial/\partial\theta_V(\theta_V \mathcal{S}_{cs}) \tag{78}$$

$$s_{cth} = S_{cth} \; \partial\theta_S/\partial V \tag{79}$$

$$\mathcal{S} = S/V \qquad \mathcal{S}' = S'/V' \tag{80}$$

$$\mathcal{S}_{cs} = S_{cs}/(V\theta_V) \qquad \mathcal{S}_{tc} = S_{tc}/(V\theta_V) \tag{81}$$

where s_{inc} = incoherent differential entropy density, s_{tc} = total coherence differential entropy density, s_{cs} = coherent space differential entropy density, s_{cth} = coherent thermodynamics differential entropy density, $\mathcal{S}$ = incoherent average entropy density, $\mathcal{S}'$ = average energy density, $\mathcal{S}_{cs}$ = coherent spacetime average entropy density, and $\mathcal{S}_{tc}$ = total coherence average entropy density, and where

$$\Phi_{SV} = \theta_S + \beta_{SS} - \theta_V - \beta_{VV} = \Phi_S - \Phi_V \tag{82}$$

$$\Phi_S = \theta_S + \beta_{SS} \qquad \Phi_V = \theta_V + \beta_{VV} \tag{83}$$

In equation (81) $V\theta_V$ is given by equation (11) with $\beta_{VV} = \pi/2$, and in equation (83)

$$\tan \beta_{SS} = S\partial\theta_S/\partial S \tag{84}$$

Sometimes it is convenient to work with an entropy density that does not include the internal phase angle of the volume so that instead of equation (73) the following entropy density is introduced

$$\bar{s}_{bs} = d\bar{S}/|d\bar{V}| = s_{bs}\, e^{j\Phi S} \qquad \bar{\bar{s}}_{bs} = \bar{s}_{bs}\, e^{-j\Phi V} \tag{85}$$

The superconducting state ($T < T_c$) has $\beta_{VV} = \pi/2$ in equations (75B) and (75C).

The broken symmetry energy density for partially coherent internal energy and partially coherent spacetime is

$$\bar{\bar{e}}_{bs} = e_{bs}\, e^{j\Phi UV} = d\bar{\bar{U}}/d\bar{\bar{V}} = e^{j(\theta_U - \theta_V)}(dU + jUd\theta_U)/(dV + jVd\theta_V) \tag{86}$$

where the energy density magnitude is written as

$$e_{bs} = |d\bar{\bar{U}}|/|d\bar{\bar{V}}| \tag{87}$$

and can have the following representations

$$e_{bs} = \sec \beta_{UU} \cos \beta_{VV}\, e_{inc} \tag{88A}$$

$$= \csc \beta_{UU} \sin \beta_{VV}\, e_{tc} \tag{88B}$$

$$= \sec \beta_{UU} \sin \beta_{VV}\, e_{cs} \tag{88C}$$

$$= \csc \beta_{UU} \cos \beta_{VV}\, e_{cth} \tag{88D}$$

where the following energy densities are defined

$$e_{inc} = \partial U/\partial V \tag{89}$$

$$e_{tc} = U_{tc}/V\ \partial\theta_U/\partial\theta_V = E_{tc}\ \theta_V \partial\theta_U/\partial\theta_V \tag{90}$$

$$e_{cs} = 1/V\ \partial U_{cs}/\partial\theta_V = \partial/\partial\theta_V(\theta_V E_{cs}) \tag{91}$$

$$e_{cth} = U_{cth}\, \partial\theta_U/\partial V \tag{92}$$

$$E = U/V \qquad E' = U'/V' \tag{93}$$

$$E_{cs} = U_{cs}/(V\theta_V) \qquad E_{tc} = U_{tc}/(V\theta_V) \tag{94}$$

where e_{inc} = incoherent differential energy density, e_{tc} = total coherent differential energy density, e_{cs} = coherent space differential energy density, e_{cth} = coherent thermodynamic differential energy density, E = incoherent average energy density, E' = average energy density, E_{cs} = coherent spacetime average energy density, E_{tc} = total coherence average energy density, and where

$$\Phi_{UV} = \theta_U + \beta_{UU} - \theta_V - \beta_{VV} = \Phi_U - \Phi_V \tag{95}$$

$$\Phi_U = \theta_U + \beta_{UU} \tag{96}$$

$$\tan \beta_{UU} = U\partial\theta_U/\partial U \tag{97}$$

An energy density can be defined that excludes the internal phase angle of space as follows

$$\bar{e}_{bs} = d\bar{U}/|d\bar{V}| = e_{bs}\, e^{j\Phi U} \qquad \bar{\bar{e}}_{bs} = \bar{e}_{bs}\, e^{-j\Phi V} \tag{98}$$

The values of U, U', U_{tc}, U_{cs} and U_{cth} are evaluated from their corresponding renormalization group trace equations of the general form given in equation (31). The superconducting state ($T < T_c$) has $\beta_{VV} = \pi/2$ in equations (88B) and (88C).

The generalized coordinate density can be written as

$$\bar{\bar{a}}_{bs} = d\bar{\alpha}/d\bar{V} = e^{j(\theta_\alpha - \theta_V)}(d\alpha + j\alpha d\theta_\alpha)/(dV + jVd\theta_V) \tag{99}$$

which can also be represented as

$$\bar{\bar{a}}_{bs} = a_{bs}\, e^{j\Phi\alpha V} \qquad a_{bs} = |d\bar{\alpha}|/|d\bar{V}| \tag{100}$$

where

$$a_{bs} = \sec \beta_{\alpha\alpha} \cos \beta_{VV}\, a_{inc} \tag{101A}$$

$$= \csc \beta_{\alpha\alpha} \sin \beta_{VV}\, a_{tc} \tag{101B}$$

where the following generalized coordinate densities are defined

$$a_{inc} = d\alpha/dV \tag{102}$$

$$a_{tc} = \alpha_{tc}/V\ \partial\theta_\alpha/\partial\theta_V = A_{tc}\, \theta_V \partial\theta_\alpha/\partial\theta_V \tag{103}$$

$$A = \alpha/V \qquad A' = \alpha'/V' \tag{104}$$

$$A_{tc} = \alpha_{tc}/(V\theta_V) \tag{105}$$

where a_{inc} = incoherent differential generalized coordinate density, a_{tc} = totally coherent differential generalized coordinate density, A = incoherent average generalized coordinate density, A' = average generalized coordinate density, A_{tc} = total coherence average generalized coordinate density, and where

$$\Phi_{\alpha V} = \theta_\alpha + \beta_{\alpha\alpha} - \theta_V - \beta_{VV} = \Phi_\alpha - \Phi_V \tag{106}$$

$$\Phi_\alpha = \theta_\alpha + \beta_{\alpha\alpha} \tag{107}$$

An alternative definition of the generalized coordinate density that does not contain the internal phase angle of the volume is given by

$$\bar{a}_{bs} = d\bar{\alpha}/|d\bar{V}| = a_{bs}\, e^{j\Phi\alpha} \qquad \bar{\bar{a}}_{bs} = \bar{a}_{bs}\, e^{-j\Phi V} \tag{108}$$

Note that it is assumed that $\bar{V}$ and $\bar{\alpha}$ behave in a homologous way so that if $\bar{V}$ is incoherent then so is $\bar{\alpha}$, and when $\bar{V}$ is coherent then so is $\bar{\alpha}$. This is why only two representations of a_{bs} appear in equation (101). The superconducting state ($T < T_c$) would have $\beta_{VV} = \pi/2$ and $\beta_{\alpha\alpha} = \pi/2$ in equation (101B), while the BCS state would have $\beta_{VV} = 0$ and $\beta_{\alpha\alpha} = 0$ in equation (101A).

B. Pressure for Partially Coherent Thermodynamics and Partially Coherent Spacetime.

This subsection determines the pressure of a partially coherent thermodynamic system in the normal state of a high-T_c superconductor. The pressure is obtained from equation (67) as

$$\bar{\bar{P}} = Pe^{j\theta'_P} = Td\bar{S}/d\bar{V} - d\bar{U}/d\bar{V} - \bar{\bar{M}}d\bar{\alpha}/d\bar{V} \tag{109}$$

$$\bar{P} = Pe^{j\theta_P} = Td\bar{S}/|d\bar{V}| - d\bar{U}/|d\bar{V}| - \bar{M}|d\bar{\alpha}|/|d\bar{V}| \tag{110}$$

so that

$$\bar{\bar{P}} = \bar{P}e^{-j\Phi V} \qquad \theta'_P = \theta_P - \Phi_V \tag{111}$$

The pressure $\bar{P}$ can be obtained from equation (110) to be

$$\bar{P} = Ts_{bs}\, e^{j\Phi S} - e_{bs}\, e^{j\Phi U} - Ma_{bs}\, e^{j\Phi M} \tag{112}$$

which can be rewritten using equations (75) through (78) and (88) through (92) as follows

$$\bar{P} = \cos\beta_{VV}\,(Ts_{inc} \sec\beta_{SS}\, e^{j\Phi S} - e_{inc} \sec\beta_{UU}\, e^{j\Phi U} - Ma_{inc} \sec\beta_{\alpha\alpha}\, e^{j\theta M}) \tag{113}$$

$$= \sin\beta_{VV}\,(Ts_{tc} \csc\beta_{SS}\, e^{j\Phi S} - e_{tc} \csc\beta_{UU}\, e^{j\Phi U} - Ma_{tc} \csc\beta_{\alpha\alpha}\, e^{j\theta M}) \tag{114}$$

$$= \sin\beta_{VV}\,(Ts_{cs} \sec\beta_{SS}\, e^{j\Phi S} - e_{cs} \sec\beta_{UU}\, e^{j\Phi U} - Ma_{tc} \csc\beta_{\alpha\alpha}\, e^{j\theta M}) \tag{115}$$

$$= \cos\beta_{VV}\,(Ts_{cth} \csc\beta_{SS}\, e^{j\Phi S} - e_{cth} \csc\beta_{UU}\, e^{j\Phi U} - Ma_{inc} \sec\beta_{\alpha\alpha}\, e^{j\theta M}) \tag{116}$$

From equations (68) and (69) it follows that

$$d\bar{P} = \sec \beta_{PP}\, dP\, e^{j\Phi_P} = \csc \beta_{PP}\, Pd\theta_P\, e^{j\Phi_P} \tag{117}$$

$$d\bar{\bar{P}} = \sec \beta_{PP}\, dP\, e^{j(\Phi_P-\Phi_V)} = \csc \beta_{PP}\, Pd\theta_P\, e^{j(\Phi_P-\Phi_V)} \tag{118}$$

where

$$\tan \beta_{PP} = P\partial\theta_P/\partial P \tag{119}$$

$$\Phi_P = \theta_P + \beta_{PP} \tag{120}$$

Equations (113) through (116) give four fundamental representations of the pressure in the normal state of a high-T_c superconductor.

From equation (67) it follows that

$$T\partial\bar{S}/\partial T = \partial\bar{U}/\partial T + \bar{\bar{M}}\partial\bar{\alpha}/\partial T \tag{121}$$

$$T\partial\bar{S}/\partial\bar{V} = \partial\bar{U}/\partial\bar{V} + \bar{\bar{P}} + \bar{\bar{M}}\partial\bar{\alpha}/\partial\bar{V} \tag{122}$$

$$\partial\bar{S}/\partial\bar{V} = \partial/\partial T(\bar{\bar{P}} + \bar{\bar{M}}\partial\bar{\alpha}/\partial\bar{V}) \tag{123}$$

$$\partial\bar{U}/\partial\bar{V} = T\partial T(\bar{\bar{P}} + \bar{\bar{M}}\partial\bar{\alpha}/\partial\bar{V}) - (\bar{\bar{P}} + \bar{\bar{M}}\partial\bar{\alpha}/\partial\bar{V}) \tag{124}$$

Equivalently, equations (122) through (124) can be written as

$$T\partial\bar{S}/|\partial\bar{V}| = \partial\bar{U}/|\partial\bar{V}| + \bar{P} + \bar{M}|\partial\bar{\alpha}|/|\partial\bar{V}| \tag{125}$$

$$\partial\bar{S}/|\partial\bar{V}| = \partial/\partial T(\bar{P} + \bar{M}|\partial\bar{\alpha}|/|\partial\bar{V}|) \tag{126}$$

$$\partial\bar{U}/|\partial\bar{V}| = T\partial/\partial T(\bar{P} + \bar{M}|\partial\bar{\alpha}|/|\partial\bar{V}|) - (\bar{P} + \bar{M}|\partial\bar{\alpha}|/|\partial\bar{V}|) \tag{127}$$

Equations (124) and (127) can be written respectively as

$$\bar{\bar{e}}_{bs} = T\partial/\partial T(\bar{\bar{P}} + \bar{\bar{M}}\bar{\bar{a}}_{bs}) - (\bar{\bar{P}} + \bar{\bar{M}}\bar{\bar{a}}_{bs}) \tag{127A}$$

$$\bar{e}_{bs} = T\partial/\partial T(\bar{P} + \bar{M}a_{bs}) - (\bar{P} + \bar{M}a_{bs}) \tag{127B}$$

where $\bar{M}a_{bs} = \bar{\bar{M}}\bar{\bar{a}}_{bs}$, and where $\bar{\bar{e}}_{bs}$, $\bar{e}_{bs}$, e_{bs}, $\bar{\bar{a}}_{bs}$, $\bar{a}_{bs}$ and a_{bs} are given by equations (86), (98), (88), (100), (108) and (101) respectively. If the generalized coordinate is independent of the volume then $a_{bs} = 0$ and

$$\bar{\bar{e}}_{bs} = T\partial\bar{\bar{P}}/\partial T - \bar{\bar{P}} \tag{127C}$$

$$\bar{e}_{bs} = T\partial\bar{P}/\partial T - \bar{P} \tag{127D}$$

For the totally coherent case with $\beta_{UU} = \pi/2$ and $\beta_{VV} = \pi/2$ it follows from equations (127C), (86), (88B), (90) and (95) that

$$\bar{U}^{tc}/\bar{V}\ \partial\theta_U/\partial\theta_V = T\partial\bar{\bar{P}}/\partial T - \bar{\bar{P}} \qquad j\bar{U}^{tc}/V\ \partial\theta_U/\partial\theta_V = T\partial\bar{P}/\partial T - \bar{P} \tag{127E}$$

From equation (121) it follows approximately after neglecting $d\bar{\alpha}$

$$\sec\beta_{SS}\ T\partial S/\partial T \sim \sec\beta_{UU}\ \partial U/\partial T \tag{128}$$

$$\csc\beta_{SS}\ TS\partial\theta_S/\partial T \sim \csc\beta_{UU}\ U\partial\theta_U/\partial T \tag{129}$$

$$\Phi_S \sim \Phi_U \qquad \theta_S + \beta_{SS} \sim \theta_U + \beta_{UU} \tag{130}$$

Equation (130) gives $\theta_S \sim \theta_U$ for incoherent thermodynamics when $\beta_{SS} = 0$ and $\beta_{UU} = 0$, and for coherent thermodynamics when $\beta_{SS} = \pi/2$ and $\beta_{UU} = \pi/2$. Equation (126) can be written approximately, after neglecting $d\bar{\alpha}$, as follows

$$s_{bs}\ e^{j\Phi_S} \sim \partial P/\partial T \sec\beta_{PP}\ e^{j\Phi_P} \tag{131}$$

$$= P\partial\theta_P/\partial T \csc\beta_{PP}\ e^{j\Phi_P}$$

where equation (85) was used to evaluate the left hand side of equation (126). Equation (131) can be rewritten as

$$s_{bs} \sim \sec\beta_{PP}\ \partial P/\partial T = \csc\beta_{PP}\ P\partial\theta_P/\partial T \tag{132}$$

$$\Phi_S \sim \Phi_U \sim \Phi_P \tag{133}$$

where s_{bs} is given by any of the expressions in equation (75), Φ_S is given by equation (83), Φ_U is given by equation (96), and Φ_P by equation (120).

Equation (67) or equation (112) gives

$$Ts_{bs}\cos\Phi_S = e_{bs}\cos\Phi_U + P\cos\theta_P + Ma_{bs}\cos\theta_M \tag{134}$$

$$Ts_{bs}\sin\Phi_S = e_{bs}\sin\Phi_U + P\sin\theta_P + Ma_{bs}\sin\theta_M \tag{135}$$

which gives the pressure as

$$P^2 = T^2s_{bs}^2 + e_{bs}^2 + M^2a_{bs}^2 - 2Ts_{bs}\,e_{bs}\cos(\Phi_S - \Phi_U) \tag{136}$$

$$- 2TMs_{bs}a_{bs}\cos(\Phi_S - \theta_M) + 2Me_{bs}a_{bs}\cos(\Phi_U - \theta_M)$$

$$\tan\theta_P = A/B \tag{137}$$

$$A = Ts_{bs}\sin\Phi_S - e_{bs}\sin\Phi_U - Ma_{bs}\sin\theta_M \tag{138}$$

$$B = Ts_{bs}\cos\Phi_S - e_{bs}\cos\Phi_U - Ma_{bs}\cos\theta_M \tag{139}$$

Assuming $\theta_M \sim \Phi_S \sim \Phi_U$ in equation (136) gives the magnitude of the pressure approximately as

$$P \sim Ts_{bs} - e_{bs} - Ma_{bs} \tag{140}$$

where s_{bs} is any form in equation (75), e_{bs} is any corresponding expression in equation (88), and a_{bs} is the corresponding value obtained from equation (101). The various expressions for P given in equation (140) can be read directly from equations (113) through (116) by assuming $\theta_P \sim \theta_M \sim \Phi_S \sim \Phi_U$.

From equation (67) or more directly from equations (113) through (116) it follows that the following equivalent pairs of equations are valid

$$Ts_{inc} \sec \beta_{SS} \cos \Phi_S = e_{inc} \sec \beta_{UU} \cos \Phi_U + P \sec \beta_{VV} \cos \theta_P + Ma_{inc} \sec \beta_{\alpha\alpha} \cos \theta_M \tag{141}$$

$$Ts_{inc} \sec \beta_{SS} \sin \Phi_S = e_{inc} \sec \beta_{UU} \sin \Phi_U + P \sec \beta_{VV} \sin \theta_P + Ma_{inc} \sec \beta_{\alpha\alpha} \sin \theta_M \tag{142}$$

$$Ts_{tc} \csc \beta_{SS} \cos \Phi_S = e_{tc} \csc \beta_{UU} \cos \Phi_U + P \csc \beta_{VV} \cos \theta_P + Ma_{tc} \csc \beta_{\alpha\alpha} \cos \theta_M \tag{143}$$

$$Ts_{tc} \csc \beta_{SS} \sin \Phi_S = e_{tc} \csc \beta_{UU} \sin \Phi_U + P \csc \beta_{VV} \sin \theta_P + Ma_{tc} \csc \beta_{\alpha\alpha} \sin \theta_M \tag{144}$$

$$Ts_{cs} \sec \beta_{SS} \cos \Phi_S = e_{cs} \sec \beta_{UU} \cos \Phi_U + P \csc \beta_{VV} \cos \theta_P + Ma_{tc} \csc \beta_{\alpha\alpha} \cos \theta_M \tag{145}$$

$$Ts_{cs} \sec \beta_{SS} \sin \Phi_S = e_{cs} \sec \beta_{UU} \sin \Phi_U + P \csc \beta_{VV} \sin \theta_P + Ma_{tc} \csc \beta_{\alpha\alpha} \sin \theta_M \tag{146}$$

$$Ts_{cth} \csc \beta_{SS} \cos \Phi_S = e_{cth} \csc \beta_{UU} \cos \Phi_U + P \sec \beta_{VV} \cos \theta_P + Ma_{inc} \sec \beta_{\alpha\alpha} \cos \theta_M \tag{147}$$

$$Ts_{cth} \csc \beta_{SS} \sin \Phi_S = e_{cth} \csc \beta_{UU} \sin \Phi_U + P \sec \beta_{VV} \sin \theta_P + Ma_{inc} \sec \beta_{\alpha\alpha} \sin \theta_M \tag{148}$$

For the development of equations (141) through (148) it has been assumed that

the variables $\bar{V}$ and $\bar{\alpha}$ behave similarly in the sense that if one is coherent so is the other, and if one is incoherent then so is the other.

From equations (141) and (142) it follows that

$$P^2 \sec^2\beta_{VV} = T^2 s^2_{inc} \sec^2\beta_{SS} + e^2_{inc} \sec^2\beta_{UU} + M^2 a^2_{inc} \sec^2\beta_{\alpha\alpha} \tag{149}$$
$$- 2Ts_{inc}e_{inc} \sec\beta_{SS} \sec\beta_{UU} \cos(\Phi_S - \Phi_U)$$
$$- 2TMs_{inc}a_{inc} \sec\beta_{SS} \sec\beta_{\alpha\alpha} \cos(\Phi_S - \theta_M)$$
$$+ 2Me_{inc}a_{inc} \sec\beta_{UU} \sec\beta_{\alpha\alpha} \cos(\Phi_U - \theta_M)$$

and

$$\tan\theta_P = C/D \tag{150}$$

$$C = Ts_{inc} \sec\beta_{SS} \sin\Phi_S - e_{inc} \sec\beta_{UU} \sin\Phi_U - Ma_{inc} \sec\beta_{\alpha\alpha} \sin\theta_M \tag{151}$$

$$D = Ts_{inc} \sec\beta_{SS} \cos\Phi_S - e_{inc} \sec\beta_{UU} \cos\Phi_U - Ma_{inc} \sec\beta_{\alpha\alpha} \cos\theta_M \tag{152}$$

Combining equations (143) and (144) gives

$$P^2 \csc^2\beta_{VV} = T^2 s^2_{tc} \csc^2\beta_{SS} + e^2_{tc} \csc^2\beta_{UU} + M^2 a^2_{tc} \csc^2\beta_{\alpha\alpha} \tag{153}$$
$$- 2Ts_{tc}e_{tc} \csc\beta_{SS} \csc\beta_{UU} \cos(\Phi_S - \Phi_U)$$
$$- 2TMs_{tc}a_{tc} \csc\beta_{SS} \csc\beta_{\alpha\alpha} \cos(\Phi_S - \theta_M)$$
$$+ 2Me_{tc}a_{tc} \csc\beta_{UU} \csc\beta_{\alpha\alpha} \cos(\Phi_U - \theta_M)$$

and

$$\tan\theta_P = E/F \tag{154}$$

$$E = Ts_{tc} \csc\beta_{SS} \sin\Phi_S - e_{tc} \csc\beta_{UU} \sin\Phi_U - Ma_{tc} \csc\beta_{\alpha\alpha} \sin\theta_M \tag{155}$$

$$F = Ts_{tc} \csc\beta_{SS} \cos\Phi_S - e_{tc} \csc\beta_{UU} \cos\Phi_U - Ma_{tc} \csc\beta_{\alpha\alpha} \cos\theta_M \tag{156}$$

From equations (145) and (146) it follows that

$$P^2 \csc^2\beta_{VV} = T^2 s^2_{cs} \sec^2\beta_{SS} + e^2_{cs} \sec^2\beta_{UU} + M^2 a^2_{tc} \csc^2\beta_{\alpha\alpha} \tag{157}$$
$$- 2Ts_{cs}e_{cs} \sec\beta_{SS} \sec\beta_{UU} \cos(\Phi_S - \Phi_U)$$
$$- 2TMs_{cs}a_{tc} \sec\beta_{SS} \csc\beta_{\alpha\alpha} \cos(\Phi_S - \theta_M)$$
$$+ 2Me_{cs}a_{tc} \sec\beta_{UU} \csc\beta_{\alpha\alpha} \cos(\Phi_U - \theta_M)$$

and

$$\tan \theta_P = G/H \tag{158}$$

$$G = Ts_{cs} \sec \beta_{SS} \sin \Phi_S - e_{cs} \sec \beta_{UU} \sin \Phi_U - Ma_{tc} \csc \beta_{\alpha\alpha} \sin \theta_M \tag{159}$$

$$H = Ts_{cs} \sec \beta_{SS} \cos \Phi_S - e_{cs} \sec \beta_{UU} \cos \Phi_U - Ma_{tc} \csc \beta_{\alpha\alpha} \cos \theta_M \tag{160}$$

Finally, combining equations (147) and (148) gives

$$\begin{aligned} P^2 \sec^2 \beta_{VV} = {} & T^2 s_{cth}^2 \csc^2 \beta_{SS} + e_{cth}^2 \csc^2 \beta_{UU} + M^2 a_{inc}^2 \sec^2 \beta_{\alpha\alpha} \\ & - 2Ts_{cth} e_{cth} \csc \beta_{SS} \csc \beta_{UU} \cos(\Phi_S - \Phi_U) \\ & - 2TMs_{cth} a_{inc} \csc \beta_{SS} \sec \beta_{\alpha\alpha} \cos(\Phi_S - \theta_M) \\ & + 2Me_{cth} a_{inc} \csc \beta_{UU} \sec \beta_{\alpha\alpha} \cos(\Phi_U - \theta_M) \end{aligned} \tag{161}$$

and

$$\tan \theta_P = I/J \tag{162}$$

$$I = Ts_{cth} \csc \beta_{SS} \sin \Phi_S - e_{cth} \csc \beta_{UU} \sin \Phi_U - Ma_{inc} \sec \beta_{\alpha\alpha} \sin \theta_M \tag{163}$$

$$J = Ts_{cth} \csc \beta_{SS} \cos \Phi_S - e_{cth} \csc \beta_{UU} \cos \Phi_U - Ma_{inc} \sec \beta_{\alpha\alpha} \cos \theta_M \tag{164}$$

Equation (130) shows that $\Phi_S \sim \Phi_U$ and this simplifies the application of equations (141) through (164). An approximate solution for equations (141) through (148) can be found by taking

$$\theta_P \sim \theta_M \sim \Phi_U \sim \Phi_S \tag{165}$$

Combining equations (133) and (165) gives $\beta_{PP} \sim 0$ which must hold if θ_P is essentially constant. The constant phase angle conditions in equation (165) simplify equations (149), (153), (157) and (161) so that the magnitude of the pressure is given approximately by any of the following equations

$$P \sim \cos \beta_{VV} (Ts_{inc} \sec \beta_{SS} - e_{inc} \sec \beta_{UU} - Ma_{inc} \sec \beta_{\alpha\alpha}) \tag{166}$$

$$\sim \sin \beta_{VV} (Ts_{tc} \csc \beta_{SS} - e_{tc} \csc \beta_{UU} - Ma_{tc} \csc \beta_{\alpha\alpha}) \tag{167}$$

$$\sim \sin \beta_{VV} (Ts_{cs} \sec \beta_{SS} - e_{cs} \sec \beta_{UU} - Ma_{tc} \csc \beta_{\alpha\alpha}) \tag{168}$$

$$\sim \cos \beta_{VV} (Ts_{cth} \csc \beta_{SS} - e_{cth} \csc \beta_{UU} - Ma_{inc} \sec \beta_{\alpha\alpha}) \tag{169}$$

which can also be determined from equation (113) through (116) by using the approximate conditions in equation (165).

From equations (130) and (133) it follows that

$$\theta_P + \tan^{-1}(P \partial\theta_P / \partial P) \sim k \tag{170}$$

$$\theta_S + \tan^{-1}(S \partial\theta_S / \partial S) \sim k \tag{171}$$

$$\theta_U + \tan^{-1}(U \partial\theta_U / \partial U) \sim k \tag{172}$$

or for small angles

$$\theta_P + P\partial\theta_P/\partial P \sim k \tag{173}$$

$$\theta_S + S\partial\theta_S/\partial S \sim k \tag{174}$$

$$\theta_U + U\partial\theta_U/\partial U \sim k \tag{175}$$

The solutions to equations (173) through (175) are given by

$$\theta_P \sim k + C_P/P \qquad \theta_S \sim k + C_S/S \qquad \theta_U \sim k + C_U/U \tag{176}$$

where k, C_P, C_S and C_U = constants. This suggests that the internal phase angles of the thermodynamic functions tend to vary inversely with magnitudes of the thermodynamic functions.

4. PHOTOELECTRIC EFFECT IN HIGH-T_c SUPERCONDUCTORS. The standard calculation of the photoelectric current in terms of the temperature of the metal which is irradiated by photons was given by Fowler who performed a statistical mechanical treatment of the gas of free electrons that is assumed to be responsible for the electrical properties of metals.[27] This calculation is well known and proceeds by calculating the partition function for a Fermi gas at finite temperature and leads to the following result for the photoelectric current[27]

$$I = A\alpha T/(m_e k) \int_{p_M}^{\infty} p \, \ell n[1 + \mu e^{-p^2/(2m_e kT)}]dp \tag{177}$$

where p = electron momentum, μ = chemical potential, m_e = electron mass, k = Boltzmann constant, T = absolute temperature, A = constant independent of frequency and temperature, and α = constant given by[27]

$$\alpha = 4\pi m_e k^2/h^3 \tag{178}$$

where h = Planck's constant. The lower limit on the integral comes from the Einstein law of the photoelectric effect[27]

$$p_M^2/(2m_e) = h(\nu - \nu_o) = h\nu - W \tag{179}$$

where p_M = maximum momentum of the emitted photoelectrons, ν = frequency of the incident light, ν_o = threshold frequency, and W = work function.

Changing the variable of integration in the following manner

$$\eta = p^2/(2m_e kT) \qquad d\eta = pdp/(m_e kT) \tag{180}$$

allows the integral in equation (177) to be written as[27]

$$I = A\alpha T^2 D(\delta) \qquad I/T^2 = A\alpha D(\delta) \tag{181}$$

where

$$D(\delta) = \int_0^{\infty} \ell n(1 + e^{-\eta+\delta})d\eta \tag{182}$$

where

$$\delta = h(\nu - \nu_o)/(kT) \tag{183}$$

and δ can be positive or negative. For $\delta \leqslant 0$ the function $D(\delta)$ in equation (182) is obtained by expanding the logarithm in a power series with the result[27]

$$D = e^{\delta} - e^{2\delta}/2^2 + e^{3\delta}/3^2 - e^{4\delta}/4^2 + \cdots \tag{184}$$

For $\delta \geqslant 0$ the integral in equation (182) is rewritten as

$$\begin{aligned} D &= \int_0^{\delta} \ell n(1 + e^{-\eta+\delta})d\eta + \int_{\delta}^{\infty} \ell n(1 + e^{-\eta+\delta})d\eta \qquad (185)\\ &= \int_0^{\delta} \ell n[e^{-\eta+\delta}(1 + e^{\eta-\delta})]d\eta + \int_{\delta}^{\infty} \ell n(1 + e^{-\eta+\delta})d\eta \\ &= \int_0^{\delta} [-\eta + \delta + \ell n(1 + e^{\eta-\delta})]d\eta + \int_{\delta}^{\infty} \ell n(1 + e^{-\eta+\delta})d\eta \\ &= \delta^2/2 + \int_0^{\delta} \ell n(1 + e^{\eta-\delta})d\eta + \int_{\delta}^{\infty} \ell n(1 + e^{-\eta+\delta})d\eta \end{aligned}$$

In this form the logarithms can be expanded in power series because the arguments of the exponentials are always negative numbers, and it is simple to show that[27]

$$D = \pi^2/6 + \delta^2/2 - (e^{-\delta} - e^{-2\delta}/2^2 + e^{-3\delta}/3^2 - e^{-4\delta}/4^2 + \cdots) \tag{186}$$

The leading temperature term for the photoemission current is T^2 as shown in equation (181), and there is good agreement between equation (181) and experimental results on many kinds of ordinary metals and various temperatures.[27]

This section generalizes the Fowler theory of photoemission to the case of high-T_c superconductors which are treated as materials having complete spacetime coherence for the superconducting state ($T < T_c$), and partial spacetime coherence for the normal state ($T > T_c$). This is done by first observing that because the electron pairing force is weak (Section 2) the assumption of free electrons can be made and the complete or partial spacetime coherence can be introduced into the photoemission theory by integrating over a complex number single particle momentum in equation (177) as follows

$$\begin{aligned} \bar{I} &= A\alpha T/(m_e k)\int \bar{p}\ \ell n[1 + \bar{\mu} e^{-\bar{p}^2/(2m_e kT)}]d\bar{p} \qquad (187)\\ &= A\alpha T^2 \bar{D}(\bar{\delta}) \end{aligned}$$

where

$$\bar{D} = \int \ell n(1 + e^{-\bar{\eta}+\bar{\delta}})d\bar{\eta} \tag{188}$$

$$\bar{\delta} = \delta e^{j\theta_\delta} = h(\bar{\nu} - \bar{\nu}_o)/(kT) = \text{constant} \tag{189}$$

$$\bar{\nu} = \nu e^{j\theta_\nu} \qquad \bar{\nu}_o = \nu e^{j\theta_\nu^o} \tag{190}$$

where the internal phase angles of the frequency are assumed to be[28]

$$\theta_\nu = \theta_\nu^o = -\theta_{tR} = -\theta_{tR}^o = \text{constant} \tag{191}$$

where θ_{tR} and θ_{tR}^o refer to the incident electromagnetic waves that eject electrons from the surface of a high-T_c superconductor, so that equation (189) can be written as

$$\delta = h(\nu - \nu_o)/(kT) \qquad \theta_\delta = \theta_\nu = \theta_\nu^o = -\theta_{tR} = -\theta_{tR}^o \tag{192}$$

where δ and θ_δ are constants. The constant internal phase angle $\theta_\delta = -\theta_{tR}$ is associated with the incident electromagnetic wave interacting with matter. The integration variable $\bar{\eta}$ can be written as

$$\bar{\eta} = \eta e^{j\theta_\eta} = \bar{p}^2/(2m_e kT) \tag{193}$$

$$\eta = p^2/(2m_e kT) \qquad \theta_\eta = 2\theta_p = 2(\theta_r - \theta_t) \tag{194}$$

The superconducting and normal states of high-T_c superconductors will now be considered. For the superconducting state $\theta_t = \pi/6$, $\theta_r = \pi/3$, and $\theta_\eta = \pi/3$ for the electrons.[23] The complex number Einstein law of photoemission can be written as[25]

$$\bar{p}_M^2/(2m_e) = h(\bar{\nu} - \bar{\nu}_o) = h\bar{\nu} - \bar{W} \tag{194A}$$

$$p_M^2/(2m_e) = h(\nu - \nu_o) \qquad 2\theta_{pM} = \theta_\nu = \theta_\nu^o = \theta_\delta = \text{constant} \tag{194B}$$

where $\bar{p}_M$ = complex number maximum momentum of ejected electron.

A. Coherent Spacetime Superconducting State.

According to the theory presented in this paper the superconducting state of a high-T_c compound has complete spacetime coherence and therefore it follows from equations (193) and (194) that the variation of $\bar{\eta}$ is given by a pure rotation in internal space as follows

$$d\bar{\eta} = j\bar{\eta}d\theta_\eta = j\bar{p}^2 d\theta_p/(m_e kT) \tag{195}$$

with the magnitudes η and p taken as constants $\eta = \eta_c$ and $p = p_c$ and where

$$\eta = \eta_c = p_c^2/(2m_e kT) = \text{constant} \tag{196}$$

where the constant momentum magnitude is associated with the constant transition temperature T_c by

$$p_c^2/(2m_e) = kT_c \tag{197}$$

so that from equations (196) and (197)

$$\eta = \eta_c = T_c/T \tag{198}$$

Note also that $d\theta_\eta = 2d\theta_p$. The integral in equation (188) can then be written as

$$\bar{D} = \int_0^{\pi/3} \ell n[1 + \exp(\bar{\delta} - \eta_c e^{j\theta_\eta})] j\eta_c \, e^{j\theta_\eta} \, d\theta_\eta \tag{199}$$

where the upper limit comes from equation (194) and the fact that for the superconducting state of a high-T_c material

$$\theta_\eta = 2\theta_p = 2\theta_v = 2(\theta_r - \theta_t) = 2(\pi/3 - \pi/6) = \pi/3 \tag{200}$$

For an electron in a coherent spacetime state of the superconducting state of a high-T_c compound the kinematic internal phase angles are given by[23]

$$\theta_r = \pi/3 \qquad \theta_t = \pi/6 \qquad \theta_p = \pi/6 \qquad \theta_{pM} = \pi/6 \tag{200A}$$

while from equations (191), (192) and (194B) it follows that a photon ejecting an electron from the surface of a high-T_c material in its superconducting state has the following internal phase angles

$$\theta_\nu = \theta_\delta = \pi/3 \qquad \theta_{rR} = \theta^o_{rR} = -\pi/3 \qquad \theta_{tR} = \theta^o_{tR} = -\pi/3 \tag{200B}$$

Note that for photons it is always true that $\theta_{tR} = \theta_{rR}$. For the superconducting state of a high-T_c superconductor it follows from equation (194A) that the internal phase angle of the work function is given by

$$\theta_W = \pi/3 \tag{200C}$$

The spacetime coordinate phase angles for electrons in the superconducting state of a high-T_c compound are $\theta_t = \pi/6$ and $\theta_r = \pi/3$, while for blackbody radiation photons in the superconducting state of a high-T_c compound the phase angles are $\theta_\nu = \pi/6$ with $\theta_{tR} = \theta_{rR} = -\pi/6$, a result which follows from the momentum conservation law for photon-electron collisions $h\bar{\nu}/c = m\bar{v}$ which gives immediately $\theta_\nu = \pi/3 - \pi/6 = \pi/6$.[28] On the other hand, the internal phase angles of the photons that are ejecting electrons from the surface of a high-T_c material in its superconducting state are given by equation (200B) as $\theta_\nu = \pi/3$ with $\theta_{tR} = \theta_{rR} = -\pi/3$ and are seen to have twice the values associated with blackbody radiation photons.

Expanding the logarithm in equation (199) gives the following result

$$\bar{D} = \eta_c \bar{D}' = \eta_c (D'_R + jD'_I) = \eta_c \sum_{\sigma=1}^{\infty} \bar{J}'_\sigma \tag{201}$$

$$= j\eta_c \int_0^{\pi/3} \{ \sum_{\sigma=1}^{\infty} (-1)^{\sigma-1}/\sigma \exp[\sigma(\bar{\delta} - \eta_c e^{j\theta_\eta})] \} e^{j\theta_\eta} \, d\theta_\eta$$

where

$$\bar{J}'_\sigma = (-1)^{\sigma-1}/\sigma \; e^{(\sigma\bar{\delta} + j\pi/2)} \int_0^{\pi/3} \exp(j\theta_\eta - \sigma\eta_c e^{j\theta_\eta}) d\theta_\eta \tag{202}$$

Simple algebra shows that

$$D'_R = \sum_{\sigma=1}^{\infty} J'_{\sigma R} \tag{203}$$

$$D'_I = \sum_{\sigma=1}^{\infty} J'_{\sigma I} \tag{204}$$

where

$$J'_{\sigma R} = \mu_\sigma^+[S_\sigma \cos(\sigma\delta \sin\theta_\delta) - C_\sigma \sin(\sigma\delta \sin\theta_\delta)] \tag{205}$$
$$= \mu_\sigma^+[S_\sigma \cos(\sigma\delta\sqrt{3}/2) - C_\sigma \sin(\sigma\delta\sqrt{3}/2)]$$

$$J'_{\sigma I} = \mu_\sigma^+[S_\sigma \sin(\sigma\delta \sin\theta_\delta) + C_\sigma \cos(\sigma\delta \sin\theta_\delta)] \tag{206}$$
$$= \mu_\sigma^+[S_\sigma \sin(\sigma\delta\sqrt{3}/2) + C_\sigma \cos(\sigma\delta\sqrt{3}/2)]$$

where

$$S_\sigma = \int_0^{\pi/3} e^{-\sigma\eta_c \cos\theta_\eta} \sin(\sigma\eta_c \sin\theta_\eta - \theta_\eta) d\theta_\eta \tag{207}$$

$$C_\sigma = \int_0^{\pi/3} e^{-\sigma\eta_c \cos\theta_\eta} \cos(\sigma\eta_c \sin\theta_\eta - \theta_\eta) d\theta_\eta \tag{208}$$

$$\mu_\sigma^+ = (-1)^{\sigma-1}/\sigma\ e^{\sigma\delta \cos\theta_\delta} = (-1)^{\sigma-1}/\sigma\ e^{\sigma\delta/2} \tag{209}$$

where the value $\theta_\delta = \pi/3$ was used for a photon undergoing an inelastic scattering with a coherent spacetime electron and ejecting it from the surface of a high-T_c compound in its superconducting state as described by the complex number forms of the Einstein equations (194A) and (194B). Finally, combining equations (187), (198) and (201) gives the following expression for the photoemission current from a high-T_c compound in its superconducting state

$$\bar{I} = A\alpha T_c T\bar{D}' \tag{210}$$

and the measured photoemission current is

$$I_m = A\alpha T_c T D'_R \tag{211}$$

where D'_R is given by equations (203) through (209). Equation (211) shows that for the superconducting state of a high-T_c material the leading term of the photoemission current is linear in T, and not quadratic in T as equation (181) shows to be the case for ordinary metals.

B. Normal State Photoemission.

The photoelectric current from the normal state ($T > T_c$) of a high-T_c superconductor material is calculated from the general complex number equations (187) through (194) by inserting the following general expression for $d\bar{\eta}$ that

is valid for partial spacetime coherence into equation (188)

$$d\bar{\eta} = e^{j(\theta_\eta + \beta_{\eta\eta})} \sec \beta_{\eta\eta}\, d\eta \tag{212}$$

where

$$\tan \beta_{\eta\eta} = \eta \partial \theta_\eta / \partial \eta \tag{213}$$

and where $\bar{\eta}$ is defined by equation (193). The evaluation of the integral in equation (188) requires the examination of two possible cases, Re $\bar{\delta} < 0$ and Re $\bar{\delta} > 0$.

Case 1. Re $\bar{\delta} < 0$ or $\nu < \nu_o$.

For this case the logarithm in equation (188) can be simply expanded in a power series with the result

$$\bar{D} = \sum_{\sigma=1}^{\infty} \bar{L}_\sigma \tag{214}$$

where

$$\bar{L}_\sigma = (-1)^{\sigma-1}/\sigma\, e^{\sigma\bar{\delta}} \int_o^\infty \exp[-\sigma\eta e^{j\theta_\eta} + j(\theta_\eta + \beta_{\eta\eta})] \sec \beta_{\eta\eta}\, d\eta \tag{215}$$

The real and imaginary parts of $\bar{D}$ are written as

$$D_R = \sum_{\sigma=1}^{\infty} L_{\sigma R} \qquad D_I = \sum_{\sigma=1}^{\infty} L_{\sigma I} \tag{216}$$

where

$$L_{\sigma R} = (-1)^{\sigma-1}/\sigma\, e^{\sigma\delta \cos \theta_\delta} \int_o^\infty H \cos(\sigma\delta \sin \theta_\delta - \sigma\eta \sin \theta_\eta + \theta_\eta + \beta_{\eta\eta}) d\eta \tag{217}$$

$$L_{\sigma I} = (-1)^{\sigma-1}/\sigma\, e^{\sigma\delta \cos \theta_\delta} \int_o^\infty H \sin(\sigma\delta \sin \theta_\delta - \sigma\eta \sin \theta_\eta + \theta_\eta + \beta_{\eta\eta}) d\eta \tag{218}$$

where

$$H = \sec \beta_{\eta\eta}\, e^{-\sigma\eta \cos \theta_\eta} \tag{218A}$$

This is the general case, and some simplifications can be made.

If $\theta_\eta = \theta_\eta^c = 2(\theta_r^c - \theta_t^c)$ = constant, then $\beta_{\eta\eta} = 0$ and the integrals simplify to the following forms

$$L_{\sigma R}^c = (-1)^{\sigma-1}/\sigma\, e^{\sigma\delta \cos \theta_\delta} \int_o^\infty e^{-\sigma\eta \cos \theta_\eta^c} \cos(\alpha_\sigma - \sigma\eta \sin \theta_\eta^c) d\eta \tag{219}$$

$$L_{\sigma I}^c = (-1)^{\sigma-1}/\sigma\, e^{\sigma\delta \cos \theta_\delta} \int_o^\infty e^{-\sigma\eta \cos \theta_\eta^c} \sin(\alpha_\sigma - \sigma\eta \sin \theta_\eta^c) d\eta \tag{220}$$

where α_σ = constant given by

$$\alpha_\sigma = \sigma\delta \sin \theta_\delta + \theta_\eta^c \tag{221}$$

The integrals in equations (219) and (220) can be written as

$$L_{\sigma R}^c = \mu_\sigma^+ (M_{c\sigma} \cos \alpha_\sigma + M_{s\sigma} \sin \alpha_\sigma) \tag{222}$$

$$L_{\sigma I}^c = \mu_\sigma^+ (M_{c\sigma} \sin \alpha_\sigma - M_{s\sigma} \cos \alpha_\sigma) \tag{223}$$

where

$$\mu_\sigma^+ = (-1)^{\sigma-1}/\sigma \; e^{\sigma\delta \cos \theta_\delta} = (-1)^{\sigma-1}/\sigma \; e^{\sigma\delta \cos \theta_{tR}} \tag{224}$$

where θ_δ is generally related to the internal phase angle of time for electromagnetism by equation (192), and where

$$M_{c\sigma}(\theta_\eta^c) = \int_0^\infty e^{-\sigma\eta \cos \theta_\eta^c} \cos(\sigma\eta \sin \theta_\eta^c) d\eta \tag{225}$$

$$M_{s\sigma}(\theta_\eta^c) = \int_0^\infty e^{-\sigma\eta \cos \theta_\eta^c} \sin(\sigma\eta \sin \theta_\eta^c) d\eta \tag{226}$$

The integrals in equations (225) and (226) can be found in tables of integrals. By defining the quantities

$$a = \cos \theta_\eta^c = \cos[2(\theta_r^c - \theta_t^c)] \tag{227}$$

$$b = \sin \theta_\eta^c = \sin[2(\theta_r^c - \theta_t^c)] \tag{228}$$

the integrals in equations (225) and (226) are[29,30]

$$M_{c\sigma} = a/\sigma \tag{229}$$

$$M_{s\sigma} = b/\sigma \tag{230}$$

The values of D_R and D_I for $\theta_\eta = \theta_\eta^c$ = constant are obtained from equations (216), (222) and (223) to be

$$D_R^c = \sum_{\sigma=1}^\infty L_{\sigma R}^c \qquad D_I^c = \sum_{\sigma=1}^\infty L_{\sigma I}^c \tag{230A}$$

For the normal state of a high-T_c superconductor $\theta_\eta \neq 0$ for the electron momentum and $\theta_\delta \neq 0$ for the electromagnetic waves interacting with the electrons, and in particular

$$\theta_\eta = 2\theta_p = 2(\theta_r + \beta_{rr} - \theta_t - \beta_{tt}) \tag{231}$$

$$\theta_\delta = -\theta_{tR} = -\theta_{tR}^o \tag{232}$$

For Re $\bar{\delta} < 0$ the measured photoemission current for the normal state is obtained from equations (187), (216) and (217) to be

$$I_m = A\alpha T^2 D_R \qquad T > T_c \tag{233}$$

For the more simple assumption that $\theta_\eta = \theta_\eta^c$ the measured photoemission current is given by

$$I_m^c = A\alpha T^2 D_R^c \qquad T > T_c \tag{234}$$

where D_R^c is given by equation (230A). Thus as with ordinary metals, the normal state of a high-T_c superconductor has a T^2 dependence in the leading term of the photoemission current.

When $\theta_r^c = 0$ and $\theta_t^c = 0$ it follows that $\theta_\eta^c = 0$ and equations (229) and (230) become with $a = 1$ and $b = 0$

$$M_{c\sigma} = 1/\sigma \qquad M_{s\sigma} = 0 \tag{235}$$

and with $\theta_\delta = -\theta_{tR} = -\theta_{tR}^o = 0$ equation (216) becomes

$$D_R^o = \sum_{\sigma=1}^{\infty} (-1)^{\sigma-1}/\sigma^2 \, e^{\sigma\delta} \qquad D_I^o = 0 \tag{235A}$$

which is just the Fowler result given in equation (184) for $\nu < \nu_o$. Equation (235A) can also be obtained directly from equation (215) by taking $\theta_\eta = 0$.

Case 2. Re $\bar{\delta} > 0$ or $\nu > \nu_o$.

For this case the integral in equation (188) must be written as a generalization of the scalar form given in equation (185)

$$\bar{D} = D_R + jD_I = \bar{\delta}^2/2 + \bar{G} + \bar{H} \tag{236}$$

where

$$\bar{G} = \int_0^{\bar{\delta}} \ell n(1 + e^{\bar{\eta}-\bar{\delta}}) d\bar{\eta} \tag{237}$$

$$\bar{H} = \int_{\bar{\delta}}^{\infty} \ell n(1 + e^{-\bar{\eta}+\bar{\delta}}) d\bar{\eta} \tag{238}$$

By expanding the logarithms in power series it is easy to show that

$$\bar{G} = \sum_{\sigma=1}^{\infty} \bar{P}_\sigma \qquad \bar{H} = \sum_{\sigma=1}^{\infty} \bar{T}_\sigma \tag{239}$$

where

$$\bar{P}_\sigma = (-1)^{\sigma-1}/\sigma \, e^{-\sigma\bar{\delta}} \int_0^{\delta} \exp[\sigma\eta e^{j\theta_\eta} + j(\theta_\eta + \beta_{\eta\eta})] \sec \beta_{\eta\eta} \, d\eta \tag{240}$$

$$\bar{T}_\sigma = (-1)^{\sigma-1}/\sigma \, e^{\sigma\bar{\delta}} \int_0^{\delta} \exp[-\sigma\eta e^{j\theta_\eta} + j(\theta_\eta + \beta_{\eta\eta})] \sec \beta_{\eta\eta} \, d\eta \tag{241}$$

The real and imaginary parts of equation (236) are written as

$$D_R = \delta^2/2 \cos(2\theta_\delta) + G_R + H_R \tag{242}$$

$$D_I = \delta^2/2 \sin(2\theta_\delta) + G_I + H_I \tag{243}$$

where

$$G_R = \sum_{\sigma=1}^{\infty} P_{\sigma R} \qquad G_I = \sum_{\sigma=1}^{\infty} P_{\sigma I} \tag{244}$$

$$H_R = \sum_{\sigma=1}^{\infty} T_{\sigma R} \qquad H_I = \sum_{\sigma=1}^{\infty} T_{\sigma I} \tag{245}$$

where

$$P_{\sigma R} = (-1)^{\sigma-1}/\sigma \; e^{-\sigma\delta \cos \theta_\delta} \int_0^{\delta} \mathfrak{C} \cos(-\sigma\delta \sin \theta_\delta + \sigma\eta \sin \theta_\eta + \theta_\eta + \beta_{\eta\eta}) d\eta \tag{246}$$

$$P_{\sigma I} = (-1)^{\sigma-1}/\sigma \; e^{-\sigma\delta \cos \theta_\delta} \int_0^{\delta} \mathfrak{C} \sin(-\sigma\delta \sin \theta_\delta + \sigma\eta \sin \theta_\eta + \theta_\eta + \beta_{\eta\eta}) d\eta \tag{247}$$

$$T_{\sigma R} = (-1)^{\sigma-1}/\sigma \; e^{\sigma\delta \cos \theta_\delta} \int_{\delta}^{\infty} \mathfrak{H} \cos(\sigma\delta \sin \theta_\delta - \sigma\eta \sin \theta_\eta + \theta_\eta + \beta_{\eta\eta}) d\eta \tag{248}$$

$$T_{\sigma I} = (-1)^{\sigma-1}/\sigma \; e^{\sigma\delta \cos \theta_\delta} \int_{\delta}^{\infty} \mathfrak{H} \sin(\sigma\delta \sin \theta_\delta - \sigma\eta \sin \theta_\eta + \theta_\eta + \beta_{\eta\eta}) d\eta \tag{249}$$

where $\mathfrak{H}$ is given by equation (218A) and $\mathfrak{C}$ is given by

$$\mathfrak{C} = \sec \beta_{\eta\eta} \; e^{\sigma\eta \cos \theta_\eta} \tag{249A}$$

Therefore combining equations (242) through (249) gives

$$D_R = \delta^2/2 \cos(2\theta_\delta) + \sum_{\sigma=1}^{\infty} (P_{\sigma R} + T_{\sigma R}) \tag{250}$$

$$D_I = \delta^2/2 \sin(2\theta_\delta) + \sum_{\sigma=1}^{\infty} (P_{\sigma I} + T_{\sigma I}) \tag{251}$$

where $\theta_\delta = -\theta_{tR} = -\theta_{tR}^o$ = constant is associated with the electromagnetic waves interacting with the normal state of the high-T_c material and producing photoelectrons.

For the special case when $\theta_\eta = \theta_\eta^c$ = constant, the integrals in equations (246) through (249) simplify as follows

$$P_{\sigma R}^c = (-1)^{\sigma-1}/\sigma \; e^{-\sigma\delta \cos \theta_\delta} \int_0^{\delta} e^{\sigma\eta \cos \theta_\eta^c} \cos(\gamma_\sigma + \sigma\eta \sin \theta_\eta^c) d\eta \tag{252}$$

$$P_{\sigma I}^c = (-1)^{\sigma-1}/\sigma \; e^{-\sigma\delta \cos \theta_\delta} \int_0^{\delta} e^{\sigma\eta \cos \theta_\eta^c} \sin(\gamma_\sigma + \sigma\eta \sin \theta_\eta^c) d\eta \tag{253}$$

$$T^c_{\sigma R} = (-1)^{\sigma-1}/\sigma\ e^{\sigma\delta\ \cos\ \theta_\delta} \int_\delta^\infty e^{-\sigma\eta\ \cos\ \theta^c_\eta} \cos(\alpha_\sigma - \sigma\eta\ \sin\ \theta^c_\eta)d\eta \tag{254}$$

$$T^c_{\sigma I} = (-1)^{\sigma-1}/\sigma\ e^{\sigma\delta\ \cos\ \theta_\delta} \int_\delta^\infty e^{-\sigma\eta\ \cos\ \theta^c_\eta} \sin(\alpha_\sigma - \sigma\eta\ \sin\ \theta^c_\eta)d\eta \tag{255}$$

where α_σ is given by equation (221) and where

$$\gamma_\sigma = -\sigma\delta\ \sin\ \theta_\delta + \theta^c_\eta \tag{256}$$

The integrals in equations (252) through (255) can be rewritten as

$$P^c_{\sigma R} = \mu^-_\sigma(N_{c\sigma}\ \cos\ \gamma_\sigma - N_{s\sigma}\ \sin\ \gamma_\sigma) \tag{257}$$

$$P^c_{\sigma I} = \mu^-_\sigma(N_{c\sigma}\ \sin\ \gamma_\sigma + N_{s\sigma}\ \cos\ \gamma_\sigma) \tag{258}$$

$$T^c_{\sigma R} = \mu^+_\sigma(Q_{c\sigma}\ \cos\ \alpha_\sigma + Q_{s\sigma}\ \sin\ \alpha_\sigma) \tag{259}$$

$$T^c_{\sigma I} = \mu^+_\sigma(Q_{c\sigma}\ \sin\ \alpha_\sigma - Q_{s\sigma}\ \cos\ \alpha_\sigma) \tag{260}$$

where μ^+_σ is given in equation (224) and where

$$\mu^-_\sigma = (-1)^{\sigma-1}/\sigma\ e^{-\sigma\delta\ \cos\ \theta_\delta} = (-1)^{\sigma-1}/\sigma\ e^{-\sigma\delta\ \cos\ \theta_{tR}} \tag{261}$$

and where

$$N_{c\sigma}(\delta,\theta^c_\eta) = \int_0^\delta e^{\sigma\eta\ \cos\ \theta^c_\eta} \cos(\sigma\eta\ \sin\ \theta^c_\eta)d\eta \tag{262}$$

$$N_{s\sigma}(\delta,\theta^c_\eta) = \int_0^\delta e^{\sigma\eta\ \cos\ \theta^c_\eta} \sin(\sigma\eta\ \sin\ \theta^c_\eta)d\eta \tag{263}$$

$$Q_{c\sigma}(\delta,\theta^c_\eta) = \int_\delta^\infty e^{-\sigma\eta\ \cos\ \theta^c_\eta} \cos(\sigma\eta\ \sin\ \theta^c_\eta)d\eta \tag{264}$$

$$Q_{s\sigma}(\delta,\theta^c_\eta) = \int_\delta^\infty e^{-\sigma\eta\ \cos\ \theta^c_\eta} \sin(\sigma\eta\ \sin\ \theta^c_\eta)d\eta \tag{265}$$

The integrals in equations (262) through (265) are found in tables of integrals and have the values[29,30]

$$N_{c\sigma}(\delta,\theta^c_\eta) = 1/\sigma\ e^{\sigma a\delta}[a\ \cos(\sigma b\delta) + b\ \sin(\sigma b\delta)] - a/\sigma \tag{266}$$

$$N_{s\sigma}(\delta,\theta^c_\eta) = 1/\sigma\ e^{\sigma a\delta}[a\ \sin(\sigma b\delta) - b\ \cos(\sigma b\delta)] + b/\sigma \tag{267}$$

$$Q_{c\sigma}(\delta,\theta^c_\eta) = 1/\sigma\ e^{-\sigma a\delta}[a\ \cos(\sigma b\delta) - b\ \sin(\sigma b\delta)] \tag{268}$$

$$Q_{s\sigma}(\delta,\theta^c_\eta) = 1/\sigma\ e^{-\sigma a\delta}[a\ \sin(\sigma b\delta) + b\ \cos(\sigma b\delta)] \tag{269}$$

where a and b are defined in equations (227) and (228) respectively. The values of D_R and D_I for $\theta_\eta = \theta_\eta^c$ = constant are obtained from equations (250) and (251) to be

$$D_R^c = \delta^2/2 \cos(2\theta_\delta) + \sum_{\sigma=1}^{\infty} (P_{\sigma R}^c + T_{\sigma R}^c) \tag{270}$$

$$D_I^c = \delta^2/2 \sin(2\theta_\delta) + \sum_{\sigma=1}^{\infty} (P_{\sigma I}^c + T_{\sigma I}^c) \tag{271}$$

where equations (257) through (269) can be used to calculate these quantities.

Finally, the photoemission current for the normal state of a high-T_c superconsuctor with Re $\bar{\delta} > 0$ is given by

$$I_m = A\alpha T^2 D_R \qquad T > T_c \tag{272}$$

where D_R is given by equation (250) for the case $\theta_\eta = \theta_\eta(\eta)$, and by

$$I_m^c = A\alpha T^2 D_R^c \qquad T > T_c \tag{273}$$

for the simplifying case where $\theta_\eta = \theta_\eta^c$ = constant and where D_R^c is given by equation (270). Therefore from equations (272) and (273) it is clear that the photoemission current from the normal state of a high-T_c compound has a leading term which is quadratic in the temperature of the material.

Equations (252) through (271), with the approximation that $\theta_r = \theta_r^c$ and $\theta_t = \theta_t^c$ are constants, are valid for the normal state of a high-T_c superconductor. It is possible to reduce these equations to the case of ordinary metallic behaviour by making the following substitutions

$$\theta_r = 0 \qquad \theta_t = 0 \qquad \theta_\eta^c = 0 \qquad \theta_\delta = 0 \tag{274}$$

$$a = 1 \qquad b = 0 \qquad \gamma_\sigma = 0 \qquad \alpha_\sigma = 0 \tag{275}$$

$$\mu_\sigma^+ = (-1)^{\sigma-1}/\sigma \; e^{\sigma\delta} \qquad \mu_\sigma^- = (-1)^{\sigma-1}/\sigma \; e^{-\sigma\delta} \tag{276}$$

Then the integrals in equations (266) through (269) become

$$N_{c\sigma} = (e^{\sigma\delta} - 1)/\sigma \qquad N_{s\sigma} = 0 \tag{277}$$

$$Q_{c\sigma} = 1/\sigma \; e^{-\sigma\delta} \qquad Q_{s\sigma} = 0 \tag{278}$$

and therefore from equations (257) through (260)

$$P_{\sigma R}^c = (-1)^{\sigma-1}/\sigma^2 \; (1 - e^{-\sigma\delta}) \qquad P_{\sigma I}^c = 0 \tag{279}$$

$$T_{\sigma R}^c = (-1)^{\sigma-1}/\sigma^2 \qquad T_{\sigma I}^c = 0 \tag{280}$$

and from equation (270)

$$D_R^c = \delta^2/2 + \sum_{\sigma=1}^{\infty} (-1)^{\sigma-1}/\sigma^2 \ (2 - e^{-\sigma\delta}) \qquad (281)$$

$$= \delta^2/2 + \pi^2/6 - \sum_{\sigma=1}^{\infty} (-1)^{\sigma-1}/\sigma^2 \ e^{-\sigma\delta}$$

where the following identity was used[31]

$$\pi^2/12 = \sum_{\sigma=1}^{\infty} (-1)^{\sigma-1}/\sigma^2 \qquad (282)$$

Equation (281) is just the Fowler result for ordinary metals that was presented in equation (186)

The values of θ_t and θ_r (or θ_t^c and θ_r^c) for the electrons in the normal state ($T > T_c$) of a high-T_c superconductor material can be obtained by comparing the measured values of the photoemission current with the predicted values given by equations (233) or (234) for the case $\nu < \nu_o$ and by equations (272) or (273) for the case $\nu > \nu_o$. The value of $\theta_\nu = -\theta_{tR}$ for the electromagnetic waves interacting with the normal state of a high-T_c compound may possibly be obtained from photoemission experiments. Note that according to equations (233), (234), (272) and (273) the leading term of the photoelectric current is proportional to T^2 for the normal state of a high-T_c compound. The leading term of the photoelectric current from the superconducting state of high-T_c matter is linear in T according to equation (211). The verification of this linear dependence on T may possibly allow the determination of $\theta_t = \pi/6$ and $\theta_r = \pi/3$ for electrons, and $\theta_\nu = -\theta_{tR} = \pi/3$ for photons ejecting electrons from the surface of a high-T_c material in its superconducting state.

5. CONCLUSION. The superconducting state of a high-T_c compound is described as being a completely coherent spacetime state of electrons in Cooper pairs. Like ordinary superconductivity, high-T_c superconductivity is a macroscopic quantum phenomenon.[32] By considering the coherent spacetime acceleration of electrons in Cooper pairs it is shown that the normalized superconductivity energy gap for a high-T_c material is given by $(6/\pi)(3.52)/(1 - 4/\pi\ \theta_a)$ where θ_a = relative phase angle of the electron acceleration. The often measured large values of the normalized superconductivity energy gap is due to the factor $(6/\pi)$ and is not necessarily associated with a strong coupling of the electrons. The $6/\pi$ factor arises from the coherent spacetime state for electrons which has $\theta_t = \pi/6$ for the value of the internal phase angle for time. The normal state ($T > T_c$) of a high-T_c material is described as a partially coherent spacetime state. Incoherent, partially coherent and coherent thermodynamic functions are introduced to describe slow, moderately fast, and ultrafast thermodynamic processes respectively. These processes can occur in the various spacetime states of ordinary or high-T_c matter. Because the electron-electron interaction is weak, a noninteracting electron gas in a coherent or partially coherent spacetime state is used to describe the photoemission of electrons from the surface of the superconducting and normal states respectively of a high-T_c superconductor. The leading term of the photoelectric current is linear in temperature for the superconducting state and quadratic in temperature for the normal state of a high-T_c material.

ACKNOWLEDGEMENT

The author would like to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. Ginsberg, D. M., editor, Physical Properties of High Temperature Superconductors I, World Scientific, Singapore, 1989.

2. Wolf, S. A. and Kresin, V. Z., editors, Novel Superconductivity, Plenum, New York, 1987.

3. Bednorz, J. G. and Müller, K. A., editors, Earlier and Recent Aspects of Superconductivity, Springer-Verlag, New York, 1990.

4. Lynn, J. W., editor, High Temperature Superconductivity, Springer-Verlag, New York, 1990.

5. Ishiguro, T. and Yamaji, K., Organic Superconductors, Springer-Verlag, New York, 1990.

6. Bednorz, J. G. and Müller, K. A., "Perovskite-Type Oxides-The New Approach to High-T_c Superconductivity," Revs. Mod. Physics, Vol. 60, No. 3, July 1988.

7. Pickett, W. E., "Electronic Structure of the High-Temperature Oxide Superconductors," Revs. Mod. Physics, Vol. 61, No. 2, April 1989.

8. Micnas, R., Ranninger, J., and Robaszkiewicz, S., "Superconductivity in Narrow-Band Systems with Local Nonretarded Attractive Interactions," Revs. Mod. Physics, Vol. 62, No. 1, January 1990.

9. Geballe, T. H. and Hulm, J. K., "Superconductivity-The State That Came in from the Cold," Science, Vol. 239, pg. 367, 22 January 1988.

10. Phillips, J. C., Physics of High-T_c Superconductors, Academic, San Diego, CA, 1989.

11. Bedel, K., Pines, D. and Schrieffer, J. R., editors, High Temperature Superconductivity, Addison-Wesley, Reading, MA, 1990.

12. Cava, R. J., "Superconductors Beyond 1-2-3," Scientific American, p. 42, Aug. 1990.

13. Anderson, P. W., Science, Vol. 235, p. 1196, 1987.

14. Schlesinger, Z., Collins, R. T., Holtzberg, F., Feild, C., Blanton, S. H., Welp, U., Crabtree, G. W., Fang, Y. and Liu, J. Z., "Superconducting Energy Gap and Normal-State Conductivity of a Single-Domain $YBa_2Cu_3O_7$ Crystal," Phys. Rev. Lett., Vol. 65, 6 Aug. 1990.

15. Ioffe, L. B. and Wiegmann, P. B., "Linear Temperature Dependence of Resistivity as Evidence of Gauge Interaction," Phys. Rev. Lett., Vol. 65, 30 July 1990.

16. Tsuei, C. C., Newns, D. M., Chi, C. C. and Pattnaik, P. C., "Anomalous Isotope Effect and van Hove Singularity in Superconducting Cu Oxides," Phys. Rev. Lett., Vol. 65, p. 2724, 19 Nov. 1990.

17. Emery, V. J., "Strong-Coupling Field Theory and Soliton Doping in a One-Dimensional Copper-Oxide Model," Phys. Rev. Lett., Vol. 65, p. 1076, 20 Aug. 1990.

18. Trugman, S. A., "Explanation of Normal-State Properties of High-Temperature Superconductors," Phys. Rev. Lett., Vol. 65, p. 500, 23 July 1990.

19. Margaritondo, G., Huber, D. L. and Olson, C. G., "Photoemission Spectroscopy of the High-Temperature Superconductivity Gap," Science, Vol. 246, p. 770, 10 Nov. 1989.

20. Friedl, B., Thomsen, C. and Cardona, M., "Determination of the Superconducting Gap in $RBa_2Cu_3O_{7-\delta}$," Phys. Rev. Lett., Vol. 65, p. 915, 13 Aug. 1990.

21. Seidel, H., Hentsch, F., Mehring, M., Bednorz, J. G. and Müller, K. A., Europhys. Lett., Vol. 5, p. 647, 1988.

22. Demuth, J. E., Persson, B. N. J., Holtzberg, F. and Chandrasekhar, C. V., "Surface and Superconducting Properties of Cleaved High-Temperature Superconductors," Phys. Rev. Lett, Vol. 64, p. 603, 29 Jan. 1990.

23. Weiss, R. A., "Electromagnetism and Gravity," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990, p. 265.

24. Weiss, R. A., "Gauge Theory of Time," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990, p. 367.

25. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.

26. Weiss, R. A., Relativistic Thermodynamics, Vols. 1 and 2, Exposition Press, New York, 1976.

27. Fowler, R. H., Statistical Mechanics, Cambridge Univ. Press, New York, p. 358, 1955.

28. Weiss, R. A., "Thermal Radiation of High-T_c Superconductors," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990, p. 399.

29. Petit Bois, G., Tables of Indefinite Integrals, Dover, New York, 1961.

30. Gradshteyn, I. S. and Ryzhik, I. M., Table of Integrals, Series, and Products, Academic, New York, 1980.

31. Bromwich, T., An Introduction to the Theory of Infinite Series, MacMillan, New York, 1955.

32. Bardeen, J., "Superconductivity and Other Macroscopic Quantum Phenomena," Physics Today, p. 25, Dec. 1990.

QUANTUM THEORY OF TIME AND THERMODYNAMICS

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. Quantum thermodynamics is introduced through a quantized relativistic trace equation. This equation describes the discrete and continuous spectra and eigenfunctions of macroscopic thermodynamic systems. For solids and quantum liquids this is equivalent to a set of coupled eigenvalue equations for the internal energy and Grüneisen parameter. Simultaneous eigenvalue equations are developed for internal energy, time, time dimension and space dimension. These equations determine the effects of real state equations on the rates and geometrical structures of physical processes such as chemical and nuclear reactions which occur in bulk matter. High-T_c superconductivity is suggested to be associated with the coherent spacetime state of electrons in Cooper pairs. A quantized relativistic thermodynamic trace equation for coherent spacetime is developed, and this equation in conjunction with the quantized coherent time, time dimension and space dimension equations are suggested to describe high-T_c superconductivity. The first order macroscopic quantum eigenvalue equations for time, time dimension and space dimension are the bulk matter equivalents of the Dirac equation which describes microscopic systems. The eigenvalue equations for time, time dimension and space dimension are solved and yield solutions that predict structured energy and pressure. The solution for a particle confined to an energy-pressure box is obtained. The eigenvalue equations suggest that time and dimension can be interpreted to be wave functions in energy-pressure space. For the case where the thermodynamic gauge parameters are nearly constant, the first order time and dimension equations assume a Schrödinger form whose solutions can also be used to determine the quantum structures of energy and pressure. These results have applications to astrophysics and geophysics because the internal processes of stars and planets are affected by the quantized time and dimension structures of energy and pressure. The macroscopic quantum equations may also be useful for the description of quantum devices that utilize differences in energy states to confine electrons in quantum wells, wires and dots.

1. INTRODUCTION. Time enters the calculations of classical and quantum physics as an independent parameter in the scores of differential equations of physics such as Newton's law of motion, the equations of Lagrange, Hamilton, Schrödinger and Dirac, and the Einstein field equations of general relativity.[1-15] These laws and equations teach us nothing about the nature of time because time enters only in the denominators of the derivatives that appear in the expression of the laws of nature.

A better understanding of time can be obtained if time can be raised from the denominators of the derivatives that appear in the equations of physics, and placed in the numerators of laws that describe the dependence of time on such basic quantities as energy density and pressure or temperature and density.[16]

In this way differential equations for time can be determined, and time assumes a more fundamental role akin to the wave function of the Dirac and Schrödinger equations. In this sense time is easier to describe than space because space, through the concepts of energy density and pressure, still remains as as independent parameter in the denominator of the time equations. Time and the dimentions of time and space are local quantities that depend on energy and pressure.[16]

The complex number internal energy, energy density, pressure, time, time dimension, and space dimension are written as[16,17]

$$\bar{U} = Ue^{j\theta U} \qquad \bar{E} = Ee^{j\theta E} \qquad \bar{P} = Pe^{j\theta P} \tag{1}$$

$$\bar{t} = te^{j\theta t} \qquad \bar{D}_t = D_t e^{j\theta Dt} \qquad \bar{D}_s = D_s e^{j\theta Ds} \tag{2}$$

where $\bar{E} = \bar{U}/V$ holds for incoherent space. Also $\theta_E = \theta_U$. The corresponding unrenormalized time, dimension, and thermodynamic functions will be designated as in equations (1) and (2) but with a superscript "a" added to the magnitudes and internal phase angles of these quantities. From equations (1) and (2) the following differential expressions are valid assuming the partial coherence of the thermodynamic functions

$$d\bar{U} = e^{j\theta U}(dU + jUd\theta_U) \qquad d\bar{t} = e^{j\theta t}(dt + jtd\theta_t) \tag{3}$$

$$d\bar{E} = e^{j\theta E}(dE + jEd\theta_E) \qquad d\bar{D}_t = e^{j\theta Dt}(dD_t + jD_t d\theta_{Dt}) \tag{4}$$

$$d\bar{P} = e^{j\theta P}(dP + jPd\theta_P) \qquad d\bar{D}_s = e^{j\theta Ds}(dD_s + jD_s d\theta_{Ds}) \tag{5}$$

These differentials can be rewritten as

$$d\bar{U} = e^{j\Phi U} \sec \beta_{UU}\, dU = e^{j\Phi U} \csc \beta_{UU}\, Ud\theta_U \tag{6}$$

$$d\bar{E} = e^{j\Phi E} \sec \beta_{EE}\, dE = e^{j\Phi E} \csc \beta_{EE}\, Ed\theta_E \tag{7}$$

$$d\bar{P} = e^{j\Phi P} \sec \beta_{PP}\, dP = e^{j\Phi P} \csc \beta_{PP}\, Pd\theta_P \tag{8}$$

$$d\bar{t} = e^{j\Phi t} \sec \beta_{tt}\, dt = e^{j\Phi t} \csc \beta_{tt}\, td\theta_t \tag{9}$$

$$d\bar{D}_t = e^{j\Phi Dt} \sec \beta_{DtDt}\, dD_t = e^{j\Phi Dt} \csc \beta_{DtDt}\, D_t d\theta_{Dt} \tag{10}$$

$$d\bar{D}_s = e^{j\Phi Ds} \sec \beta_{DsDs}\, dD_s = e^{j\Phi Ds} \csc \beta_{DsDs}\, D_s d\theta_{Ds} \tag{11}$$

where

$$\tan \beta_{UU} = U\partial\theta_U/\partial U \qquad \tan \beta_{PP} = P\partial\theta_P/\partial P \qquad \tan \beta_{tt} = t\partial\theta_t/\partial t \tag{12}$$

$$\tan \beta_{DtDt} = D_t\partial\theta_{Dt}/\partial D_t \qquad \tan \beta_{DsDs} = D_s\partial\theta_{Ds}/\partial D_s \tag{13}$$

$$\tan \beta_{EE} = E\partial\theta_E/\partial E \tag{13A}$$

$$\Phi_U = \theta_U + \beta_{UU} \qquad \Phi_P = \theta_P + \beta_{PP} \qquad \Phi_t = \theta_t + \beta_{tt} \tag{14}$$

$$\Phi_{Dt} = \theta_{Dt} + \beta_{DtDt} \qquad \Phi_{Ds} = \theta_{Ds} + \beta_{DsDs} \qquad \Phi_E = \theta_E + \beta_{EE} \tag{15}$$

From equations (5) and (8) it follows that the bulk modulus is given by

$$\bar{K}_T = K_T e^{j\theta KT} = n\partial\bar{P}/\partial n \tag{15A}$$

where

$$K_T = \sec \beta_{PP}\, n\partial P/\partial n = \csc \beta_{PP}\, P\partial\theta_P/\partial n \tag{15B}$$

$$\theta_{KT} = \Phi_P \tag{15C}$$

where Φ_P is given in equation (14), and $n = N/V$.

The renormalization group equations for energy, time, time dimension, and space dimension are for incoherent spacetime written in complex number form as[16-18]

$$\bar{E} + \bar{D}_t\bar{\beta}_E - \bar{D}_s\bar{\beta}_P = E^a + D_t^a\beta_E^a \tag{16}$$

$$\bar{t} - \bar{D}_t\bar{\beta}_E\, \partial\bar{t}/\partial\bar{E} + \bar{D}_s\bar{\beta}_P\, \partial\bar{t}/\partial\bar{P} = t^a - D_t^a\beta_E^a\, \partial t^a/\partial E^a \tag{17}$$

$$\bar{D}_t + \bar{D}_t\bar{\beta}_E\, \partial\bar{D}_t/\partial\bar{E} - \bar{D}_s\bar{\beta}_P\, \partial\bar{D}_t/\partial\bar{P} = D_t^a + D_t^a\beta_E^a\, \partial D_t^a/\partial E^a \tag{18}$$

$$\bar{D}_s + \bar{D}_t\bar{\beta}_E\, \partial\bar{D}_s/\partial\bar{E} - \bar{D}_s\bar{\beta}_P\, \partial\bar{D}_s/\partial\bar{P} = D_s^a + D_t^a\beta_E^a\, \partial D_s^a/\partial E^a \tag{19}$$

where $\bar{E}$, $\bar{t}$, $\bar{D}_t$ and $\bar{D}_s$ = complex number values of the renormalized energy density, time, time dimension and space dimension respectively, E^a , t^a , D_t^a and D_s^a = unrenormalized values respectively of the energy density, time, time dimension and space dimension. Equations (16) through (19) are coupled nonlinear partial differential equations whose solutions are difficult to obtain unless simplifying assumptions are introduced. The gauge parameters for incoherent space are defined as follows[17,18]

$$\bar{\beta}_E = T/V(d\bar{U}/dT)_{\bar{P}V} \qquad \beta_E^a = T/V(dU^a/dT)_{P^aV} \tag{20}$$

$$\bar{\beta}_P = d/dV(\bar{P}V)_{\bar{U}} \qquad \beta_P^a = d/dV(P^aV)_{U^a} \tag{21}$$

where for incoherent space $\bar{U} = \bar{E}V$ and $U^a = E^aV$. For the unrenromalized vacuum $D_t^{av} = 1$ and $D_s^{av} = 3$. If for the renormalized vacuum $D_t^v < 1$ and $D_s^v < 3$, then the vacuum is fractal. If for the spacetime in matter $D_t < 1$ and $D_s < 3$ then this spacetime is fractal. The renormalization group equations (16) through (19) determine the effects of the gauge parameters of real state equations on the energy, process rates, and geometrical structure of spacetime that are associated with bulk matter. These equations are gauge and conformal invariant.[16,17] The time evolution equations for relativistic thermodynamics can be obtained by

requiring conformal and gauge invariance of equation (17) which is the renormalization group equation for time.[17] This brings the time down into the denominator as required for conventional rate equations.

Equation (16) can be written in terms of the internal energy as[17,18]

$$\bar{U} + \bar{D}_t T(d\bar{U}/dT)_{\bar{P}V} - \bar{D}_s V d/dV(\bar{P}V)_{\bar{U}} = U^a + D_t^a T(dU^a/dT)_{P^a V} \tag{22}$$

The form of equation (22) follows from gauge invariance and conformal invariance.[17] The energy density and pressure of solids and low temperature quantum liquids are written in the following form[17,18]

$$\bar{E} = \bar{E}_o + \bar{E}_\sigma T^\sigma \tag{23}$$

$$\bar{P} = \bar{P}_o + \bar{P}_\sigma T^\sigma \qquad \bar{P}_o = - V d\bar{E}_o/dV - \bar{E}_o \tag{24}$$

$$(\sigma - 1)\bar{P}_\sigma = V d\bar{E}_\sigma/dV + \bar{E}_\sigma \qquad \bar{\gamma}_o = \bar{P}_\sigma/\bar{E}_\sigma \tag{25}$$

where $\bar{E}$ and $\bar{P}$ = renormalized energy density and pressure respectively, $\bar{E}_o$ and $\bar{P}_o$ = renormalized zero-temperature values of the energy density and pressure, $\bar{E}_\sigma$ = thermal energy density coefficient, $\bar{\gamma}_o$ = Grüneisen parameter which is independent of temperature, and σ = number that describes the temperature variation. The zero-temperature forms of equation (22) can be written in any of the following ways with $D_t^o = 1$ and $\bar{D}_s^o = 3$ as the zero-temperature values of D_t and D_s[17,18]

$$\bar{E}_o - 3[(1 + \bar{\gamma}_o)\bar{P}_o - \bar{K}_o] = E_o^a \tag{26}$$

$$3V^2 d^2\bar{U}_o/dV^2 + 3(1 + \bar{\gamma}_o)V d\bar{U}_o/dV + \bar{U}_o = U_o^a \tag{27}$$

$$3V^2 d^2\bar{E}_o/dV^2 + 3(3 + \bar{\gamma}_o)V d\bar{E}_o/dV + (3\bar{\gamma}_o + 4)\bar{E}_o = E_o^a \tag{28}$$

$$3V^2 d^2\bar{P}_o/dV^2 + 3(3 + \bar{\gamma}_o)V d\bar{P}_o/dV + [3(\bar{\gamma}_o + V d\bar{\gamma}_o/dV) + 4]\bar{P}_o = P_o^a \tag{29}$$

$$3n^2 d^2\bar{P}_o/dn^2 - 3(1 + \bar{\gamma}_o)n d\bar{P}_o/dn + [3(\bar{\gamma}_o - n d\bar{\gamma}_o/dn) + 4]\bar{P}_o = P_o^a \tag{30}$$

where K_o = zero-temperature value of the bulk modulus given by

$$\bar{K}_o = n d\bar{P}_o/dn = - V d\bar{P}_o/dV \tag{31}$$

and where $\bar{E}_o = \bar{U}_o/V$ = incoherent average energy density at zero temperature. The corresponding T^σ component of equation (16) is given for $\bar{D}_t^o = 1$ and $\bar{D}_s^o = 3$ by[17,18]

$$\bar{E}_\sigma[1 + \sigma + \sigma\bar{\gamma}_o\bar{P}_o/(\bar{P}_o - \bar{K}_o) + 3n d\bar{\gamma}_o/dn] = E_\sigma^a[1 + \sigma + \sigma\gamma_o^a P_o^a/(P_o^a - K_o^a)] \tag{32}$$

The unrenormalized energy density, pressure, bulk modulus and Grüneisen parameter are given by equations (23) through (25) and equation (31) with the superscript "a" added to the equations.

The derivatives in equations (17) through (19) can be written for the general case of the partial coherence of time and dimension and the partial coherence of energy and pressure as

$$\partial\bar{t}/\partial\bar{E} = \sec \beta_{tt} \cos \beta_{EE} \, \partial t/\partial E \, e^{j\Phi tE} \tag{33}$$
$$= \csc \beta_{tt} \sin \beta_{EE} \, t/E \, \partial\theta_t/\partial\theta_E \, e^{j\Phi tE}$$

$$\partial\bar{t}/\partial\bar{P} = \sec \beta_{tt} \cos \beta_{PP} \, \partial t/\partial P \, e^{j\Phi tP} \tag{34}$$
$$= \csc \beta_{tt} \sin \beta_{PP} \, t/P \, \partial\theta_t/\partial\theta_P \, e^{j\Phi tP}$$

$$\partial\bar{D}_t/\partial\bar{E} = \sec \beta_{DtDt} \cos \beta_{EE} \, \partial D_t/\partial E \, e^{j\Phi DtE} \tag{35}$$
$$= \csc \beta_{DtDt} \sin \beta_{EE} \, D_t/E \, \partial\theta_{Dt}/\partial\theta_E \, e^{j\Phi DtE}$$

$$\partial\bar{D}_t/\partial\bar{P} = \sec \beta_{DtDt} \cos \beta_{PP} \, \partial D_t/\partial P \, e^{j\Phi DtP} \tag{36}$$
$$= \csc \beta_{DtDt} \sin \beta_{PP} \, D_t/P \, \partial\theta_{Dt}/\partial\theta_P \, e^{j\Phi DtP}$$

$$\partial\bar{D}_s/\partial\bar{E} = \sec \beta_{DsDs} \cos \beta_{EE} \, \partial D_s/\partial E \, e^{j\Phi DsE} \tag{37}$$
$$= \csc \beta_{DsDs} \sin \beta_{EE} \, D_s/E \, \partial\theta_{Ds}/\partial\theta_E \, e^{j\Phi DsE}$$

$$\partial\bar{D}_s/\partial\bar{P} = \sec \beta_{DsDs} \cos \beta_{PP} \, \partial D_s/\partial P \, e^{j\Phi DsP} \tag{38}$$
$$= \csc \beta_{DsDs} \sin \beta_{PP} \, D_s/P \, \partial\theta_{Ds}/\partial\theta_P \, e^{j\Phi DsP}$$

where

$$\Phi_{tE} = \theta_t + \beta_{tt} - \theta_E - \beta_{EE} \qquad \Phi_{tP} = \theta_t + \beta_{tt} - \theta_P - \beta_{PP} \tag{39}$$

$$\Phi_{DtE} = \theta_{Dt} + \beta_{DtDt} - \theta_E - \beta_{EE} \qquad \Phi_{DtP} = \theta_{Dt} + \beta_{DtDt} - \theta_P - \beta_{PP} \tag{40}$$

$$\Phi_{DsE} = \theta_{Ds} + \beta_{DsDs} - \theta_E - \beta_{EE} \qquad \Phi_{DsP} = \theta_{Ds} + \beta_{DsDs} - \theta_P - \beta_{PP} \tag{41}$$

Equations (17) through (19) are written in terms of the independent variables $\bar{E}$, $\bar{P}$, E^a and P^a but can be written in terms of particle number density n and temperature T so that the set of equations (16) through (19) can be written as follows[16]

$$\bar{E} + \bar{D}_t\bar{\beta}_E - \bar{D}_s\bar{\beta}_P = E^a + D_t^a\beta_E^a \tag{42}$$

$$\bar{t} - \bar{q}_2\ \partial\bar{t}/\partial T + \bar{s}_2\ \partial\bar{t}/\partial n = t^a - q_D^a\ \partial t^a/\partial T + s_D^a\ \partial t^a/\partial n \tag{43}$$

$$\bar{D}_t + \bar{q}_2\ \partial\bar{D}_t/\partial T - \bar{s}_2\ \partial\bar{D}_t/\partial n = D_t^a + q_D^a\ \partial D_t^a/\partial T - s_D^a\ \partial D_t^a/\partial n \tag{44}$$

$$\bar{D}_s + \bar{q}_2\ \partial\bar{D}_s/\partial T - \bar{s}_2\ \partial\bar{D}_s/\partial n = D_s^a + q_D^a\ \partial D_s^a/\partial T - s_D^a\ \partial D_s^a/\partial n \tag{45}$$

where[16]

$$\bar{q}_2 = \bar{h}\bar{D}_t\bar{\beta}_E + \bar{f}\bar{D}_s\bar{\beta}_P \qquad q_D^a = h^a D_t^a\beta_E^a \tag{46}$$

$$\bar{s}_2 = \bar{g}\bar{D}_t\bar{\beta}_E + \bar{e}\bar{D}_s\bar{\beta}_P \qquad s_D^a = g^a D_t^a\beta_E^a \tag{47}$$

$$\bar{e} = 1/\bar{D}_e\ \partial\bar{E}/\partial T \qquad \bar{f} = 1/\bar{D}_e\ \partial\bar{E}/\partial n \tag{48}$$

$$\bar{h} = 1/\bar{D}_e\ \partial\bar{P}/\partial n \qquad \bar{g} = 1/\bar{D}_e\ \partial\bar{P}/\partial T \tag{49}$$

$$\bar{D}_e = \partial\bar{P}/\partial n\ \partial\bar{E}/\partial T - \partial\bar{P}/\partial T\ \partial\bar{E}/\partial n \tag{50}$$

and where h^a and g^a are calculated in the same manner as h and g except that now the superscript "a" is added to $\bar{E}$ and P to indicate the renormalized calculation.

Equations (16) through (19) simplify for the case of incoherent energy and pressure where the energy density and pressure can be taken as real numbers

$$E + D_t\beta_E - D_s\beta_P = E^a + D_t^a\beta_E^a \tag{51}$$

$$\bar{t} - D_t\beta_E\ \partial\bar{t}/\partial E + D_s\beta_P\ \partial\bar{t}/\partial P = t^a - D_t^a\beta_E^a\ \partial t^a/\partial E^a \tag{52}$$

$$\bar{D}_t + D_t\beta_E\ \partial\bar{D}_t/\partial E - D_s\beta_P\ \partial\bar{D}_t/\partial P = D_t^a + D_t^a\beta_E^a\ \partial D^a/\partial E^a \tag{53}$$

$$\bar{D}_s + D_t\beta_E\ \partial\bar{D}_s/\partial E - D_s\beta_P\ \partial\bar{D}_s/\partial P = D_s^a + D_t^a\beta_E^a\ \partial D_s^a/\partial E^a \tag{54}$$

while equations (42) through (45) are written as

$$E + D_t\beta_E - D_s\beta_P = E^a + D_t^a\beta_E^a \tag{55}$$

$$\bar{t} - q_2\ \partial\bar{t}/\partial T + s_2\ \partial\bar{t}/\partial n = t^a - q_D^a\ \partial t^a/\partial T + s_D^a\ \partial t^a/\partial n \tag{56}$$

$$\bar{D}_t + q_2\ \partial\bar{D}_t/\partial T - s_2\ \partial\bar{D}_t/\partial n = D_t^a + q_D^a\ \partial D_t^a/\partial T - s_D^a\ \partial D_t^a/\partial n \tag{57}$$

$$\bar{D}_s + q_2\ \partial\bar{D}_s/\partial T - s_2\ \partial\bar{D}_s/\partial n = D_s^a + q_D^a\ \partial D_s^a/\partial T - s_D^a\ \partial D_s^a/\partial n \tag{58}$$

where q_2 and s_2 are given by equations (46) and (47) with the bars removed. For incoherent space $E = U/V$ and β_E and β_P are given by the real number equivalents of equations (20) and (21).

For coherent time, coherent dimension, but incoherent space, energy and pressure it follows that equations (51) through (54) become[16]

$$E + D_t\beta_E - D_s\beta_P = E^a + D_t^a\beta_E^a \tag{59}$$

$$\bar{t}(1 - jD_t\beta_E\,\partial\theta_t/\partial E + jD_s\beta_P\,\partial\theta_t/\partial P) = t^a - D_t^a\beta_E^a\,\partial t^a/\partial E^a \tag{60}$$

$$\bar{D}_t(1 + jD_t\beta_E\,\partial\theta_{Dt}/\partial E - jD_s\beta_P\,\partial\theta_{Dt}/\partial P) = D_t^a + D_t^a\beta_E^a\,\partial D_t^a/\partial E^a \tag{61}$$

$$\bar{D}_s(1 + jD_t\beta_E\,\partial\theta_{Ds}/\partial E - jD_s\beta_P\,\partial\theta_{Ds}/\partial P) = D_s^a + D_t^a\beta_E^a\,\partial D_s^a/\partial E^a \tag{62}$$

where for coherent time, time dimension and space dimension it follows from equations (3) through (5) that[16]

$$d\bar{t} = j\bar{t}d\theta_t \qquad d\bar{D}_t = j\bar{D}_t d\theta_{Dt} \qquad d\bar{D}_s = j\bar{D}_s d\theta_{Ds} \tag{63}$$

but for incoherent space and incoherent energy β_E and β_P are given by equations (20) and (21) respectively with the bars removed. Equations (60) through (62) describe coherent time and dimension states (for incoherent space and energy) whose coherency arises from spacetime interactions on an incoherent energy and spacetime state. For this case the time and dimensions of the unrenormalized state described by the right hand sides of equations (60) through (62) are incoherent. Both the magnitude and the internal phase angle of the time can be obtained from equation (60) to be[16]

$$\tan\theta_t = D_t\beta_E\,\partial\theta_t/\partial E - D_s\beta_P\,\partial\theta_t/\partial P \tag{64}$$

$$t = \cos\theta_t\,(t^a - D_t^a\beta_E^a\,\partial t^a/\partial E^a) \tag{65}$$

where t, D_t and D_s are taken to be constants. Similarly from equations (61) and (62) it follows that

$$\tan\theta_{Dt} = D_s\beta_P\,\partial\theta_{Dt}/\partial P - D_t\beta_E\,\partial\theta_{Dt}/\partial E \tag{66}$$

$$D_t = \cos\theta_{Dt}\,(D_t^a + D_t^a\beta_E^a\,\partial D_t^a/\partial E^a) \tag{67}$$

$$\tan\theta_{Ds} = D_s\beta_P\,\partial\theta_{Ds}/\partial P - D_t\beta_E\,\partial\theta_{Ds}/\partial E \tag{68}$$

$$D_s = \cos\theta_{Ds}\,(D_s^a + D_t^a\beta_E^a\,\partial D_s^a/\partial E^a) \tag{69}$$

For equations (60) through (69) the renormalized time and dimensions are coherent but space and energy remain incoherent. Equations (60) through (69)

describe a spacetime interaction induced broken symmetry of time, time dimension and space dimension. This system may describe a special form of high-T_c superconductivity where the internal phase angle of space has a constant value $\theta_r = \pi/3$ and the internal phase angle of time is a variable which may possibly be engineered to have the value $\theta_t = \pi/6$ which is associated with high-T_c superconductivity.[16,19] A constant phase angle of the space coordinates would drop out of the energy density and pressure terms in equations (17) through (19) and yield the simplified equations (52) through (54) where E and P are taken to be real numbers. For this type of superconductor the unrenormalized state (E^a, t^a, D_t^a, D_s^a) does not have an intrinsic broken symmetry. It may be possible to engineer the vacuum, by the introduction of external fields, in such a way that ordinary materials become high-T_c superconductors.

For some physical systems the nuclear, atomic or molecular structure induces a broken symmetry in the local spacetime and in the thermodynamic functions of the unrenormalized state. For the case where spacetime has a broken symmetry the energy trace equation is written as

$$\bar{E}' + \bar{D}_t'\bar{\beta}_{\bar{E}}' - \bar{D}_s'\bar{\beta}_P' = \bar{E}^{a'} + \bar{D}_t^{a'}\bar{\beta}_{\bar{E}}^{a'} \tag{70}$$

where now

$$\bar{\beta}_{\bar{E}}' = T/V'(d\bar{U}'/dT)_{\bar{P}'V'} \qquad \bar{\beta}_{\bar{E}}^{a'} = T/V'(d\bar{U}^{a'}/dT)_{\bar{P}^{a'}V'} \tag{71}$$

$$\bar{\beta}_P' = d/dV'(\bar{P}'V')_{\bar{U}'} \qquad \bar{\beta}_P^{a'} = d/dV'(\bar{P}^{a'}V')_{\bar{U}^{a'}} \tag{72}$$

Equivalently, equation (70) and $\bar{U}' = \bar{E}'V'$ and $\bar{U}^{a'} = \bar{E}^{a'}V'$ gives

$$\bar{U}' + \bar{D}_t'T(d\bar{U}'/dT)_{\bar{P}'V'} - \bar{D}_s'V'd/dV'(\bar{P}'V')_{\bar{U}'} = \bar{U}^{a'} + \bar{D}_t^{a'}T(d\bar{U}^{a'}/dT)_{\bar{P}^{a'}V'} \tag{73}$$

where for partially coherent spacetime[16,17]

$$V' = \int|d\bar{V}| = \int \sec\beta_{VV}\, dV = \int \csc\beta_{VV}\, Vd\theta_V \tag{74}$$

$$\bar{V} = Ve^{j\theta V} \qquad d\bar{V} = e^{j\Phi V}\sec\beta_{VV}\, dV = e^{j\Phi V}\csc\beta_{VV}\, Vd\theta_V \tag{75}$$

$$\tan\beta_{VV} = V\partial\theta_V/\partial V \qquad \Phi_V = \theta_V + \beta_{VV} \tag{76}$$

and where now for the first time the possibility of complex number unrenormalized thermodynamic functions on the right hand sides of equations (70) or (73) is considered. The broken symmetry of the unrenormalized state is due to a special structure of matter. From equation (73) it follows that, for the state equations of solids and quantum liquids as in equations (23) through (25), the same form of the relations given in equations (26) through (32) are valid with the replacements $V \to V'$ and $n \to n'$.

The average energy density for partially coherent spacetime is given by

$$\bar{E}' = \bar{U}'/V' \tag{77}$$

where $\bar{E}'$ = average energy density for partially coherent spacetime. For the general case of partially coherent spacetime and partially coherent energy, equations (16) through (19) become

$$\bar{E}' + \bar{D}_t'\bar{\beta}_E' - \bar{D}_s'\bar{\beta}_P' = \bar{E}^{a'} + \bar{D}_t^{a'}\bar{\beta}_E^{a'} \tag{78}$$

$$\bar{t}' - \bar{D}_t'\bar{\beta}_E' \, \partial\bar{t}'/\partial\bar{E}' + \bar{D}_s'\bar{\beta}_P' \, \partial\bar{t}'/\partial\bar{P}' = \bar{t}^{a'} - \bar{D}_t^{a'}\bar{\beta}_E^{a'} \, \partial\bar{t}^{a'}/\partial\bar{E}^{a'} \tag{79}$$

$$\bar{D}_t' + \bar{D}_t'\bar{\beta}_E' \, \partial\bar{D}_t'/\partial\bar{E}' - \bar{D}_s'\bar{\beta}_P' \, \partial\bar{D}_t'/\partial\bar{P}' = \bar{D}_t^{a'} + \bar{D}_t^{a'}\bar{\beta}_E^{a'} \, \partial\bar{D}_t^{a'}/\partial\bar{E}^{a'} \tag{80}$$

$$\bar{D}_s' + \bar{D}_t'\bar{\beta}_E' \, \partial\bar{D}_s'/\partial\bar{E}' - \bar{D}_s'\bar{\beta}_P' \, \partial\bar{D}_s'/\partial\bar{P}' = \bar{D}_s^{a'} + \bar{D}_t^{a'}\bar{\beta}_E^{a'} \, \partial\bar{D}_s^{a'}/\partial\bar{E}^{a'} \tag{81}$$

where $\bar{\beta}_E'$ and $\bar{\beta}_P'$ are given by equations (71) and (72) respectively, and where the unrenormalized values of time, time dimension, space dimension, energy and pressure are now complex numbers because of the special nuclear, atomic or molecular structure of matter as, for example, in the case of high-T_c superconductors. The renormalization group equations (78) through (81) can be recast in terms of particle number density $n' = N/V'$ and temperature T in a form analogous to equations (42) through (45).

The volumes for incoherent space with $\beta_{VV} = 0$, and for coherent space with $\beta_{VV} = \pi/2$, are obtained from equation (74) and are given respectively by

$$V' = V \qquad\qquad V' = V\theta_V \tag{82}$$

where for coherent space V = constant. For coherent spacetime the average energy density is given by

$$\bar{E}^{cs} = \bar{U}^{cs}/(V\theta_V) \qquad\qquad \bar{E}^{csa} = \bar{U}^{csa}/(V\theta_V) \tag{83}$$

where $\bar{U}^{cs}$ and $\bar{E}^{cs}$ = coherent spacetime internal energy and average energy density respectively. For the case of coherent spacetime and partially coherent energy, the gauge functions are obtained from equations (71) and (72) to be

$$\bar{\beta}_E^{cs} = T/(V\theta_V)(d\bar{U}^{cs}/dT)_{\bar{P}^{cs}V\theta_V} \qquad \bar{\beta}_E^{csa} = T/(V\theta_V)(d\bar{U}^{csa}/dT)_{\bar{P}^{csa}V\theta_V} \tag{84}$$

$$\bar{\beta}_P^{cs} = d/d\theta_V(\bar{P}^{cs}\theta_V)_{\bar{U}^{cs}} \qquad \bar{\beta}_P^{csa} = d/d\theta_V(\bar{P}^{csa}\theta_V)_{\bar{U}^{csa}} \tag{85}$$

where V = constant. The renormalization group equations for coherent space, time and dimensions are then obtained from the general set of equations (78) through (81) as follows for a slow process in the superconducting state of a high-T_c superconductor

$$\bar{E}^{cs} + \bar{D}_t^{cs}\bar{\beta}_E^{cs} - \bar{D}_s^{cs}\bar{\beta}_P^{cs} = \bar{E}^{csa} + \bar{D}_t^{csa}\bar{\beta}_E^{csa} \tag{86}$$

$$\bar{t}^{cs}(1 - j\bar{D}_t^{cs}\bar{\beta}_E^{cs}\ \partial\theta_t/\partial\bar{E}^{cs} + j\bar{D}_s^{cs}\bar{\beta}_P^{cs}\ \partial\theta_t/\partial\bar{P}^{cs}) = \bar{t}^{csa}(1 - j\bar{D}_t^{csa}\bar{\beta}_E^{csa}\ \partial\theta_t/\partial\bar{E}^{csa}) \quad (87)$$

$$\bar{D}_t^{cs}(1 + j\bar{D}_t^{cs}\bar{\beta}_E^{cs}\ \partial\theta_{Dt}/\partial\bar{E}^{cs} - j\bar{D}_s^{cs}\bar{\beta}_P^{cs}\ \partial\theta_{Dt}/\partial\bar{P}^{cs}) = \bar{D}_t^{csa}(1 + j\bar{D}_t^{csa}\bar{\beta}_E^{csa}\ \partial\theta_{Dt}^{a}/\partial\bar{E}^{csa}) \quad (88)$$

$$\bar{D}_s^{cs}(1 + j\bar{D}_t^{cs}\bar{\beta}_E^{cs}\ \partial\theta_{Ds}/\partial\bar{E}^{cs} - j\bar{D}_s^{cs}\bar{\beta}_P^{cs}\ \partial\theta_{Ds}/\partial\bar{P}^{cs}) = \bar{D}_s^{csa}(1 + j\bar{D}_t^{csa}\bar{\beta}_E^{csa}\ \partial\theta_{Ds}^{a}/\partial\bar{E}^{csa}) \quad (89)$$

where equation (63) was used to describe the variation of coherent time and dimensions. For this case time, space and dimension are coherent, and the coherency occurs also in the unrenormalized state due to material structure. The energy is partially coherent.

For the case of totally coherent time, space, dimension and internal energy, corresponding to an ultrafast process in the superconducting state of a high-T_c superconductor, the conditions are

$$\beta_{tt} = \pi/2 \qquad \beta_{VV} = \pi/2 \qquad \beta_{UU} = \pi/2 \quad (90)$$

$$\beta_{DtDt} = \pi/2 \qquad \beta_{DsDs} = \pi/2 \quad (91)$$

The average energy density for this case is written as

$$\bar{E}^{tc} = \bar{U}^{tc}/(V\theta_V) \qquad \bar{E}^{tca} = \bar{U}^{tca}/(V\theta_V) \quad (92)$$

where $\bar{U}^{tc}$ and $\bar{E}^{tc}$ = totally coherent internal energy and energy density respectively. The differential of the internal energy is written for pure rotation as the following ultrafast process condition[17]

$$d\bar{U}^{tc} = j\bar{U}^{tc}d\theta_U \qquad d\bar{E}^{tc} = j\bar{E}^{tc}d\theta_E - \bar{E}^{tc}/\theta_V\ d\theta_V \quad (93)$$

where U^{tc} = constant. Equation (63) gives the differentials $d\bar{t}$, $d\bar{D}_t$ and $d\bar{D}_s$ and for this case the derivatives in equations (79) through (81) are written as

$$d\bar{t}^{tc}/d\bar{E}^{tc} = j\bar{t}^{tc}\ d\theta_t/d\bar{E}^{tc} \qquad d\bar{t}^{tc}/d\bar{P}^{tc} = j\bar{t}^{tc}\ d\theta_t/d\bar{P}^{tc} \quad (94)$$

$$d\bar{D}_t^{tc}/d\bar{E}^{tc} = j\bar{D}_t^{tc}\ d\theta_{Dt}/d\bar{E}^{tc} \qquad d\bar{D}_t^{tc}/d\bar{P}^{tc} = j\bar{D}_t^{tc}\ d\theta_{Dt}/d\bar{P}^{tc} \quad (95)$$

$$d\bar{D}_s^{tc}/d\bar{E}^{tc} = j\bar{D}_s^{tc}\ d\theta_{Ds}/d\bar{E}^{tc} \qquad d\bar{D}_s^{tc}/d\bar{P}^{tc} = j\bar{D}_s^{tc}\ d\theta_{Ds}/d\bar{P}^{tc} \quad (96)$$

with analogous expressions for the unrenormalized derivatives. Then the renormalization group equations (78) through (81) are written for total coherence as

$$\bar{E}^{tc} + \bar{D}_t^{tc}\bar{\beta}_E^{tc} - \bar{D}_s^{tc}\bar{\beta}_P^{tc} = \bar{E}^{tca} + \bar{D}_t^{tca}\bar{\beta}_E^{tca} \quad (97)$$

$$\bar{t}^{tc}(1 - j\bar{D}_t^{tc}\bar{\beta}_E^{tc}\ \partial\theta_t/\partial\bar{E}^{tc} + j\bar{D}_s^{tc}\bar{\beta}_P^{tc}\ \partial\theta_t/\partial\bar{P}^{tc}) = \bar{t}^{tca}(1 - j\bar{D}_t^{tca}\bar{\beta}_E^{tca}\ \partial\theta_t^{a}/\partial\bar{E}^{tca}) \quad (98)$$

$$\bar{D}_t^{tc}(1+j\bar{D}_t^{tc}\bar{\beta}_E^{tc}\,\partial\theta_{Dt}/\partial\bar{E}^{tc}-j\bar{D}_s^{tc}\bar{\beta}_P^{tc}\,\partial\theta_{Dt}/\partial\bar{P}^{tc}) = \bar{D}_t^{tca}(1+j\bar{D}_t^{tca}\bar{\beta}_E^{tca}\,\partial\theta_{Dt}^{a}/\partial\bar{E}^{tca}) \quad (99)$$

$$\bar{D}_s^{tc}(1+j\bar{D}_t^{tc}\bar{\beta}_E^{tc}\,\partial\theta_{Ds}/\partial\bar{E}^{tc}-j\bar{D}_s^{tc}\bar{\beta}_P^{tc}\,\partial\theta_{Ds}/\partial\bar{P}^{tc}) = \bar{D}_s^{tca}(1+j\bar{D}_t^{tca}\bar{\beta}_E^{tca}\,\partial\theta_{Ds}^{a}/\partial\bar{E}^{tca}) \quad (100)$$

which are the renormalization group equations for an ultrafast process occurring in the superconducting state of a high-T_c superconductor. The gauge functions for coherent spacetime and coherent internal energy are given by

$$\bar{\beta}_E^{tc} = T/(V\theta_V)(d\bar{U}^{tc}/dT)_{\bar{P}^{tc}V\theta_V} = j\bar{E}^{tc}(Td\theta_U/dT)_{\bar{P}^{tc}V\theta_V} \quad (101)$$

$$\bar{\beta}_P^{tc} = d/d\theta_V(\bar{P}^{tc}\theta_V)_{\bar{U}^{tc}} = \bar{P}^{tc} + (\theta_V d\bar{P}^{tc}/d\theta_V)_{\bar{U}^{tc}} \quad (102)$$

with similar expressions for the unrenormalized gauge functions.

In this paper a material is described as ordinary if the spacetime has a zero or constant broken symmetry. The normal state of a high-T_c material is described by partially coherent spacetime, and the superconducting state is described by a coherent spacetime state. Thermodynamic processes are described as being slow for zero or constant broken symmetry of the thermodynamic functions, moderately fast for partial coherence of the thermodynamic functions, and ultrafast if the thermodynamic functions change coherently.

The effects of the gauge parameters β_E and β_P in the renormalization group equations for energy, time, time dimension and space dimension is greatest for systems that have real state equations with large departures from ideal systems. Therefore at ordinary pressures the renormalization group equations will have significant effects for liquids and solids but the effects on gases will be small except at high pressures. The geometry and reaction rates of chemical processes in liquids and solids will be affected by the gauge parameters which appear in the renormalization group equations that determine $\bar{E}$, t, D_t and D_s. This is true, for instance, for the Belousov-Zhabotinskii reaction in liquids. The reaction rates and the fractal nature of the reaction product geometry for a real system will be described by the renormalized values of the time t and space dimension D_s, and these will be different from the predictions of conventional calculations which give the unrenormalized results t^a and D_s^a respectively for the time and fractal space dimension of the reaction.

This paper develops the quantum eigenvalue equations corresponding to the renormalization group equations for energy, time, time dimension and space dimension that appear in equations (78) through (81) and their variations. Briefly, the summary of the paper is as follows: Section 2 derives the necessary thermodynamic equations for application to quantum thermodynamics, Section 3 derives relativistic trace equations for broken spacetime symmetry, Section 4 introduces quantum thermodynamics and the thermodynamic eigenvalue equations, Section 5 studies the quantum theory of time and dimension and derives first order Dirac-like eigenvalue equations for time and dimension, Section 6 obtains solutions to the first order time and dimension eigenvalue equations, Section 7 treats the substructure of time and dimension, and finally Section 8 considers the quantized time and dimension structures of pressure and energy that can be derived from a Schrödinger-like form of the time and dimension eigenvalue equations.

2. THERMODYNAMICS AND BROKEN SPACETIME SYMMETRY. This section gives a very brief review of broken symmetry thermodynamics. For broken spacetime symmetry the first and second laws of thermodynamics can be written as[17]

$$T d\bar{S} = d\bar{U} + \bar{\bar{P}} d\bar{V} + \bar{\bar{M}} d\bar{\alpha} \tag{103}$$

$$= d\bar{U} + \bar{P}|d\bar{V}| + \bar{M}|d\bar{\alpha}|$$

where[17]

$$\bar{S} = S e^{j\theta_S} \tag{104}$$

$$d\bar{S} = (dS + jS d\theta_S) e^{j\Phi_S}$$

$$= \sec \beta_{SS}\, dS\, e^{j\Phi_S} = \csc \beta_{SS}\, S d\theta_S\, e^{j\Phi_S} \tag{105}$$

where

$$\tan \beta_{SS} = S\partial\theta_S/\partial S \qquad \Phi_S = \theta_S + \beta_{SS} \tag{106}$$

The pressure $\bar{\bar{P}}$ and $\bar{P}$ and the generalized forces $\bar{\bar{M}}$ and $\bar{M}$ are represented by

$$\bar{P} = P e^{j\theta_P} \qquad \bar{\bar{P}} = P e^{j\theta'_P} \tag{107}$$

$$\bar{M} = M e^{j\theta_M} \qquad \bar{\bar{M}} = M e^{j\theta'_M} \tag{108}$$

$$\bar{\bar{P}} d\bar{V} = \bar{P}|d\bar{V}| = \bar{P} dV' \qquad \bar{\bar{M}} d\bar{\alpha} = \bar{M}|d\bar{\alpha}| \tag{109}$$

$$\bar{\alpha} = \alpha e^{j\theta_\alpha} \qquad d\bar{\alpha} = e^{j\Phi_\alpha} \sec \beta_{\alpha\alpha}\, d\alpha = e^{j\Phi_\alpha} \csc \beta_{\alpha\alpha}\, \alpha d\theta_\alpha \tag{110}$$

$$\tan \beta_{\alpha\alpha} = \alpha\partial\theta_\alpha/\partial\alpha \qquad \Phi_\alpha = \theta_\alpha + \beta_{\alpha\alpha} \tag{111}$$

where

$$\theta_P = \theta'_P + \theta_V + \beta_{VV} \qquad \theta_M = \theta'_M + \theta_\alpha + \beta_{\alpha\alpha} \tag{112}$$

From equation (103) the following basic thermodynamic equations can be derived by neglection $d\bar{\alpha}$

$$T\partial\bar{S}/\partial T = \partial\bar{U}/\partial T \tag{113}$$

$$T\partial\bar{S}/|\partial\bar{V}| = \partial\bar{U}/|\partial\bar{V}| + \bar{P} \tag{114}$$

$$\partial\bar{S}/|\partial\bar{V}| = T\partial\bar{P}/\partial T \tag{115}$$

$$\partial\bar{U}/|\partial\bar{V}| = T\partial\bar{P}/\partial T - \bar{P} \tag{116}$$

where $d\bar{P}$ is given by equations (5) and (8). Often these equations are written in terms of the average density of the thermodynamic functions.

The incoherent average energy density $\bar{\mathcal{E}}$ and the average energy density for

broken symmetry spacetime $\bar{E}'$ are given by equations (1) and (77) respectively. Similarly, the incoherent average entropy density $\bar{\mathcal{S}}$ and the average entropy $\bar{\mathcal{S}}'$ for spacetime with partial broken symmetry are given by

$$\bar{\mathcal{S}} = \bar{S}/V \qquad \bar{\mathcal{S}}' = \bar{S}'/V' \tag{117}$$

where V' is given by equation (74). The average energy density for coherent spacetime $\bar{\mathcal{E}}^{cs}$ and the average energy density for coherent thermodynamic functions and coherent spacetime $\bar{\mathcal{E}}^{tc}$ are given by equations (83) and (92). In a similar fashion the average entropy density for coherent spacetime $\bar{\mathcal{S}}^{cs}$ and the average entropy density for both coherent spacetime and coherent thermodynamic functions are given by

$$\bar{\mathcal{S}}^{cs} = \bar{S}^{cs}/(V\theta_V) \qquad \bar{\mathcal{S}}^{tc} = \bar{S}^{tc}/(V\theta_V) \tag{118}$$

The corresponding average densities of the generalized coordinates for incoherent spacetime and for partially coherent spacetime are given by

$$\bar{A} = \bar{\alpha}/V \qquad \bar{A}' = \bar{\alpha}'/V' \tag{119}$$

and the corresponding average values of the generalized coordinate density for coherent spacetime and for total coherence of both spacetime and thermodynamic functions are respectively

$$\bar{A}^{cs} = \bar{\alpha}^{cs}/(V\theta_V) \qquad \bar{A}^{tc} = \bar{\alpha}^{tc}/(V\theta_V) \tag{120}$$

Coherent thermodynamics in coherent spacetime is associated with an ultrafast thermodynamic process in the superconducting state of a high-T_c superconductor. For this case, in addition to $\beta_{tt} = \pi/2$, the following conditions hold

$$\beta_{UU} = \pi/2 \qquad \beta_{SS} = \pi/2 \qquad \beta_{VV} = \pi/2 \qquad \beta_{\alpha\alpha} = \pi/2 \tag{121}$$

$$d\bar{U}^{tc} = j\bar{U}^{tc}d\theta_U \qquad d\bar{S}^{tc} = j\bar{S}^{tc}d\theta_S \qquad d\bar{V} = j\bar{V}d\theta_V \qquad d\bar{\alpha} = j\bar{\alpha}d\theta_\alpha \tag{122}$$

with t , U , S , V and α = constants. In this case equation (103) gives for total coherence

$$\bar{P} = Ts^{tc}\, e^{j(\theta_S+\pi/2)} - e^{tc}\, e^{j(\theta_U+\pi/2)} - Ma^{tc}\, e^{j\theta_M} \tag{123}$$

where the differential totally coherent entropy density, energy density, and generalized coordinate density are given respectively by

$$s^{tc} = \mathcal{S}^{tc}(\theta_V d\theta_S/d\theta_V)_{S,V} \tag{124}$$

$$e^{tc} = E^{tc}(\theta_V d\theta_U/d\theta_V)_{U,V} \tag{125}$$

$$a^{tc} = A^{tc}(\theta_V d\theta_\alpha/d\theta_V)_{\alpha,V} \tag{126}$$

where $\mathcal{S}^{tc}$, E^{tc} and A^{tc} are the magnitudes of the complex numbers given in equations (118), (92) and (120) respectively. A simple solution to equation (123)

uses the approximation

$$\theta_P \sim \theta_M \sim \theta_S + \pi/2 \sim \theta_U + \pi/2 \tag{127}$$

and gives the pressure magnitude as

$$P \sim Ts^{tc} - e^{tc} - Ma^{tc} \tag{128}$$

Further insight into the approximate solutions in equations (127) and (128) can be obtained by noting that for coherent thermodynamics and coherent spacetime associated with ultrafast processes in high-T_c superconductors, equations (113) through (116) can be written as

$$TS\partial\theta_S/\partial T \sim U\partial\theta_U/\partial T \qquad \theta_S \sim \theta_U \qquad TS \sim U \tag{129}$$

$$jT\bar{\mathcal{S}}^{tc}\theta_V\partial\theta_S/\partial\theta_V \sim j\bar{E}^{tc}\theta_V\partial\theta_U/\partial\theta_V + \bar{P}^{tc} \tag{130}$$

$$j\bar{\mathcal{S}}^{tc}\theta_V\partial\theta_S/\partial\theta_V \sim \partial\bar{P}^{tc}/\partial T \tag{131}$$

$$j\bar{E}^{tc}\theta_V\partial\theta_U/\partial\theta_V \sim T\partial\bar{P}^{tc}/\partial T - \bar{P}^{tc} \tag{132}$$

in conjunction with equation (121). From equation (131) it follows that

$$\theta_S + \pi/2 \sim \theta_P + \beta_{PP} \tag{133}$$

A comparison of equations (127) and (133) shows that $\beta_{PP} \sim 0$, and therefore the pressure behaves in an approximately incoherent manner.

3. RELATIVISTIC TRACE EQUATIONS. This section develops relativistic trace equations for four thermodynamic and spacetime conditions: A) partially coherent energy and partially coherent spacetime, B) incoherent energy and incoherent spacetime, C) partially coherent energy and coherent spacetime, and D) coherent energy and coherent spacetime. The case of coherent energy and incoherent spacetime is formally identical to case D. The trace equations are the energy renormalization group equations for bulk matter with real state equations.

Case A. Partially Coherent Energy and Partially Coherent Spacetime.

For partially coherent energy and partially coherent spacetime, corresponding to a moderately fast thermodynamic process in the normal state of a high-T_c superconductor, the spatial volume is represented by V' given by equation (74), and equation (73) is the trace equation for this case

$$\bar{U}' + \bar{D}_t'T(d\bar{U}'/dT)_{\bar{P}'V'} - \bar{D}_s'V'd/dV'(\bar{P}'V')_{\bar{U}'} = \bar{U}^{a'} + \bar{D}_t^{a'}T(d\bar{U}^{a'}/dT)_{\bar{P}^{a'}V'} \tag{134A}$$

or

$$\bar{E}' + \bar{D}_t'\bar{\beta}_{\bar{E}}' - \bar{D}_s'\bar{\beta}_P' = \bar{E}^{a'} + \bar{D}_t^{a'}\bar{\beta}_{\bar{E}}^{a'} \tag{134B}$$

where $d\bar{U}'$ and $d\bar{U}^{a'}$ are given by equation (6). If the space and time dimensions

are written analogous to equations (23) and (24) for solids and low temperature quantum liquids as

$$\bar{D}'_s = \bar{D}'_{so} + \bar{D}'_{s\sigma}T^{\sigma} + \cdots \qquad \bar{D}'_t = \bar{D}'_{to} + \bar{D}'_{t\sigma}T^{\sigma} + \cdots \tag{135}$$

and the internal energy, energy density and pressure as

$$\bar{U}' = \bar{U}'_o + \bar{U}'_\sigma T^{\sigma} \tag{136}$$

$$\bar{E}' = \bar{E}'_o + \bar{E}'_\sigma T^{\sigma} \qquad \bar{P}' = \bar{P}'_o + \bar{P}'_\sigma T^{\sigma} \tag{137}$$

where

$$\bar{E}'_o = \bar{U}'_o/V' \qquad \bar{E}'_\sigma = \bar{U}'_\sigma/V' \tag{138}$$

$$\bar{\gamma}' = (\partial\bar{P}'/\partial T)/(\partial\bar{E}'/\partial T) \qquad \bar{\gamma}'_o = \bar{P}'_\sigma/\bar{E}'_\sigma \tag{139}$$

where $\bar{\gamma}'$ = Grüneisen function associated with the broken symmetry volume V'.

For the case of solids and quantum liquids described by equations (23), (24) and (135), the trace equation (134) has a zero-temperature component and a T^{σ} component. The zero-temperature component is analogous to equations (26) and (30) and can be written in the following equivalent forms[30]

$$\bar{E}'_o - \bar{D}'_{so}[(1 + \bar{\gamma}'_o)\bar{P}'_o - \bar{K}'_o] = \bar{E}^{a'}_o \tag{140}$$

$$\bar{D}'_{so}V'^2 d^2\bar{U}'_o/dV'^2 + \bar{D}'_{so}(1 + \bar{\gamma}'_o)V'd\bar{U}'_o/dV' + \bar{U}'_o = \bar{U}^{a'}_o \tag{141}$$

$$\bar{D}'_{so}V'^2 d^2\bar{E}'_o/dV'^2 + \bar{D}'_{so}(3 + \bar{\gamma}'_o)V'd\bar{E}'_o/dV' + [\bar{D}'_{so}(1 + \bar{\gamma}'_o) + 1]\bar{E}'_o = \bar{E}^{a'}_o \tag{142}$$

$$\bar{D}'_{so}V'^2 d^2\bar{P}'_o/dV'^2 + \bar{D}'_{so}(3 + \bar{\gamma}'_o)V'd\bar{P}'_o/dV' + [\bar{D}'_{so}(\bar{\gamma}'_o + V'd\bar{\gamma}'_o/dV') + \bar{D}'_{so} + 1]\bar{P}'_o = \bar{P}^{a'}_o \tag{143}$$

$$\bar{D}'_{so}n'^2 d^2\bar{P}'_o/dn'^2 - \bar{D}'_{so}(1 + \bar{\gamma}'_o)n'd\bar{P}'_o/dn' + [\bar{D}'_{so}(\bar{\gamma}'_o - n'd\bar{\gamma}'_o/dn') + \bar{D}'_{so} + 1]\bar{P}'_o = \bar{P}^{a'}_o \tag{144}$$

where n' = N/V' and where V' is given by equation (74). For $\bar{D}'_{so} = 3$ equations (140) through (144) become equations (26) through (30). The T^{σ} term of equation (134) is given by

$$\bar{E}'_\sigma[1 + \sigma\bar{D}'_{to} + \sigma\bar{D}'_{to}\bar{\gamma}'_o\bar{P}'_o/(\bar{P}'_o - \bar{K}'_o) - \bar{D}'_{so}V'd\bar{\gamma}'_o/dV' - \bar{D}'_{s\sigma}\bar{\beta}'_{Po}] \tag{145}$$

$$= \bar{E}^{a'}_\sigma[1 + \sigma\bar{D}^{a'}_{to} + \sigma\bar{D}^{a'}_{to}\bar{\gamma}^{a'}_o\bar{P}^{a'}_o/(\bar{P}^{a'}_o - \bar{K}^{a'}_o)]$$

Equation (140) arises from the fact that

$$\bar{\beta}'_{Po} = (1 + \bar{\gamma}'_o)\bar{P}'_o - \bar{K}'_o \tag{146}$$

where $\bar{\beta}'_{Po}$ = zero-temperature value of $\bar{\beta}'_P$ given by equation (72). Equation (145) reduces to equation (32) for the case $\bar{D}'_{to} = 1$, $\bar{D}^{a'}_{to} = 1$, $\bar{D}'_{so} = 3$ and $\bar{D}'_{s\sigma} = 0$.

In these equations the zero-temperature pressures and bulk moduli are given by

$$\bar{P}_o' = - V'd\bar{E}_o'/dV' - \bar{E}_o' \qquad \bar{P}_o^{a\,\prime} = - V'd\bar{E}_o^{a\,\prime}/dV' - \bar{E}_o^{a\,\prime} \tag{147}$$

$$\bar{K}_o' = - V'd\bar{P}_o'/dV' = 2V'd\bar{E}_o'/dV' + V'^2d^2\bar{E}_o'/dV'^2 \tag{148}$$

$$\bar{K}_o^{a\,\prime} = - V'd\bar{P}_o^{a\,\prime}/dV' = 2V'd\bar{E}_o^{a\,\prime}/dV' + V'^2d^2\bar{E}_o^{a\,\prime}/dV'^2 \tag{149}$$

Case B. Incoherent Energy and Incoherent Spacetime.

For incoherent energy and incoherent spacetime, corresponding to a very slow thermodynamic process in ordinary matter, $\theta_U = 0$, $\theta_P = 0$, $\theta_V = 0$ and $V' = V$, and the trace equation is written as

$$U + D_tT(dU/dT)_{PV} - D_sVd/dV(PV)_U = U^a + D_t^aT(dU^a/dT)_{pa_V} \tag{150}$$

or equivalently as

$$E + D_t\beta_E - D_s\beta_P = E^a + D_t^a\beta_E^a \tag{151}$$

All of the equations (140) through (144) are valid for this case if the replacement $V' \to V$ is made and if the bars are removed.

Case C. Partially Coherent Energy in Coherent Spacetime.

This case corresponds to a moderately fast thermodynamic process taking place in the superconducting state of a high-T_c superconductor. For coherent spacetime it follows from equation (74) and the conditions $\beta_{VV} = \pi/2$ and V = constant that $V' = V\theta_V$. Then from equation (73) it follows that

$$\bar{U}^{cs} + \bar{D}_t^{cs}T(d\bar{U}^{cs}/dT)_{\bar{P}^{cs}\theta_V} - \bar{D}_s^{cs}V\theta_V\partial/\partial\theta_V(\bar{P}^{cs}\theta_V)_{\bar{U}^{cs}} = \bar{U}^{csa} + \bar{D}_t^{csa}T(d\bar{U}^{csa}/dT)_{\bar{P}^{csa}\theta_V} \tag{152A}$$

or equivalently as

$$\bar{E}^{cs} + \bar{D}_t^{cs}\bar{\beta}_E^{cs} - \bar{D}_s^{cs}\bar{\beta}_P^{cs} = \bar{E}^{csa} + \bar{D}_t^{csa}\bar{\beta}_E^{csa} \tag{152B}$$

where

$$d\bar{U}^{cs} = e^{j\theta}U(dU^{cs} + jU^{cs}d\theta_U) \qquad dV' = Vd\theta_V \tag{153}$$

The possibility that the unrenormalized functions $\bar{U}^{csa}$, $\bar{P}^{csa}$ and D_t^{csa} in equations (152) are structurally induced complex numbers is now considered. Equation (152B) is obtained from equation (152A) by dividing through by $V\theta_V$, where the resulting energy densities are given by equation (83), and where the gauge parameters $\bar{\beta}_E^{cs}$ and $\bar{\beta}_P^{cs}$ are given by equations (84) and (85) respectively. The solution of equation (152B) begins with the determination of the gauge parameters.

The expression for $\bar{\beta}_P^{cs}$ is given in equation (85). The calculation of $\bar{\beta}_P^{cs}$

begins by using the chain rule for derivatives as follows

$$(d\bar{P}/d\theta_V)_{\bar{U}^{cs}} = (\partial\bar{P}^{cs}/\partial\theta_V)_T + (\partial\bar{P}^{cs}/\partial T)_{\theta_V}(dT/d\theta_V)_{\bar{U}^{cs}} \tag{154}$$

The derivative $dT/d\theta_V$ in equation (154) is obtained from the condition $\bar{U}^{cs}$ = constant as follows

$$d\bar{U}^{cs} = (\partial\bar{U}^{cs}/\partial\theta_V)_T d\theta_V + (\partial\bar{U}^{cs}/\partial T)_{\theta_V} dT = 0 \tag{155}$$

so that

$$(dT/d\theta_V)_{\bar{U}^{cs}} = -(\partial\bar{U}^{cs}/\partial\theta_V)/(\partial\bar{U}^{cs}/\partial T) \tag{156}$$

so that from equation (85) it follows that

$$\bar{\beta}_P^{cs} = \bar{P}^{cs} + \theta_V(\partial\bar{P}^{cs}/\partial\theta_V)_T - \theta_V(\partial\bar{P}^{cs}/\partial T)_{\theta_V}(\partial\bar{U}^{cs}/\partial\theta_V)_T/(\partial\bar{U}^{cs}/\partial T)_{\theta_V} \tag{157}$$

Equation (157) can be rewritten as

$$\bar{\beta}_P^{cs} = \bar{P}^{cs} + \theta_V(\partial\bar{P}^{cs}/\partial\theta_V)_T - \bar{\gamma}^{cs}/V\ (\partial\bar{U}^{cs}/\partial\theta_V)_T \tag{158}$$

where the Grüneisen parameter is given by

$$\bar{\gamma}^{cs} = (\partial\bar{P}^{cs}/\partial T)/(\partial\bar{E}^{cs}/\partial T) \tag{159}$$

where $\bar{U}^{cs}$ and $\bar{E}^{cs}$ are related by equation (83). If equation (116) is written in the form

$$1/V(\bar{U}^{cs}/\partial\theta_V) = T\partial\bar{P}^{cs}/\partial T - \bar{P}^{cs} \qquad \partial/\partial\theta_V(\theta_V\bar{E}^{cs}) = T\partial\bar{P}^{cs}/\partial T - \bar{P}^{cs} \tag{160}$$

then equation (158) gives

$$\bar{\beta}_P^{cs} = \bar{P}^{cs} - \bar{K}_T^{cs} - \bar{\gamma}^{cs}(T\partial\bar{P}^{cs}/\partial T - \bar{P}^{cs}) \tag{161}$$

$$= (1 + \bar{\gamma}^{cs})\bar{P}^{cs} - \bar{K}_T^{cs} - \bar{\gamma}^{cs}(T\partial\bar{P}^{cs}/\partial T)_{\theta_V}$$

where

$$\bar{K}_T^{cs} = -\theta_V(\partial\bar{P}^{cs}/\partial\theta_V)_T \tag{162}$$

and this gives the desired value of $\bar{\beta}_P^{cs}$.

The value of $\bar{\beta}_E^{cs}$ is obtained from equation (84) to be

$$\bar{\beta}_E^{cs} = T/(V\theta_V)[\bar{C}_{V\theta_V}^{cs} + (\partial\bar{U}^{cs}/\partial\theta_V)_T(d\theta_V/dT)_{\bar{P}^{cs}\theta_V}] \tag{163}$$

where the coherent space heat capacity at constant $V\theta_V$ is given by

$$\bar{C}_{V\theta_V}^{cs} = (\partial\bar{U}^{cs}/\partial T)_{V\theta_V} = (\partial\bar{U}^{cs}/\partial T)_{\theta_V} \tag{164}$$

From the condition $\bar{P}^{cs}\theta_V$ = constant it follows that

$$[\bar{P}^{cs} + \theta_V(\partial\bar{P}^{cs}/\partial\theta_V)_T]d\theta_V + \theta_V(\partial\bar{P}^{cs}/\partial T)_{\theta_V} dT = 0 \tag{165}$$

and

$$(d\theta_V/dT)_{\bar{P}^{cs}\theta_V} = - \theta_V(\partial\bar{P}^{cs}/\partial T)_{\theta_V} /[\bar{P}^{cs} + \theta_V(\partial\bar{P}^{cs}/\partial\theta_V)_T] \tag{166}$$

so that

$$\bar{\beta}_E^{cs} = T/(V\theta_V)\bar{C}_{V\theta_V}^{cs} - \bar{\gamma}^{cs}T(\partial\bar{P}^{cs}/\partial T)_{\theta_V}(T\partial\bar{P}^{cs}/\partial T - \bar{P}^{cs})/[\bar{P}^{cs} + \theta_V(\partial\bar{P}^{cs}/\partial\theta_V)_T] \tag{167}$$

which is the required form of $\bar{\beta}_E^{cs}$.

Consider now the case of a solid and low temperature quantum liquid state equation of the form

$$\bar{E}^{cs} = \bar{E}_o^{cs} + \bar{E}_\sigma^{cs}T^\sigma \qquad \bar{P}^{cs} = \bar{P}_o^{cs} + \bar{P}_\sigma^{cs}T^\sigma \tag{168}$$

$$\bar{U}^{cs} = \bar{U}_o^{cs} + \bar{U}_\sigma^{cs}T^\sigma \tag{169}$$

Then the T = 0 component of equation (161) is given by

$$\bar{\beta}_{Po}^{cs} = (1 + \bar{\gamma}_o^{cs})\bar{P}_o^{cs} - \bar{K}_o^{cs} \tag{170}$$

and the energy density components by

$$\bar{E}_o^{cs} = \bar{U}_o^{cs}/(V\theta_V) \qquad \bar{E}_o^{csa} = \bar{U}_o^{csa}/(V\theta_V) \tag{171}$$

$$\bar{E}_\sigma^{cs} = \bar{U}_\sigma^{cs}/(V\theta_V) \qquad \bar{E}_\sigma^{csa} = \bar{U}_\sigma^{csa}/(V\theta_V) \tag{172}$$

Placing equation (168) into equation (160) gives the zero-temperature pressure and the thermal pressure respectively as

$$\bar{P}_o^{cs} = - 1/Vd\bar{U}_o^{cs}/d\theta_V = - d/d\theta_V(\theta_V\bar{E}_o^{cs}) = - \theta_V d\bar{E}_o^{cs}/d\theta_V - \bar{E}_o^{cs} \tag{173}$$

$$(\sigma - 1)\bar{P}_\sigma^{cs} = 1/Vd\bar{U}_\sigma^{cs}/d\theta_V = d/d\theta_V(\theta_V\bar{E}_\sigma^{cs}) = \theta_V d\bar{E}_\sigma^{cs}/d\theta_V + \bar{E}_\sigma^{cs} \tag{174}$$

The zero-temperature bulk modulus for coherent space is given by

$$\bar{K}_o^{cs} = - \theta_V d\bar{P}_o^{cs}/d\theta_V = 2\theta_V d\bar{E}_o^{cs}/d\theta_V + \theta_V^2 d^2\bar{E}_o^{cs}/d\theta_V^2 \tag{175}$$

The corresponding unrenormalized values of the zero-temperature pressure and bulk modulus are given respectively by

$$\bar{P}_o^{csa} = - \theta_V d\bar{E}_o^{csa}/d\theta_V - \bar{E}_o^{csa} \tag{176}$$

$$\bar{K}_o^{csa} = 2\theta_V d\bar{E}_o^{csa}/d\theta_V + \theta_V^2 d^2\bar{E}_o^{csa}/d\theta_V^2 \tag{177}$$

From equations (159) and (168) the Grüneisen parameter is obtained as

$$\bar{\gamma}^{cs} = \bar{P}^{cs}_{\sigma}/\bar{E}^{cs}_{\sigma} + \cdots \tag{178}$$

and therefore

$$\bar{\gamma}^{cs}_{o} = \bar{P}^{cs}_{\sigma}/\bar{E}^{cs}_{\sigma} \tag{179}$$

Analogous to equations (26) and (140) the T = 0 component of equations (86) or (152) is

$$\bar{E}^{cs}_{o} - \bar{D}_{so}[(1 + \bar{\gamma}^{cs}_{o})\bar{P}^{cs}_{o} - \bar{K}^{cs}_{o}] = \bar{E}^{csa}_{o} \tag{180}$$

Equation (180) is equivalent to the following forms

$$\bar{D}_{so}\theta_V^2 d^2\bar{U}^{cs}_{o}/d\theta_V^2 + \bar{D}_{so}(1 + \bar{\gamma}^{cs}_{o})\theta_V d\bar{U}^{cs}_{o}/d\theta_V + \bar{U}^{cs}_{o} = \bar{U}^{csa}_{o} \tag{181}$$

$$\bar{D}_{so}\theta_V^2 d^2\bar{E}^{cs}_{o}/d\theta_V^2 + \bar{D}_{so}(3 + \bar{\gamma}^{cs}_{o})\theta_V d\bar{E}^{cs}_{o}/d\theta_V + [\bar{D}^{cs}_{so}(1 + \bar{\gamma}^{cs}_{o}) + 1]\bar{E}^{cs}_{o} = \bar{E}^{csa}_{o} \tag{182}$$

$$\bar{D}_{so}\theta_V^2 d^2\bar{P}^{cs}_{o}/d\theta_V^2 + \bar{D}_{so}(3 + \bar{\gamma}^{cs}_{o})\theta_V d\bar{P}^{cs}_{o}/d\theta_V + [\bar{D}^{cs}_{so}(\bar{\gamma}^{cs}_{o} + \theta_V d\bar{\gamma}^{cs}_{o}/d\theta_V) + \bar{D}_{so} + 1]\bar{P}^{cs}_{o} = \bar{P}^{csa}_{o} \tag{183}$$

The T^{σ} component of equations (86) or (152) is obtained first by combining equations (167) and (168) to get

$$\bar{\beta}^{cs}_{E} = \bar{E}^{cs}_{\sigma}T^{\sigma}[\sigma + \sigma\bar{\gamma}^{cs}_{o}\bar{P}^{cs}_{o}/(\bar{P}^{cs}_{o} + \theta_V d\bar{P}^{cs}_{o}/d\theta_V)] + \cdots \tag{184}$$

Correspondingly, the value $\bar{\beta}^{cs}_{P}$ can be obtained from equations (161) and (168) to give

$$\bar{\beta}^{cs}_{P} = \bar{\beta}^{cs}_{Po} + \{\bar{E}^{cs}_{\sigma}[\bar{\gamma}^{cs}_{o} - (\bar{\gamma}^{cs}_{o})^2(\sigma - 1)] + \theta_V d\bar{P}^{cs}_{\sigma}/d\theta_V\}T^{\sigma} + \cdots \tag{185}$$

where $\bar{\beta}^{cs}_{Po}$ is given by equation (170). From the definition of $\bar{\gamma}^{cs}_{o}$ given in equation (178), and using equation (174) it is easy to show that

$$\theta_V d\bar{P}^{cs}_{\sigma}/d\theta_V = \bar{E}^{cs}_{\sigma}[\theta_V d\bar{\gamma}^{cs}_{o}/d\theta_V + (\bar{\gamma}^{cs}_{o})^2(\sigma - 1) - \bar{\gamma}^{cs}_{o}] \tag{186}$$

Finally combining equation (185) and (186) gives

$$\bar{\beta}^{cs}_{P} = \bar{\beta}^{cs}_{Po} + \bar{E}^{cs}_{\sigma}\theta_V d\bar{\gamma}^{cs}_{o}/d\theta_V T^{\sigma} \tag{187}$$

Combining equations (86), (184) and (187) gives the T^{σ} component of equation (152) as

$$\bar{E}^{cs}_{\sigma}[1 + \sigma\bar{D}^{cs}_{to} + \sigma\bar{D}^{cs}_{to}\bar{\gamma}^{cs}_{o}\bar{P}^{cs}_{o}/(\bar{P}^{cs}_{o} - \bar{K}^{cs}_{o}) - \bar{D}^{cs}_{so}\theta_V d\bar{\gamma}^{cs}_{o}/d\theta_V - \bar{D}^{cs}_{s\sigma}\bar{\beta}^{cs}_{Po}] \tag{188}$$

$$= \bar{E}^{csa}_{\sigma}[1 + \sigma\bar{D}^{csa}_{to} + \sigma\bar{D}^{csa}_{to}\bar{\gamma}^{csa}_{o}\bar{P}^{csa}_{o}/(\bar{P}^{csa}_{o} - \bar{K}^{csa}_{o})]$$

From equations (174) and (179) it follows that

$$\bar{E}_{\sigma}^{cs}/\bar{E}_{\sigma}^{csa} = \exp[(\sigma-1)\int(\bar{\gamma}_{o}^{cs} - \bar{\gamma}_{o}^{csa})d\theta_{V}/\theta_{V}] \tag{189}$$

Equation (188) is similar to equation (32) which describes incoherent spacetime. In equation (188) $\bar{P}_{o}^{cs}$ and $\bar{P}_{o}^{csa}$ are given by coherent space equations (173) and (176) while $\bar{K}_{o}^{cs}$ and $\bar{K}_{o}^{csa}$ are given by equations (175) and (177), whereas in equation (32) $\bar{P}_{o}$ and P_{o}^{a} are given by the incoherent space equations (23) and (24) while $\bar{K}_{o}$ and K_{o}^{a} are given by equation (31). Likewise, equation (188) is similar to equation (145) except that in equation (145) $\bar{P}_{o}'$, $\bar{K}_{o}'$, $\bar{P}_{o}^{a'}$ and $\bar{K}_{o}^{a'}$ are given by equations (147) through (149).

D. Coherent Energy and Coherent Spacetime.

For this case of an ultrafast thermodynamic process occurring in the superconducting state of a high-T_c superconductor, the variation of the internal energy and spatial volume is given in equation (122) and the relativistic trace is obtained from equations (73) and (122) to be

$$\bar{U}^{tc} + j\bar{D}_{t}^{tc}\bar{U}^{tc}(d\theta_{U}/dT)_{\bar{P}^{tc}\theta_{V}} - \bar{D}_{s}^{tc}V\theta_{V}\partial/\partial\theta_{V}(\bar{P}^{tc}\theta_{V})_{\bar{U}^{tc}} = \bar{U}^{tca} + j\bar{D}_{t}^{tca}\bar{U}^{tca}(d\theta_{U}^{a}/dT)_{\bar{P}^{tca}\theta_{V}} \tag{190A}$$

or equivalently as

$$\bar{E}^{tc} + \bar{D}_{t}^{tc}\bar{\beta}_{\bar{E}}^{tc} - \bar{D}_{s}^{tc}\bar{\beta}_{P}^{tc} = \bar{E}^{tca} + \bar{D}_{t}^{tca}\bar{\beta}_{\bar{E}}^{tca} \tag{190B}$$

with U^{tc} = constant and V = constant. Thus the problem here is to determine $\bar{\beta}_{\bar{E}}^{tc}$ and $\bar{\beta}_{P}^{tc}$.

For solids and low temperature quantum liquids the pressure, internal energy, Grüneisen parameter and average energy density are given for the case at hand by

$$\bar{P}^{tc} = \bar{P}_{o}^{tc} + \bar{P}_{\sigma}^{tc}T^{\sigma} \qquad \bar{P}^{tca} = \bar{P}_{o}^{tca} + \bar{P}_{\sigma}^{tca}T^{\sigma} \tag{191}$$

$$\bar{U}^{tc} = \bar{U}_{o}^{tc} + \bar{U}_{\sigma}^{tc}T^{\sigma} \qquad \bar{U}^{tca} = \bar{U}_{o}^{tca} + \bar{U}_{\sigma}^{tca}T^{\sigma} \tag{192}$$

$$\bar{E}^{tc} = \bar{E}_{o}^{tc} + \bar{E}_{\sigma}^{tc}T^{\sigma} \qquad \bar{E}^{tca} = \bar{E}_{o}^{tca} + \bar{E}_{\sigma}^{tca}T^{\sigma} \tag{193}$$

$$\bar{\gamma}^{tc} = (\partial\bar{P}^{tc}/\partial T)/(\partial\bar{E}^{tc}/\partial T) \qquad \bar{\gamma}_{o}^{tc} = \bar{P}_{\sigma}^{tc}/\bar{E}_{\sigma}^{tc} \qquad \bar{\gamma}_{o}^{tca} = \bar{P}_{\sigma}^{tca}/\bar{E}_{\sigma}^{tca} \tag{194}$$

$$\bar{E}^{tc} = \bar{U}^{tc}/(V\theta_{V}) \qquad \bar{E}^{tca} = \bar{U}^{tca}/(V\theta_{V}) \tag{195}$$

$$\bar{E}_{o}^{tc} = \bar{U}_{o}^{tc}/(V\theta_{V}) \qquad \bar{E}_{o}^{tca} = \bar{U}_{o}^{tca}/(V\theta_{V}) \tag{196}$$

$$\bar{E}_{\sigma}^{tc} = \bar{U}_{\sigma}^{tc}/(V\theta_{V}) \qquad \bar{E}_{\sigma}^{tca} = \bar{U}_{\sigma}^{tca}/(V\theta_{V}) \tag{197}$$

where

$$\bar{U}^{tc} = U^{tc}e^{j\theta U} \qquad \bar{U}_o^{tc} = U_o^{tc}e^{j\theta Uo} \qquad \bar{U}_\sigma^{tc} = U_\sigma^{tc}e^{j\theta U\sigma} \tag{198}$$

$$\bar{U}^{tca} = U^{tca}e^{j\theta^a_U} \qquad \bar{U}_o^{tca} = U_o^{tca}e^{j\theta^a_{Uo}} \qquad \bar{U}_\sigma^{tca} = U_\sigma^{tca}e^{j\theta^a_{U\sigma}} \tag{199}$$

For total coherence equations (116) and (132) are written in the form

$$1/V\partial\bar{U}^{tc}/\partial\theta_V = j\bar{E}^{tc}\theta_V\partial\theta_U/\partial\theta_V = T\partial\bar{P}^{tc}/\partial T - \bar{P}^{tc} \tag{200}$$

For T = 0 equations (191), (192) and (200 give the zero-temperature value of the totally coherent pressure as

$$\bar{P}_o^{tc} = -1/Vj\bar{U}_o^{tc}d\theta_{Uo}/d\theta_V = -j\bar{E}_o^{tc}\theta_V d\theta_{Uo}/d\theta_V \tag{201}$$

$$\bar{P}_o^{tca} = -1/Vj\bar{U}_o^{tca}d\theta^a_{Uo}/d\theta_V = -j\bar{E}_o^{tca}\theta_V d\theta^a_{Uo}/d\theta_V \tag{202}$$

where the phase angle of the internal energy is independent of temperature in this case. The thermal component of the pressure is given by equations (191), (192) and (200) as

$$(\sigma-1)\bar{P}_\sigma^{tc} = 1/Vd\bar{U}_\sigma^{tc}/d\theta_V = j\bar{E}_\sigma^{tc}\theta_V d\theta_{U\sigma}/d\theta_V \tag{203}$$

The Gruneisen parameters $\bar{\gamma}_o^{tc}$ and $\bar{\gamma}_o^{tca}$ are obtained from equations (193) and (203) to be

$$\bar{\gamma}_o^{tc} = 1/(\sigma-1)\theta_V/\bar{U}_\sigma^{tc}d\bar{U}_\sigma^{tc}/d\theta_V = 1/(\sigma-1)j\theta_V d\theta_{U\sigma}/d\theta_V \tag{204}$$

$$\bar{\gamma}_o^{tca} = 1/(\sigma-1)\theta_V/\bar{U}_\sigma^{tca}d\bar{U}_\sigma^{tca}/d\theta_V = 1/(\sigma-1)j\theta_V d\theta^a_{U\sigma}/d\theta_V \tag{205}$$

Therefore for total coherence of spacetime and thermodynamic functions

$$\gamma_o^{tc} = 1/(\sigma-1)\theta_V d\theta_{U\sigma}/d\theta_V \qquad \theta_\gamma^{tc} = \pi/2 \tag{205A}$$

The calculation of $\bar{\beta}_E^{tc}$ is obtained from equation (101) to be

$$\bar{\beta}_E^{tc} = T/(V\theta_V)[\bar{C}_{V\theta_V}^{tc} + (\partial\bar{U}^{tc}/\partial\theta_V)_T(d\theta_V/dT)_{\bar{P}^{tc}\theta_V}] \tag{206}$$

where the totally coherent heat capacity for constant $V\theta_V$ is given by

$$\bar{C}_{V\theta_V}^{tc} = j\bar{U}^{tc}\partial\theta_U/\partial T \tag{207}$$

and

$$T/(V\theta_V)\bar{C}_{V\theta_V}^{tc} = j\bar{E}^{tc}T\partial\theta_U/\partial T \tag{208}$$

by equation (92). Then in analogy to equation (167)

$$\bar{\beta}_E^{tc} = T/(V\theta_V)\bar{C}_{V\theta_V}^{tc} - \bar{\gamma}^{tc}(T\partial\bar{P}^{tc}/\partial T)_{\theta_V}(T\partial\bar{P}^{tc}/\partial T - \bar{P}^{tc})/[\bar{P}^{tc} + \theta_V(\partial\bar{P}^{tc}/\partial\theta_V)_T] \tag{209}$$

Using the definition in equation (102) and the expression for the Grüneisen parameter in equation (193) allows $\bar{\beta}_P^{tc}$ to be calculated as follows

$$\bar{\beta}_P^{tc} = \partial/\partial\theta_V(\bar{P}^{tc}\theta_V)_{\bar{U}^{tc}} = \bar{P}^{tc} + \theta_V(\partial\bar{P}^{tc}/\partial\theta_V)_{\bar{U}^{tc}} \tag{210}$$

$$= \bar{P}^{tc} + \theta_V(\partial\bar{P}^{tc}/\partial\theta_V)_T - \bar{\gamma}^{tc}/V(\partial\bar{U}^{tc}/\partial\theta_V)_T$$

$$= (1 + \bar{\gamma}^{tc})\bar{P}^{tc} - \bar{K}_T^{tc} - \bar{\gamma}^{tc}(T\partial\bar{P}^{tc}/\partial T)_{\theta_V}$$

where now

$$\bar{K}_T^{tc} = - \theta_V(\partial\bar{P}^{tc}/\partial\theta_V)_T \tag{211}$$

The zero-temperature version of equation (210) is

$$\bar{\beta}_{Po}^{tc} = (1 + \bar{\gamma}_o^{tc})\bar{P}_o^{tc} - \bar{K}_o^{tc} \tag{212}$$

with the zero-temperature bulk modulus given by

$$\bar{K}_o^{tc} = - \theta_V(d\bar{P}_o^{tc}/d\theta_V) \qquad \bar{K}_o^{tca} = - \theta_V(d\bar{P}_o^{tca}/d\theta_V) \tag{213}$$

For total coherence the zero-temperature bulk modulus can be rewritten using equations (201), (202) and (213) as

$$\bar{K}_o^{tc} = j\theta_V d/d\theta_V(\bar{E}_o^{tc}\theta_V d\theta_{Uo}/d\theta_V) \tag{214}$$

$$\bar{K}_o^{tca} = j\theta_V d/d\theta_V(\bar{E}_o^{tca}\theta_V d\theta_{Uo}^a/d\theta_V) \tag{215}$$

From equation (93) and remembering that V = constant for the case of total coherence it follows that

$$d\bar{E}_o^{tc}/d\theta_V = j\bar{E}_o^{tc}d\theta_{Uo}/d\theta_V - \bar{E}_o^{tc}/\theta_V \tag{216}$$

Combining equations (214), (215) and (216) gives

$$\bar{K}_o^{tc} = \bar{E}_o^{tc}[j\theta_V^2 d^2\theta_{Uo}/d\theta_V^2 - (\theta_V d\theta_{Uo}/d\theta_V)^2] \tag{217A}$$

$$\bar{K}_o^{tca} = \bar{E}_o^{tca}[j\theta_V^2 d^2\theta_{Uo}^a/d\theta_V^2 - (\theta_V d\theta_{Uo}^a/d\theta_V)^2] \tag{217B}$$

Combining equations (191) through (193) and equation (210) gives

$$\bar{\beta}_P^{tc} = \bar{\beta}_{Po}^{tc} + \bar{E}_\sigma^{tc}\theta_V d\bar{\gamma}_o^{tc}/d\theta_V T^\sigma \tag{217C}$$

where equation (204) can be used to evaluate the derivative in equation (217C).

The zero-temperature trace equation for total coherence is obtained from equation (190) to be

$$\bar{E}_o^{tc} - \bar{D}_{so}^{tc}[(1 + \bar{\gamma}_o^{tc})\bar{P}_o^{tc} - \bar{K}_o^{tc}] = \bar{E}_o^{tca} \tag{218}$$

Combining equations (201), (217A) and (218) gives

$$\bar{E}_o^{tc}\{\bar{D}_{so}^{tc}[j\theta_V^2 d^2\theta_{Uo}/d\theta_V^2 - (\theta_V d\theta_{Uo}/d\theta_V)^2] + j\bar{D}_{so}^{tc}(1+\bar{\gamma}_o^{tc})\theta_V d\theta_{Uo}/d\theta_V + 1\} = \bar{E}_o^{tca} \quad (219)$$

The T^σ component of equation (190) for total coherence is

$$\bar{E}_\sigma^{tc}[1 + \sigma\bar{D}_{to}^{tc} + \sigma\bar{D}_{to}^{tc}\bar{\gamma}_o^{tc}\bar{P}_o^{tc}/(\bar{P}_o^{tc} - \bar{K}_o^{tc}) - \bar{D}_{so}^{tc}\theta_V d\bar{\gamma}_o^{tc}/d\theta_V - \bar{D}_{s\sigma}^{tc}\bar{\beta}_{Po}^{tc}] \quad (220)$$

$$= \bar{E}_\sigma^{tca}[1 + \sigma\bar{D}_{to}^{tca} + \sigma\bar{D}_{to}^{tc}\bar{\gamma}_o^{tca}\bar{P}_o^{tca}/(\bar{P}_o^{tca} - \bar{K}_o^{tca})]$$

where $\bar{P}_o^{tc}$ and $\bar{P}_o^{tca}$ are given by equations (201) and (202), and $\bar{K}_o^{tc}$ and $\bar{K}_o^{tca}$ are given by equations (214) and (215). Equation (220) can be solved by noting that from equations (204) and (205) it follows that

$$\bar{U}_\sigma^{tc}/\bar{U}_\sigma^{tca} = \bar{E}_\sigma^{tc}/\bar{E}_\sigma^{tca} = \exp[(\sigma-1)\int(\bar{\gamma}_o^{tc} - \bar{\gamma}_o^{tca})d\theta_V/\theta_V] \quad (221)$$

$$\theta_{U\sigma} = -j(\sigma-1)\int\bar{\gamma}_o^{tc}d\theta_V/\theta_V + c \qquad \theta_{U\sigma}^a = -j(\sigma-1)\int\bar{\gamma}_o^{tca}d\theta_V/\theta_V + c^a \quad (222)$$

$$= (\sigma-1)\int\gamma_o^{tc}d\theta_V/\theta_V + c \qquad = (\sigma-1)\int\gamma_o^{tc}d\theta_V/\theta_V + c^a$$

In this way the zero-temperature and thermal portions of the internal energy are calculated for coherent thermodynamics and coherent spacetime.

The values of the internal energy, Grüneisen parameter, and dimensions depend on the type of broken symmetry exhibited by the thermodynamic functions and by the spacetime. Thus $\bar{U}^{tc}$, $\bar{\gamma}^{tc}$, $\bar{D}_t^{tc}$ and $\bar{D}_s^{tc}$ are different from $\bar{U}^{cs}$, $\bar{\gamma}^{cs}$, $\bar{D}_t^{cs}$ and $\bar{D}_s^{cs}$, and from $\bar{U}'$, $\bar{\gamma}'$, $\bar{D}_t'$ and $\bar{D}_s'$. Each set of functions is determined from its own form of relativistic trace equation combined with the equations of time and dimensions. The case of coherent thermodynamics in an incoherent space, corresponding to an ultrafast process in ordinary matter, can be obtained from the results of Case D by making the substitution $V\theta_V \to V$ in equations (190) through (222).

4. QUANTUM THERMODYMANICS. In this section relativistic thermodynamic eigenvalue equations are developed that describe the discrete and continuous spectra of states of a thermodynamic system. This quantum structure can exist for the cases where the thermodynamic functions are either incoherent or coherent and when the spacetime is either coherent or incoherent. The eigenvalue equations can be written for either the internal energy, energy density or pressure. The quantum states are self activated in the sense that the source terms in the right hand sides of the relativistic energy trace equations (16), (78), (86) and (97) are assumed to be proportional to the corresponding renormalized quantities. Macroscopic quantum systems with real state equations are expected to be described by eigenvalue equations of this type.

Case A. Partially Coherent Energy and Partially Coherent Spacetime.

The eigenvalue equation corresponding to the trace equation (78) is written for a comparitively fast thermodynamic process in the normal state of a

high-T_c superconductor as

$$\bar{E}' + \bar{D}'_t\bar{\beta}'_E - \bar{D}'_s\bar{\beta}'_P = \bar{\mu}'(\bar{E}' + \bar{D}'_t\bar{\beta}'_E) \tag{223}$$

or

$$(1 - \bar{\mu}')(\bar{E}' + \bar{D}'_t\bar{\beta}'_E) - \bar{D}'_s\bar{\beta}'_P = 0 \tag{224}$$

where $\bar{E}' = \bar{U}'/V'$ and $\bar{\beta}'_E$ and $\bar{\beta}'_P$ are given by equations (71) and (72) respectively. Equation (224) may be generalized with the addition of an external potential yielding

$$(1 - \bar{\mu}')(\bar{E}' + \bar{D}'_t\bar{\beta}'_E) - \bar{D}'_s\bar{\beta}'_P + \bar{W}'_E\bar{E}' = 0 \tag{225}$$

Equation (225) yields eigenfunctions for discrete or continuous eigenvalues $\bar{\mu}'$. For a noninteracting system with $\bar{\beta}'_E = 0$ and $\bar{\beta}'_P = 0$ it follows from equation (224) that $\bar{\mu}' = 1$, while $\bar{\mu}' = 1 + \bar{W}'_E$ for a noninteracting system in an external potential.

The zero-temperature eigenvalue equation corresponding to equation (225) is written as

$$(1 - \bar{\mu}' + \bar{W}'_E)\bar{E}'_o - \bar{D}'_{so}[(1 + \bar{\gamma}'_o)\bar{P}'_o - \bar{K}'_o] = 0 \tag{226}$$

where $\bar{P}'_o$ and $\bar{K}'_o$ are given by equations (147) and (148) respectively. Equation (226) is equivalent to any of the following eigenvalue equations

$$\bar{D}'_{so}V'^2d^2\bar{U}'_o/dV'^2 + \bar{D}'_{so}(1 + \bar{\gamma}'_o)V'd\bar{U}'_o/dV' + (1 - \bar{\mu}' + \bar{W}'_E)\bar{U}'_o = 0 \tag{227}$$

$$D'_{so}V'^2d^2\bar{E}'_o/dV'^2 + \bar{D}'_{so}(3 + \bar{\gamma}'_o)V'd\bar{E}'_o/dV' + [\bar{D}'_{so}(\bar{\gamma}'_o + 1) + 1 - \bar{\mu}' + \bar{W}'_E]\bar{E}'_o = 0 \tag{228}$$

$$\bar{D}'_{so}n'^2d^2\bar{E}'_o/dn'^2 - \bar{D}'_{so}(1 + \bar{\gamma}'_o)n'd\bar{E}'_o/dn' + [\bar{D}'_{so}(\bar{\gamma}'_o + 1) + 1 - \bar{\mu}' + \bar{W}'_E]\bar{E}'_o = 0 \tag{229}$$

$$\bar{D}'_{so}V'^2d^2\bar{P}'_o/dV'^2 + \bar{D}'_{so}(3 + \bar{\gamma}'_o)V'd\bar{P}'_o/dV' + [\bar{D}'_{so}(\bar{\gamma}'_o + V'd\bar{\gamma}'_o/dV') + \bar{D}'_{so} + 1 - \bar{\mu}' + \bar{W}'_E]\bar{P}'_o = 0 \tag{230}$$

$$\bar{D}'_{so}n'^2d^2\bar{P}'_o/dn'^2 - \bar{D}'_{so}(1 + \bar{\gamma}'_o)n'd\bar{P}'_o/dn' + [\bar{D}'_{so}(\bar{\gamma}'_o - n'd\bar{\gamma}'_o/dn') + \bar{D}'_{so} + 1 - \bar{\mu}' + \bar{W}'_E]\bar{P}'_o = 0 \tag{231}$$

The general expressions for the space and time dimensions are given in equation (135). For a three dimensional space $\bar{D}'_{so} = 3$ and $\bar{D}'_{s\sigma} = 0$ and equations (226) through (231) become more simple. For example, equation (229) becomes

$$3n'^2d^2\bar{E}'_o/dn'^2 - 3(1 + \bar{\gamma}'_o)n'd\bar{E}'_o/dn + (3\bar{\gamma}'_o + 4 - \bar{\mu}' + \bar{W}'_E)\bar{E}'_o = 0 \tag{232}$$

as the zero-temperature energy eigenvalue equation.

The T^σ component of the eigenvalue equation (225) is easily obtained from equation (145) to be

$$(1 - \bar{\mu}')[1 + \sigma\bar{D}'_{to} + \sigma\bar{D}'_{to}\bar{\gamma}'_o\bar{P}'_o/(\bar{P}'_o - \bar{K}'_o)] - \bar{D}'_{so}V'd\bar{\gamma}'_o/dV' - \bar{D}'_{s\sigma}\bar{\beta}'_{Po} + \bar{W}'_E = 0 \tag{233}$$

where $\bar{P}_o'$ and $\bar{K}_o'$ are given by equations (147) and (148) respectively. For a 3 + 1 dimensional space $\bar{D}_{so}' = 3$, $\bar{D}_{to}' = 1$, $\bar{D}_{s\sigma}' = 0$ [$D_{t\sigma}'$ does not enter the trace equation (78) or the eigenvalue equation (225) in the T^{σ} term and so it it does not appear in equation (233)], and for this case equation (233) becomes

$$(1 - \bar{\mu}')[1 + \sigma + \sigma\bar{\gamma}_o'\bar{P}_o'/(\bar{P}_o' - \bar{K}_o')] - 3V'd\bar{\gamma}_o'/dV' + \bar{W}_E' = 0 \tag{234}$$

It should be noted that by the definition of an eigenvalue, $\bar{\mu}'$ is simply a number and not a temperature and density dependent thermodynamic function. Also it is assumed that the external potential $\bar{W}_E'$ is not a function of temperature and density.

Case B. Partially Coherent Energy and Incoherent Spacetime.

For this case the eigenvalue equations corresponding to the trace equation (16) can be obtained from the general case equations (223) through (234) by taking V' = V (corresponding to $\theta_V = 0$) and eliminating the primes from all quantities. This case corresponds to a comparitively fast thermodynamic process in an ordinary material.

Case C. Partially Coherent Energy and Coherent Spacetime.

For this case the trace equation (152) gives the following eigenvalue equation for the superconducting state of high-T_c materials

$$(1 - \bar{\mu}^{cs})\left[\bar{U}^{cs} + \bar{D}_t^{cs}T(d\bar{U}^{cs}/dT)_{\bar{P}^{cs}V\theta_V}\right] - \bar{D}_s^{cs}\theta_V d/d\theta_V(\bar{P}^{cs}V\theta_V)_{\bar{U}^{cs}} + \bar{W}_E^{cs}\bar{U}^{cs} = 0 \tag{235}$$

where for generality an external potential term is added. In equation (235) V = constant. Equation (235) can be rewritten as

$$(1 - \bar{\mu}^{cs})(\bar{E}^{cs} + \bar{D}_t^{cs}\bar{\beta}_E^{cs}) - \bar{D}_s^{cs}\bar{\beta}_P^{cs} + \bar{W}_E^{cs}\bar{E}^{cs} = 0 \tag{236}$$

where $\bar{E}^{cs}$, $\bar{\beta}_E^{cs}$ and $\bar{\beta}_P^{cs}$ are given by equations (83) through (85) respectively. For solids and low temperature quantum liquids the zero-temperature form of equation (236) is given by

$$(1 - \bar{\mu}^{cs} + \bar{W}_E^{cs})\bar{E}_o^{cs} - \bar{D}_{so}^{cs}[(1 + \bar{\gamma}_o^{cs})\bar{P}_o^{cs} - \bar{K}_o^{cs}] = 0 \tag{237}$$

where $\bar{P}_o^{cs}$ and $\bar{K}_o^{cs}$ are given by equations (173) and (175) respectively. Equation (237) can be written in the following equivalent forms

$$\bar{D}_{so}^{cs}\theta_V^2 d^2\bar{U}_o^{cs}/d\theta_V^2 + \bar{D}_{so}^{cs}(1 + \bar{\gamma}_o^{cs})\theta_V d\bar{U}_o^{cs}/d\theta_V + (1 - \bar{\mu}^{cs} + \bar{W}_E^{cs})\bar{U}_o^{cs} = 0 \tag{238}$$

$$\bar{D}_{so}^{cs}\theta_V^2 d^2\bar{E}_o^{cs}/d\theta_V^2 + \bar{D}_{so}^{cs}(3 + \bar{\gamma}_o^{cs})\theta_V d\bar{E}_o^{cs}/d\theta_V + [\bar{D}_{so}^{cs}(\bar{\gamma}_o^{cs} + 1) + 1 - \bar{\mu}^{cs} + \bar{W}_E^{cs}]\bar{E}_o^{cs} = 0 \tag{239}$$

$$\bar{D}_{so}^{cs}\xi^2 d^2\bar{E}_o^{cs}/d\xi^2 - \bar{D}_{so}^{cs}(1 + \bar{\gamma}_o^{cs})\xi d\bar{E}_o^{cs}/d\xi + [\bar{D}_{so}^{cs}(\bar{\gamma}_o^{cs} + 1) + 1 - \bar{\mu}^{cs} + \bar{W}_E^{cs}]\bar{E}_o^{cs} = 0 \tag{240}$$

$$\bar{D}^{cs}_{so}\theta_V^2 d^2\bar{P}^{cs}_o/d\theta_V^2 + \bar{D}^{cs}_{so}(3 + \bar{\gamma}^{cs}_o)\theta_V d\bar{P}^{cs}_o/d\theta_V \tag{241}$$

$$+ [\bar{D}_{so}(\bar{\gamma}^{cs}_o + \theta_V d\bar{\gamma}^{cs}_o/d\theta_V) + \bar{D}^{cs}_{so} + 1 - \bar{\mu}^{cs} + \bar{W}^{cs}_E]\bar{P}^{cs}_o = 0$$

$$\bar{D}^{cs}_{so}\xi^2 d^2\bar{P}^{cs}_o/d\xi^2 - \bar{D}^{cs}_{so}(1 + \bar{\gamma}^{cs}_o)\xi d\bar{P}^{cs}_o/d\xi \tag{242}$$

$$+ [\bar{D}^{cs}_{so}(\bar{\gamma}^{cs}_o - \xi d\bar{\gamma}^{cs}_o/d\xi) + \bar{D}^{cs}_{so} + 1 - \bar{\mu}^{cs} + \bar{W}^{cs}_E]\bar{P}^{cs}_o = 0$$

where $\xi = N/(V\theta_V)$ and $\bar{E}^{cs}_o = \bar{U}^{cs}_o/(V\theta_V)$.

The T^σ component of the eigenvalue equation (236) is written as

$$(1 - \bar{\mu}^{cs})[1 + \sigma\bar{D}^{cs}_{to} + \sigma\bar{D}^{cs}_{to}\bar{\gamma}^{cs}_o\bar{P}^{cs}_o/(\bar{P}^{cs}_o - \bar{K}^{cs}_o)] \tag{243}$$

$$- \bar{D}^{cs}_{so}\theta_V d\bar{\gamma}^{cs}_o/d\theta_V - \bar{D}^{cs}_{s\sigma}\bar{\beta}^{cs}_{Po} + \bar{W}^{cs}_E = 0$$

Equations (235) through (243) simplify for the case of 3 + 1 spacetime for which $\bar{D}^{cs}_{to} = 1$, $\bar{D}^{cs}_{so} = 3$ and $\bar{D}^{cs}_{s\sigma} = 0$. The external potential is assumed to be independent of θ_V and T .

Case D. Coherent Energy and Coherent Spacetime.

The eigenvalue equations for this case are easily obtainable from equations (219) and (220) which give immediately for an ultrafast process in the superconducting state of a high-T_c superconductor

$$\bar{D}^{tc}_{so}[j\theta_V^2 d^2\theta_{Uo}/d\theta_V^2 - (\theta_V d\theta_{Uo}/d\theta_V)^2] \tag{244}$$

$$+ j\bar{D}^{tc}_{so}(1 + \bar{\gamma}^{tc}_o)\theta_V d\theta_{Uo}/d\theta_V + 1 - \bar{\mu}^{tc} + \bar{W}^{tc}_E = 0$$

$$(1 - \bar{\mu}^{tc})[1 + \sigma\bar{D}^{tc}_{to} + \sigma\bar{D}^{tc}_{to}\bar{\gamma}^{tc}_o\bar{P}^{tc}_o/(\bar{P}^{tc}_o - \bar{K}^{tc}_o)] \tag{245}$$

$$- \bar{D}^{tc}_{so}\theta_V d\bar{\gamma}^{tc}_o/d\theta_V - \bar{D}_{s\sigma}\bar{\beta}^{tc}_{Po} + \bar{W}^{tc}_E = 0$$

where to be completely general a potential term has been added, and where $\bar{P}^{tc}_o$, $\bar{K}^{tc}_o$, $\bar{\gamma}^{tc}_o$ and $\bar{\beta}^{tc}_o$ are given by equations (201), (214), (204) and (212) respectively. The external potential is taken to be independent of θ_V and T .

5. QUANTUM THEORY OF TIME AND DIMENSION. Renormalization group equations for time and dimension have been presented in Section 1. Now in this section a set of first order Dirac-like differential eigenvalue equations for time and dimension are developed which may describe the discrete and continuous states of time and dimension in bulk matter. These equations describe the time and dimension structure of energy and pressure. The quantized time (rate), space dimension, and time dimension equations determine the effects of real state equations on the quantized rates and quantized geometrical structure of chemical and nuclear

reaction processes that occur in bulk matter.

Case A. Partially Coherent Spacetime, Dimension and Energy.

This case corresponds to the normal state of a high-T_c compound. Following the same procedure as in Section 4 to quantize the energy trace equation, the source terms in the right hand sides of equations (78) through (81) are assumed to be self actuated fields and the corresponding eigenvalue equations are written along with the energy eigenvalue equation (224) as

$$(1 - \bar{\mu}')(\bar{E}' + \bar{D}_t'\bar{\beta}_E') - \bar{D}_s'\bar{\beta}_P' = 0 \tag{246}$$

$$(1 - \bar{\omega}')(\bar{t}' - \bar{D}_t'\bar{\beta}_E' \, \partial\bar{t}'/\partial\bar{E}') + \bar{D}_s'\bar{\beta}_P' \, \partial\bar{t}'/\partial\bar{P}' = 0 \tag{247}$$

$$(1 - \bar{\lambda}')(\bar{D}_t' + \bar{D}_t'\bar{\beta}_E' \, \partial\bar{D}_t'/\partial\bar{E}') - \bar{D}_s'\bar{\beta}_P' \, \partial\bar{D}_t'/\partial\bar{P}' = 0 \tag{248}$$

$$(1 - \bar{\delta}')(\bar{D}_s' + \bar{D}_t'\bar{\beta}_E' \, \partial\bar{D}_s'/\partial\bar{E}') - \bar{D}_s'\bar{\beta}_P' \, \partial\bar{D}_s'/\partial\bar{P}' = 0 \tag{249}$$

For a non-interacting system $\bar{\beta}_E' = 0$, $\bar{\beta}_P' = 0$, $\bar{\mu}' = 1$, $\bar{\omega}' = 1$, $\bar{\lambda}' = 1$ and $\bar{\delta}' = 1$. Equations (246) through (249) are four simultaneous eigenvalue equations for the internal energy density, time, time dimension and space dimension. Coupled eigenvalue equations occur in many physical situations and are generally difficult to solve.[20] Equations (246) through (249) can be generalized to include external potentials and are written in the form

$$(1 - \bar{\mu}')(\bar{E}' + \bar{D}_t'\bar{\beta}_E') - \bar{D}_s'\bar{\beta}_P' + \bar{W}_E'\bar{E}' = 0 \tag{250}$$

$$(1 - \bar{\omega}')(\bar{t}' - \bar{D}_t'\bar{\beta}_E' \, \partial\bar{t}'/\partial\bar{E}') + \bar{D}_s'\bar{\beta}_P' \, \partial\bar{t}'/\partial\bar{P}' + \bar{W}_t'\bar{t}' = 0 \tag{251}$$

$$(1 - \bar{\lambda}')(\bar{D}_t' + \bar{D}_t'\bar{\beta}_E' \, \partial\bar{D}_t'/\partial\bar{E}') - \bar{D}_s'\bar{\beta}_P' \, \partial\bar{D}_t'/\partial\bar{P}' + \bar{W}_{Dt}'\bar{D}_t' = 0 \tag{252}$$

$$(1 - \bar{\delta}')(\bar{D}_s' + \bar{D}_t'\bar{\beta}_E' \, \partial\bar{D}_s'/\partial\bar{E}') - \bar{D}_s'\bar{\beta}_P' \, \partial\bar{D}_s'/\partial\bar{P}' + \bar{W}_{Ds}'\bar{D}_s' = 0 \tag{253}$$

where $\bar{W}_E'$, $\bar{W}_t'$, $\bar{W}_{Dt}'$ and $\bar{W}_{Ds}'$ are dimensionless external potentials. The case of incoherent space can be regained from equations (246) through (253) by the substitution $W' \to W$ and dropping the primes. The quantized time equation can easily be cast into a quantized rate equation by taking the rate $\sim t^{-1}$ or more generally as rate = dN/dt.

The eigenvalue equations (250) through (253) determine the internal energy, time, time dimension and space dimension of matter. The eigenvalues $\bar{\mu}^{(\nu)}$, $\bar{\omega}^{(\nu)}$, $\bar{\lambda}^{(\nu)}$ and $\bar{\delta}^{(\nu)}$ and eigenfunctions $\bar{E}^{(\nu)}$, $\bar{t}^{(\nu)}$, $\bar{D}_t^{(\nu)}$ and $\bar{D}_s^{(\nu)}$ are associated with a parameter ν that can be discrete or continuous. Thus the internal energy, time, time dimension and space dimension can have a discrete (line) or continuous spectrum. It is possible to have various combinations of discrete and continuous values for these quantities. For instance at ordinary temperatures and densities $D_t' \sim 1$ and $D_s' \sim 3$, while time and energy are not constants and appear as continuous functions at macroscopic dimensions. This may not be the case at high temperatures and densities where time and bulk matter energy may have only a limited range of values (Section 6). At these high temperatures and densities

D_t' and D_s' may be continuous functions of temperature and density.[16] The Dirac and Schrödinger equations describe microscopic systems such as molecules, atoms, atomic nuclei and the fundamental particles. The first order differential equations (250) through (253) are the bulk matter analogs of the Dirac equation.

The eigenvalue equations (250) through (253) can be rewritten in terms of temperature and particle number density as independent variables by writing[16]

$$\partial\bar{t}'/\partial\bar{P}' = \bar{e}'\partial\bar{t}'/\partial n' - \bar{f}'\partial\bar{t}'/\partial T \qquad \partial\bar{t}'/\partial\bar{E}' = \bar{h}'\partial\bar{t}'/\partial T - \bar{g}'\partial\bar{t}'/\partial n' \tag{254}$$

$$\partial\bar{D}_t'/\partial\bar{P}' = \bar{e}'\partial\bar{D}_t'/\partial n' - \bar{f}'\partial\bar{D}_t'/\partial T \qquad \partial\bar{D}_t'/\partial\bar{E}' = \bar{h}'\partial\bar{D}_t'/\partial T - \bar{g}'\partial\bar{D}_t'/\partial n' \tag{255}$$

$$\partial\bar{D}_s'/\partial\bar{P}' = \bar{e}'\partial\bar{D}_s'/\partial n' - \bar{f}'\partial\bar{D}_s'/\partial T \qquad \partial\bar{D}_s'/\partial\bar{E}' = \bar{h}'\partial\bar{D}_s'/\partial T - \bar{g}'\partial\bar{D}_s'/\partial n' \tag{256}$$

where

$$\bar{e}' = 1/\bar{D}_e'\ \partial\bar{E}'/\partial T \qquad \bar{f}' = 1/\bar{D}_e'\ \partial\bar{E}'/\partial n' \tag{257}$$

$$\bar{g}' = 1/\bar{D}_e'\ \partial\bar{P}'/\partial T \qquad \bar{h}' = 1/\bar{D}_e'\ \partial\bar{P}'/\partial n' \tag{258}$$

$$\bar{D}_e' = \partial\bar{P}'/\partial n'\ \partial\bar{E}'/\partial T - \partial\bar{P}'/\partial T\ \partial\bar{E}'/\partial n' \tag{259}$$

and where

$$\bar{q}_2' = \bar{h}'\bar{D}_t'\bar{\beta}_{\bar{E}}' + \bar{f}'\bar{D}_s'\bar{\beta}_P' \qquad \bar{q}_D' = \bar{h}'\bar{D}_t'\bar{\beta}_{\bar{E}}' \tag{260}$$

$$\bar{s}_2' = \bar{g}'\bar{D}_t'\bar{\beta}_{\bar{E}}' + \bar{e}'\bar{D}_s'\bar{\beta}_P' \qquad \bar{s}_D' = \bar{g}'\bar{D}_t'\bar{\beta}_{\bar{E}}' \tag{261}$$

Then the eigenvalue equations (250) through (253) can be written as

$$(1 - \bar{\mu}')(\bar{E}' + \bar{D}_t'\bar{\beta}_{\bar{E}}') - \bar{D}_s'\bar{\beta}_P' + \bar{W}_{\bar{E}}'\bar{E}' = 0 \tag{262}$$

$$(1 - \bar{\omega}')\bar{t}' - (\bar{q}_2' - \bar{\omega}'\bar{q}_D')\partial\bar{t}'/\partial T + (\bar{s}_2' - \bar{\omega}'\bar{s}_D')\partial\bar{t}'/\partial n' + \bar{W}_t'\bar{t}' = 0 \tag{263}$$

$$(1 - \bar{\lambda}')\bar{D}_t' + (\bar{q}_2' - \bar{\lambda}'\bar{q}_D')\partial\bar{D}_t'/\partial T - (\bar{s}_2' - \bar{\lambda}'\bar{s}_D')\partial\bar{D}_t'/\partial n' + \bar{W}_{Dt}'\bar{D}_t' = 0 \tag{264}$$

$$(1 - \bar{\delta}')\bar{D}_s' + (\bar{q}_2' - \bar{\delta}'\bar{q}_D')\partial\bar{D}_s'/\partial T - (\bar{s}_2' - \bar{\delta}'\bar{s}_D')\partial\bar{D}_s'/\partial n' + \bar{W}_{Ds}'\bar{D}_s' = 0 \tag{265}$$

In these equations $n' = N/V'$ where V' is given by equation (74).

Case B. Coherent Spacetime and Dimension and Partially Coherent Energy.

This is the case of a thermodynamic process in the superconducting state of a high-T_c compound. The eigenvalue equations for partially coherent energy and coherent spacetime and coherent dimensions can be obtained from equations (86) through (89) with added external potential terms to be

$$(1 - \bar{\mu}^{cs})(\bar{E}^{cs} + \bar{D}_t^{cs}\bar{\beta}_{\bar{E}}^{cs}) - \bar{D}_s^{cs}\bar{\beta}_P^{cs} + \bar{W}_{\bar{E}}\bar{E}^{cs} = 0 \tag{266}$$

$$(1 - \bar{\omega}^{cs})(1 - j\bar{D}_t^{cs}\bar{\beta}_{\bar{E}}^{cs}\ \partial\theta_t/\partial\bar{E}^{cs}) + j\bar{D}_s^{cs}\bar{\beta}^{cs}\ \partial\theta_t/\partial\bar{P}^{cs} + \bar{W}_t^{cs} = 0 \tag{267}$$

$$(1 - \bar{\lambda}^{cs})(1 + j\bar{D}_t^{cs}\bar{\beta}_E^{cs}\,\partial\theta_{Dt}/\partial\bar{E}^{cs}) - j\bar{D}_s^{cs}\bar{\beta}_P^{cs}\,\partial\theta_{Dt}/\partial\bar{P}^{cs} + \bar{W}_{Dt}^{cs} = 0 \qquad (268)$$

$$(1 - \bar{\delta}^{cs})(1 + j\bar{D}_t^{cs}\bar{\beta}_E^{cs}\,\partial\theta_{Ds}/\partial\bar{E}^{cs}) - j\bar{D}_s^{cs}\bar{\beta}_P^{cs}\,\partial\theta_{Ds}/\partial\bar{P}^{cs} + \bar{W}_{Ds}^{cs} = 0 \qquad (269)$$

where the magnitudes t^{cs}, D_t^{cs} and D_s^{cs} = constants, and where $\bar{E}^{cs}$, $\bar{\beta}_E^{cs}$ and $\bar{\beta}_P^{cs}$ are given by equations (83) through (85) respectively.

By introducing the parameter $\xi = N/(V\theta_V)$ = particle number density for coherent space, and using the relationships

$$\partial\theta_t/\partial\bar{P}^{cs} = \bar{e}^{cs}\partial\theta_t/\partial\xi - \bar{f}^{cs}\partial\theta_t/\partial T \qquad \partial\theta_t/\partial\bar{E}^{cs} = \bar{h}^{cs}\partial\theta_t/\partial T - \bar{g}^{cs}\partial\theta_t/\partial\xi \qquad (270)$$

$$\partial\theta_{Dt}/\partial\bar{P}^{cs} = \bar{e}^{cs}\partial\theta_{Dt}/\partial\xi - \bar{f}^{cs}\partial\theta_{Dt}/\partial T \qquad \partial\theta_{Dt}/\partial\bar{E}^{cs} = \bar{h}^{cs}\partial\theta_{Dt}/\partial T - \bar{g}^{cs}\partial\theta_{Dt}/\partial\xi \qquad (271)$$

$$\partial\theta_{Ds}/\partial\bar{P}^{cs} = \bar{e}^{cs}\partial\theta_{Ds}/\partial\xi - \bar{f}^{cs}\partial\theta_{Ds}/\partial T \qquad \partial\theta_{Ds}/\partial\bar{E}^{cs} = \bar{h}^{cs}\partial\theta_{Ds}/\partial T - \bar{g}^{cs}\partial\theta_{Ds}/\partial\xi \qquad (272)$$

where

$$\bar{e}^{cs} = 1/\bar{D}_e^{cs}\;\partial\bar{E}^{cs}/\partial T \qquad \bar{f}^{cs} = 1/\bar{D}_e^{cs}\;\partial\bar{E}^{cs}/\partial\xi \qquad (273)$$

$$\bar{g}^{cs} = 1/\bar{D}_e^{cs}\;\partial\bar{P}^{cs}/\partial T \qquad \bar{h}^{cs} = 1/\bar{D}_e^{cs}\;\partial\bar{P}^{cs}/\partial\xi \qquad (274)$$

$$\bar{D}_e^{cs} = \partial\bar{P}^{cs}/\partial\xi\;\partial\bar{E}^{cs}/\partial T - \partial\bar{P}^{cs}/\partial T\;\partial\bar{E}^{cs}/\partial\xi \qquad (275)$$

and introducing

$$\bar{q}_2^{cs} = \bar{h}^{cs}\bar{D}_t^{cs}\bar{\beta}_E^{cs} + \bar{f}^{cs}\bar{D}_s^{cs}\bar{\beta}_P^{cs} \qquad \bar{q}_D^{cs} = \bar{h}^{cs}\bar{D}_t^{cs}\bar{\beta}_E^{cs} \qquad (276)$$

$$\bar{s}_2^{cs} = \bar{g}^{cs}\bar{D}_t^{cs}\bar{\beta}_E^{cs} + \bar{e}^{cs}\bar{D}_s^{cs}\bar{\beta}_P^{cs} \qquad \bar{s}_D^{cs} = \bar{g}^{cs}\bar{D}_t^{cs}\bar{\beta}_E^{cs} \qquad (277)$$

allows the eigenvalue equations (266) through (269) to be rewritten as

$$(1 - \bar{\mu}^{cs})(\bar{E}^{cs} + \bar{D}_t^{cs}\bar{\beta}_E^{cs}) - \bar{D}_s^{cs}\bar{\beta}_P^{cs} + \bar{W}_E^{cs}\bar{E}^{cs} = 0 \qquad (278)$$

$$1 - \bar{\omega}^{cs} - j(\bar{q}_2^{cs} - \bar{\omega}^{cs}\bar{q}_D^{cs})\partial\theta_t/\partial T + j(\bar{s}_2^{cs} - \bar{\omega}^{cs}\bar{s}_D^{cs})\partial\theta_t/\partial\xi + \bar{W}_t^{cs} = 0 \qquad (279)$$

$$1 - \bar{\lambda}^{cs} + j(\bar{q}_2^{cs} - \bar{\lambda}^{cs}\bar{q}_D^{cs})\partial\theta_{Dt}/\partial T - j(\bar{s}_2^{cs} - \bar{\lambda}^{cs}\bar{s}_D^{cs})\partial\theta_{Dt}/\partial\xi + \bar{W}_{Dt}^{cs} = 0 \qquad (280)$$

$$1 - \bar{\delta}^{cs} + j(\bar{q}_2^{cs} - \bar{\delta}^{cs}\bar{q}_D^{cs})\partial\theta_{Ds}/\partial T - j(\bar{s}_2^{cs} - \bar{\delta}^{cs}\bar{s}_D^{cs})\partial\theta_{Ds}/\partial\xi + \bar{W}_{Ds}^{cs} = 0 \qquad (281)$$

Instead of the variable $\xi = n/(V\theta_V)$ it is possible to use the independent variable θ_V directly by making the replacement $\xi \rightarrow \theta_V$ in equations (270) through (281).

Case C. Coherent Spacetime, Dimensions and Energy.

This corresponds to an ultrafast thermodynamic process in the supercon-

ducting state of a high-T_c compound. For coherent energy as well as coherent spacetime and coherent dimensions, the appropriate eigenvalue equations are obtained from equations (97) through (100) with added external potentials as follows

$$(1 - \bar{\mu}^{tc})(\bar{E}^{tc} + \bar{D}_t^{tc}\bar{\beta}_E^{tc}) - \bar{D}_s^{tc}\bar{\beta}_P^{tc} + \bar{W}_E\bar{E}^{tc} = 0 \tag{282}$$

$$(1 - \bar{\omega}^{tc})(1 - j\bar{D}_t^{tc}\bar{\beta}_E^{tc}\,\partial\theta_t/\partial\bar{E}^{tc}) + j\bar{D}_s^{tc}\bar{\beta}_P^{tc}\,\partial\theta_t/\partial\bar{P}^{tc} + \bar{W}_t^{tc} = 0 \tag{283}$$

$$(1 - \bar{\lambda}^{tc})(1 + j\bar{D}_t^{tc}\bar{\beta}_E^{tc}\,\partial\theta_{Dt}/\partial\bar{E}^{tc}) - j\bar{D}_s^{tc}\bar{\beta}_P^{tc}\,\partial\theta_{Dt}/\partial\bar{P}^{tc} + \bar{W}_{Dt}^{tc} = 0 \tag{284}$$

$$(1 - \bar{\delta}^{tc})(1 + j\bar{D}_t^{tc}\bar{\beta}_E^{tc}\,\partial\theta_{Ds}/\partial\bar{E}^{tc}) - j\bar{D}_s^{tc}\bar{\beta}_P^{tc}\,\partial\theta_{Ds}/\partial\bar{P}^{tc} + \bar{W}_{Ds}^{tc} = 0 \tag{285}$$

where the magnitudes U^{tc}, t^{tc}, D_t^{tc}, D_s^{tc} and V = constants, and where $\bar{E}^{tc}$, $\bar{\beta}_E^{tc}$ and $\bar{\beta}^{tc}$ are given by equations (92), (101) and (102) respectively. Note that $\bar{E}^{tc}$ and $\bar{P}^{tc}$ are not coherent quantities as can be seen from equation (93).

From the chain rule for derivatives it is possible to replace the independent variables $\bar{E}^{tc}$ and $\bar{P}^{tc}$ by the variables θ_V and T as follows

$$\partial\theta_t/\partial T = \partial\theta_t/\partial\bar{P}^{tc}\,\partial\bar{P}^{tc}/\partial T + \partial\theta_t/\partial\bar{E}^{tc}\,\partial\bar{E}^{tc}/\partial T \tag{286}$$

$$\partial\theta_t/\partial\theta_V = \partial\theta_t/\partial\bar{P}^{tc}\,\partial\bar{P}^{tc}/\partial\theta_V + \partial\theta_t/\partial\bar{E}^{tc}\,\partial\bar{E}^{tc}/\partial\theta_V \tag{287}$$

with similar equations for the derivatives of θ_{Dt} and θ_{Ds}. Then the following relations can be derived

$$\partial\theta_t/\partial\bar{P}^{tc} = \bar{e}^{tc}\partial\theta_t/\partial\theta_V - \bar{f}^{tc}\partial\theta_t/\partial T \qquad \partial\theta_t/\partial\bar{E}^{tc} = \bar{h}^{tc}\partial\theta_t/\partial T - \bar{g}^{tc}\partial\theta_t/\partial\theta_V \tag{288}$$

$$\partial\theta_{Dt}/\partial\bar{P}^{tc} = \bar{e}^{tc}\partial\theta_{Dt}/\partial\theta_V - \bar{f}^{tc}\partial\theta_{Dt}/\partial T \qquad \partial\theta_{Dt}/\partial\bar{E}^{tc} = \bar{h}^{tc}\partial\theta_{Dt}/\partial T - \bar{g}^{tc}\partial\theta_{Dt}/\partial\theta_V \tag{289}$$

$$\partial\theta_{Ds}/\partial\bar{P}^{tc} = \bar{e}^{tc}\partial\theta_{Ds}/\partial\theta_V - \bar{f}^{tc}\partial\theta_{Ds}/\partial T \qquad \partial\theta_{Ds}/\partial\bar{E}^{tc} = \bar{h}^{tc}\partial\theta_{Ds}/\partial T - \bar{g}^{tc}\partial\theta_{Ds}/\partial\theta_V \tag{290}$$

where

$$\bar{e}^{tc} = 1/\bar{D}_e^{tc}\,\partial\bar{E}^{tc}/\partial T \qquad \bar{f}^{tc} = 1/\bar{D}_e^{tc}\,\partial\bar{E}^{tc}/\partial\theta_V \tag{291}$$

$$\bar{g}^{tc} = 1/\bar{D}_e^{tc}\,\partial\bar{P}^{tc}/\partial T \qquad \bar{h}^{tc} = 1/\bar{D}_e^{tc}\,\partial\bar{P}^{tc}/\partial\theta_V \tag{292}$$

$$\bar{D}_e^{tc} = \partial\bar{P}^{tc}/\partial\theta_V\,\partial\bar{E}^{tc}/\partial T - \partial\bar{P}^{tc}/\partial T\,\partial\bar{E}^{tc}/\partial\theta_V \tag{293}$$

Equation (93) gives the following derivatives of the energy density

$$\partial\bar{E}^{tc}/\partial T = j\bar{E}^{tc}\partial\theta_U/\partial T \qquad \partial\bar{E}^{tc}/\partial\theta_V = j\bar{E}^{tc}\partial\theta_U/\partial\theta_V - \bar{E}^{tc}/\theta_V \tag{294A}$$

and the following derivatives of the pressure

$$T\partial\bar{P}^{tc}/\partial T = j\bar{E}^{tc}\theta_V\partial\theta_U/\partial\theta_V + \bar{P}^{tc} \tag{294B}$$

Introducing the quantities

$$\bar{q}_2^{tc} = \bar{h}^{tc}\bar{D}_t^{tc}\bar{\beta}_E^{tc} + \bar{f}^{tc}\bar{D}_s^{tc}\bar{\beta}_P^{tc} \qquad \bar{q}_D^{tc} = \bar{h}^{tc}\bar{D}_t^{tc}\bar{\beta}_E^{tc} \tag{295}$$

$$\bar{s}_2^{tc} = \bar{g}^{tc}\bar{D}_t^{tc}\bar{\beta}_E^{tc} + \bar{e}^{tc}\bar{D}_s^{tc}\bar{\beta}_P^{tc} \qquad \bar{s}_D^{tc} = \bar{g}^{tc}\bar{D}_t^{tc}\bar{\beta}_E^{tc} \tag{296}$$

lets the eigenvalue equations (282) through (285) be written as

$$(1 - \bar{\mu}^{tc})(\bar{E}^{tc} + \bar{D}_t^{tc}\bar{\beta}_E^{tc}) - \bar{D}_s^{tc}\bar{\beta}_P^{tc} + \bar{W}_E^{tc}\bar{E}^{tc} = 0 \tag{297}$$

$$1 - \bar{\omega}^{tc} - j(\bar{q}_2^{tc} - \bar{\omega}^{tc}\bar{q}_D^{tc})\partial\theta_t/\partial T + j(\bar{s}_2^{tc} - \bar{\omega}^{tc}\bar{s}_D^{tc})\partial\theta_t/\partial\theta_V + \bar{W}_t^{tc} = 0 \tag{298}$$

$$1 - \bar{\lambda}^{tc} + j(\bar{q}_2^{tc} - \bar{\lambda}^{tc}\bar{q}_D^{tc})\partial\theta_{Dt}/\partial T - j(\bar{s}_2^{tc} - \bar{\lambda}^{tc}\bar{s}_D^{tc})\partial\theta_{Dt}/\partial\theta_V + \bar{W}_{Dt}^{tc} = 0 \tag{299}$$

$$1 - \bar{\delta}^{tc} + j(\bar{q}_2^{tc} - \bar{\delta}^{tc}\bar{q}_D^{tc})\partial\theta_{Ds}/\partial T - j(\bar{s}_2^{tc} - \bar{\delta}^{tc}\bar{s}_D^{tc})\partial\theta_{Ds}/\partial\theta_V + \bar{W}_{Ds}^{tc} = 0 \tag{300}$$

Equations (282) through (285) or equations (297) through (300) are the macroscopic quantum eigenvalue equations for coherent bulk matter in coherent spacetime.

Equations (282) through (285) or (297) through (300) are coupled nonlinear complex number eigenvalue equations which are in general difficult to solve. A simpler set of equations can be obtained by taking

$$j\bar{D}_t^{tc}\bar{\beta}_E^{tc}/\bar{E}^{tc} = \bar{c}_1 = \text{constant} \tag{301}$$

$$j\bar{D}_s^{tc}\bar{\beta}_P^{tc}/\bar{P}^{tc} = \bar{c}_2 = \text{constant} \tag{302}$$

which allows equations (282) through (285) to be written as

$$(1 - \bar{\mu}^{tc})(1 - j\bar{c}_1)\bar{E}^{tc} + j\bar{c}_2\bar{P}^{tc} + \bar{W}_E^{tc}\bar{E}^{tc} = 0 \tag{303}$$

$$(1 - \bar{\omega}^{tc})(1 - \bar{c}_1\bar{E}^{tc}\partial\theta_t/\partial\bar{E}^{tc}) + \bar{c}_2\bar{P}^{tc}\partial\theta_t/\partial\bar{P}^{tc} + \bar{W}_t^{tc} = 0 \tag{304}$$

$$(1 - \bar{\lambda}^{tc})(1 + \bar{c}_1\bar{E}^{tc}\partial\theta_{Dt}/\partial\bar{E}^{tc}) - \bar{c}_2\bar{P}^{tc}\partial\theta_{Dt}\partial\bar{P}^{tc} + \bar{W}_{Dt}^{tc} = 0 \tag{305}$$

$$(1 - \bar{\delta}^{tc})(1 + \bar{c}_1\bar{E}^{tc}\partial\theta_{Ds}/\partial\bar{E}^{tc}) - \bar{c}_2\bar{P}^{tc}\partial\theta_{Ds}\partial\bar{P}^{tc} + \bar{W}_{Ds}^{tc} = 0 \tag{306}$$

where now equations (304) through (306) are linear differential equations in θ_t, θ_{Dt} and θ_{Ds}. Alternatively it is possible to start with equations (297) through (300) and write

$$(1 - \bar{\mu}^{tc})(\bar{E}^{tc} + \bar{D}_t^{tc}\bar{\beta}_E^{tc}) - \bar{D}_s^{tc}\bar{\beta}_P^{tc} + \bar{W}^{tc}\bar{E}^{tc} = 0 \tag{307}$$

$$1 - \bar{\omega}^{tc} - j\bar{c}_3\partial\theta_t/\partial T + j\bar{c}_4\partial\theta_t/\partial\theta_V + \bar{W}_t^{tc} = 0 \tag{308}$$

$$1 - \bar{\lambda}^{tc} + j\bar{c}_5 \partial\theta_{Dt}/\partial T - j\bar{c}_6 \partial\theta_{Dt}/\partial\theta_V + \bar{W}_{Dt}^{tc} = 0 \tag{309}$$

$$1 - \bar{\delta}^{tc} + j\bar{c}_7 \partial\theta_{Ds}/\partial T - j\bar{c}_8 \partial\theta_{Ds}/\partial\theta_V + \bar{W}_{Ds}^{tc} = 0 \tag{310}$$

where the constants $\bar{c}_3$ through $\bar{c}_8$ are given by

$$\bar{c}_3 = \bar{q}_2^{tc} - \bar{\omega}^{tc}\bar{q}_D^{tc} \qquad \bar{c}_4 = \bar{s}_2^{tc} - \bar{\omega}^{tc}\bar{s}_D^{tc} \tag{311}$$

$$\bar{c}_5 = \bar{q}_2^{tc} - \bar{\lambda}^{tc}\bar{q}_D^{tc} \qquad \bar{c}_6 = \bar{s}_2^{tc} - \bar{\lambda}^{tc}\bar{s}_D^{tc} \tag{312}$$

$$\bar{c}_7 = \bar{q}_2^{tc} - \bar{\delta}^{tc}\bar{q}_D^{tc} \qquad \bar{c}_8 = \bar{s}_2^{tc} - \bar{\delta}^{tc}\bar{s}_D^{tc} \tag{313}$$

where $\bar{q}_2^{tc}$, $\bar{q}_D^{tc}$, $\bar{s}_2^{tc}$ and $\bar{s}_D^{tc}$ are given by equations (295) and (296) respectively. No simplification of the trace equation (307) is obtained by using this alternative procedure, but equations (308) through (310) are now linear differential equations.

6. SOLUTION OF THE TIME AND DIMENSION EIGENVALUE EQUATIONS. It is difficult to obtain a general solution for the set of energy, time, time dimension and space dimension eigenvalue equations (250) through (253) because they are coupled nonlinear differential eigenvalue equations.[20] Therefore the original set of equations are not solved in this paper. Instead, the coupled equations are decoupled by taking constant values $\bar{D}_t' = \bar{D}_{tk}'$ and $\bar{D}_s' = \bar{D}_{sk}'$ for the time and space dimensions when they appear as coefficients in equations (250) through (253) as follows

$$(1 - \bar{\mu}')(\bar{E}' + \bar{D}_{tk}'\bar{\beta}_E') - \bar{D}_{sk}'\bar{\beta}_P' + \bar{W}_E'\bar{E}' = 0 \tag{314}$$

$$(1 - \bar{\omega}')(\bar{t}' - \bar{D}_{tk}'\bar{\beta}_E' \partial\bar{t}'/\partial\bar{E}') + \bar{D}_{sk}'\bar{\beta}_P' \partial\bar{t}'/\partial\bar{P}' + \bar{W}_t'\bar{t}' = 0 \tag{315}$$

$$(1 - \bar{\lambda}')(\bar{D}_t' + \bar{D}_{tk}'\bar{\beta}_E' \partial\bar{D}_t'/\partial\bar{E}') - \bar{D}_{sk}'\bar{\beta}_P' \partial\bar{D}_t'/\partial\bar{P}' + \bar{W}_{Dt}'\bar{D}_t' = 0 \tag{316}$$

$$(1 - \bar{\delta}')(\bar{D}_s' + \bar{D}_{tk}'\bar{\beta}_E' \partial\bar{D}_s'/\partial\bar{E}') - \bar{D}_{sk}'\bar{\beta}_P' \partial\bar{D}_s'/\partial\bar{P}' + \bar{W}_{Ds}'\bar{D}_s' = 0 \tag{317}$$

where $\bar{E}' = \bar{U}'/V'$ and V' is given by equation (74). Within this approximation equations (315) through (317) are linear differential equations, however, equation (314) is still fundamentally nonlinear. The solutions to the approximate equations (314) through (317) are reasonable only for values $\bar{D}_t' \sim \bar{D}_{tk}'$ and $\bar{D}_s' \sim \bar{D}_{sk}'$, and any significant departure of $\bar{D}_t'$ and $\bar{D}_s'$ from these values requires the solution of the nonlinear coupled equations (250) through (253). For a noninteracting system $\bar{\beta}_E' = 0$ and $\bar{\beta}_P' = 0$ and

$$\bar{\mu}' = 1 + \bar{W}_E' \qquad \bar{\omega}' = 1 + \bar{W}_t' \qquad \bar{\lambda}' = 1 + \bar{W}_{Dt}' \qquad \bar{\delta}' = 1 + \bar{W}_{Ds}' \tag{317A}$$

The solution to the energy eigenvalue equation (314) was considered in Section 4.

By the technique of separation of variables the solutions to the decoupled equations (315) through (317) with $\bar{W}_t = 0$, $\bar{W}_{Dt} = 0$ and $\bar{W}_{Ds} = 0$ are respectively

$$\bar{t}' = \bar{t}^b \exp(\bar{F}' \int d\bar{E}'/\bar{\beta}'_{\bar{E}} + \bar{G}' \int d\bar{P}'/\bar{\beta}'_{\bar{P}}) \tag{317B}$$

$$\bar{D}'_t = \bar{D}^b_t \exp(-\bar{H}' \int d\bar{E}'/\bar{\beta}'_{\bar{E}} - \bar{I}' \int d\bar{P}'/\bar{\beta}'_{\bar{P}}) \tag{317C}$$

$$\bar{D}'_s = \bar{D}^b_s \exp(-\bar{J}' \int d\bar{E}'/\bar{\beta}'_{\bar{E}} - \bar{K}' \int d\bar{P}'/\bar{\beta}'_{\bar{P}}) \tag{317D}$$

where

$$\bar{F}' = \bar{\kappa}'/[(1 - \bar{\omega}')\bar{D}'_{tk}] \qquad \bar{G}' = \bar{\sigma}'/\bar{D}'_{sk} \tag{317E}$$

$$\bar{H}' = \bar{\varepsilon}'/[(1 - \bar{\lambda}')\bar{D}'_{tk}] \qquad \bar{I}' = \bar{\nu}'/\bar{D}'_{sk} \tag{317F}$$

$$\bar{J}' = \bar{\tau}'/[(1 - \bar{\delta}')\bar{D}'_{tk}] \qquad \bar{K}' = \bar{\eta}'/\bar{D}'_{sk} \tag{317G}$$

where $\bar{t}^b$, $\bar{D}^b_t$ and $\bar{D}^b_s$ = constants, and where the constants $\bar{\kappa}'$, $\bar{\varepsilon}'$, $\bar{\tau}'$, $\bar{\sigma}'$, $\bar{\nu}'$ and $\bar{\eta}'$ are related by

$$\bar{\kappa}' + \bar{\omega}' - 1 = \bar{\sigma}' \tag{317H}$$

$$\bar{\varepsilon}' + \bar{\lambda}' - 1 = \bar{\nu}' \tag{317I}$$

$$\bar{\tau}' + \bar{\delta}' - 1 = \bar{\eta}' \tag{317J}$$

The relations in equations (317H) through (317J) follow from the technique of the separation of variables which, for the time equation (315), involves writing the solution as $\bar{t}' = \bar{\psi}(\bar{E}')\bar{\phi}(\bar{P}')$ and getting

$$\bar{D}'_{sk}\bar{\beta}'_{\bar{P}}(1/\bar{\phi}d\bar{\phi}/d\bar{P}') = \bar{\sigma}' \tag{318}$$

$$(1 - \bar{\omega}')\bar{D}'_{tk}\bar{\beta}'_{\bar{E}}(1/\bar{\psi}d\bar{\psi}/d\bar{E}') = \bar{\sigma}' + 1 - \bar{\omega}' = \bar{\kappa}' \tag{319}$$

which obviously yields equation (317B). Note that equations (317H) through (317J) are equivalent to

$$\kappa'_R + \omega'_R - 1 = \sigma'_R \qquad \kappa'_I + \omega'_I = \sigma'_I \tag{320}$$

$$\varepsilon'_R + \lambda'_R - 1 = \nu'_R \qquad \varepsilon'_I + \lambda'_I = \nu'_I \tag{321}$$

$$\tau'_R + \delta'_R - 1 = \eta'_R \qquad \tau'_I + \delta'_I = \eta'_I \tag{322}$$

Case A. Incoherent Energy and Partial Coherence of Spacetime and Dimension.

This case refers to a slow thermodynamic process in the normal state of a high-T_c superconductor. For this case the energy density and constant dimension coefficients are taken as real numbers and equations (314) through (317) are written as

$$(1 - \mu')(\bar{E}' + D'_{tk}\beta'_{\bar{E}}) - D'_{sk}\beta'_{\bar{P}} + W'_{\bar{E}}\bar{E}' = 0 \tag{323}$$

$$(1 - \bar{\omega}')(\bar{t}' - D'_{tk}\beta'_{\bar{E}}\,\partial\bar{t}'/\partial\bar{E}') + D'_{sk}\beta'_{\bar{P}}\,\partial\bar{t}'/\partial\bar{P}' + \bar{W}'_t\bar{t}' = 0 \tag{324}$$

$$(1 - \bar{\lambda}')(\bar{D}'_t + D'_{tk}\beta'_E\,\partial\bar{D}'_t/\partial E') - D'_{sk}\beta'_P\,\partial\bar{D}'_t/\partial P' + \bar{W}'_{Dt}\bar{D}'_t = 0 \quad (325)$$

$$(1 - \bar{\delta}')(\bar{D}'_s + D'_{tk}\beta'_E\,\partial\bar{D}'_s/\partial E') - D'_{sk}\beta'_P\,\partial\bar{D}'_s/\partial P' + \bar{W}'_{Ds}\bar{D}'_s = 0 \quad (326)$$

The solutions to equations (323) through (326) for zero external potentials are

$$\bar{t}' = \bar{t}^b \exp(\bar{F}''\textstyle\int dE'/\beta'_E + \bar{G}''\int dP'/\beta'_P) \quad (327)$$

$$\bar{D}'_t = \bar{D}^b_t \exp(-\bar{H}''\textstyle\int dE'/\beta'_E - \bar{I}''\int dP'/\beta'_P) \quad (328)$$

$$\bar{D}'_s = \bar{D}^b_s \exp(-\bar{J}''\textstyle\int dE'/\beta'_E - \bar{K}''\int dP'/\beta'_P) \quad (329)$$

where $E' = U'/V'$, and where

$$\bar{F}'' = \bar{\kappa}'/[(1 - \bar{\omega}')D'_{tk}] \qquad \bar{G}'' = \bar{\sigma}'/D'_{sk} \quad (329A)$$

$$\bar{H}'' = \bar{\varepsilon}'/[(1 - \bar{\lambda}')D'_{tk}] \qquad \bar{I}'' = \bar{\nu}'/D'_{sk} \quad (329B)$$

$$\bar{J}'' = \bar{\tau}'/[(1 - \bar{\delta}')D'_{tk}] \qquad \bar{K}'' = \bar{\eta}'/D'_{sk} \quad (329C)$$

Case A also describes partially coherent energy.

The constants that appear in equations (317H) through (317J) can be written as

$$\bar{\kappa}' = \kappa'_R + j\kappa'_I \qquad \bar{\varepsilon}' = \varepsilon'_R + j\varepsilon'_I \qquad \bar{\tau}' = \tau'_R + j\tau'_I \quad (330)$$

$$\bar{\omega}' = \omega'_R + j\omega'_I \qquad \bar{\lambda}' = \lambda'_R + j\lambda'_I \qquad \bar{\delta}' = \delta'_R + j\delta'_I \quad (331)$$

$$\bar{\sigma}' = \sigma'_R + j\sigma'_I \qquad \bar{\nu}' = \nu'_R + j\nu'_I \qquad \bar{\eta}' = \eta'_R + j\eta'_I \quad (332)$$

Then from equations (329A) through (329C) the following relations are calculated

$$\bar{\kappa}'/(1-\bar{\omega}') = a' + jb' \qquad \bar{\varepsilon}'/(1-\bar{\lambda}') = c' + j\ell' \qquad \bar{\tau}'/(1-\bar{\delta}') = e' + jf' \quad (333)$$

where

$$a' = [\kappa'_R(1 - \omega'_R) - \kappa'_I\omega'_I]/[(1 - \omega'_R)^2 + \omega'^2_I] \quad (334)$$

$$b' = [\kappa'_I(1 - \omega'_R) + \kappa'_R\omega'_I]/[(1 - \omega'_R)^2 + \omega'^2_I] \quad (335)$$

$$c' = [\varepsilon'_R(1 - \lambda'_R) - \varepsilon'_I\lambda'_I]/[(1 - \lambda'_R)^2 + \lambda'^2_I] \quad (336)$$

$$\ell' = [\varepsilon'_I(1 - \lambda'_R) + \varepsilon'_R\lambda'_I]/[(1 - \lambda'_R)^2 + \lambda'^2_I] \quad (337)$$

$$e' = [\tau'_R(1 - \delta'_R) - \tau'_I\delta'_I]/[(1 - \delta'_R)^2 + \delta'^2_I] \quad (338)$$

$$f' = [\tau'_I(1 - \delta'_R) + \tau'_R\delta'_I]/[(1 - \delta'_R)^2 + \delta'^2_I] \quad (339)$$

With these relations, the real and imaginary parts of the solutions in equations (327) through (329) can be obtained.

The zero-potential solutions in equations (327) through (329) can then be written as

$$\bar{t}' = \bar{t}^{b} \exp(\Phi_t + j\theta_t) \qquad \bar{D}_t' = \bar{D}_t^{b} \exp(\Phi_{Dt} + j\theta_{Dt}) \qquad \bar{D}_s' = \bar{D}_s^{b} \exp(\Phi_{Ds} + j\theta_{Ds}) \quad (340)$$

where

$$\theta_t = b'/D_{tk}' \int dE'/\beta_E' + \sigma_I'/D_{sk}' \int dP'/\beta_P' \quad (341)$$

$$\Phi_t = a'/D_{tk}' \int dE'/\beta_E' + \sigma_R'/D_{sk}' \int dP'/\beta_P' \quad (342)$$

$$\theta_{Dt} = -\ell'/D_{tk}' \int dE'/\beta_E' - \nu_I'/D_{sk}' \int dP'/\beta_P' \quad (343$$

$$\Phi_{Dt} = -c'/D_{tk}' \int dE'/\beta_E' - \nu_R'/D_{sk}' \int dP'/\beta_P' \quad (344)$$

$$\theta_{Ds} = -f'/D_{tk}' \int dE'/\beta_E' - \eta_I'/D_{sk}' \int dP'/\beta_P' \quad (345)$$

$$\Phi_{Ds} = -e'/D_{tk}' \int dE'/\beta_E' - \eta_R'/D_{sk}' \int dP'/\beta_P' \quad (346)$$

For a limited region in energy-pressure space where β_E' and β_P' can be taken as approximately constant the solutions in equations (341) through (346) can be written as

$$\theta_t = b'E'/(D_{tk}'\beta_{Ek}') + \sigma_I'P'/(D_{sk}'\beta_{Pk}') \quad (347)$$

$$\Phi_t = a'E'/(D_{tk}'\beta_{Ek}') + \sigma_R'P'/(D_{sk}'\beta_{Pk}') \quad (348)$$

$$\theta_{Dt} = -\ell'E'/(D_{tk}'\beta_{Ek}') - \nu_I'P'/(D_{sk}'\beta_{Pk}') \quad (349)$$

$$\Phi_{Dt} = -c'E'/(D_{tk}'\beta_{Ek}') - \nu_R'P'/(D_{sk}'\beta_{Pk}') \quad (350)$$

$$\theta_{Ds} = -f'E'/(D_{tk}'\beta_{Ek}') - \eta_I'P'/(D_{sk}'\beta_{Pk}') \quad (351)$$

$$\Phi_{Ds} = -e'E'/(D_{tk}'\beta_{Ek}') - \eta_R'P'/(D_{sk}'\beta_{Pk}') \quad (352)$$

These equations are based on simplifying assumptions but they have heuristic value.

The solutions presented in equation (340) and within the approximations given in equations (347) through (352) are valid only within a limited range of density and temperature where β_E' and β_P' are constants. Within this approximation simple algebra shows that the real parts of the solution given in equation (340) can always be written as

$$t'_R = A \exp \Phi_t \sin[b'(E' + \nu_c)/(D'_{tk}\beta'_{Ek})]\sin[\sigma'_I(P' + \eta_c)/(D'_{sk}\beta'_{Pk})] \quad (353)$$

$$D'_{tR} = B \exp \Phi_{Dt} \sin[\ell'(E' + \xi_c)/(D'_{tk}\beta'_{Ek})]\sin[\nu'_I(P' + \omega_c)/(D'_{sk}\beta'_{Pk})] \quad (354)$$

$$D'_{sR} = C \exp \Phi_{Ds} \sin[f'(E' + \tau_c)/(D'_{tk}\beta'_{Ek})]\sin[\eta'_I(P' + \rho_c)/(D'_{sk}\beta'_{Pk})] \quad (355)$$

where ν_c, η_c, ξ_c, ω_c, τ_c and ρ_c = constants. These solutions suggest the possibility that time and dimension can exhibit structure within definite ranges of energy density and pressure. The solutions may have applications to quantum electronic devices.[21]

Consider now the possibility of box structures in energy-pressure space wherein time and dimension are localized in structures with rigid walls where the time and dimensions vanish. Suppose that time vanishes at two boundaries in energy space E'_1 and E'_2 and at two boundaries in pressure space P'_1 and P'_2. Similarly, the time dimension vanishes at two boundaries E'_3 and E'_4 in energy density space and at two boundaries P'_3 and P'_4 in pressure space. Finally, the space dimension vanishes at two boundaries in energy density space denoted by E'_5 and E'_6 and at two boundaries P'_5 and P'_6 in pressure space. Then it follows from equations (353) through (355) that

$$E'_1 + \nu_c = 0 \qquad E'_3 + \xi_c = 0 \qquad E'_5 + \tau_c = 0 \quad (356)$$

$$P'_1 + \eta_c = 0 \qquad P'_3 + \omega_c = 0 \qquad P'_5 + \rho_c = 0 \quad (357)$$

$$b'(E'_2 + \nu_c)/(D'_{tk}\beta'_{Ek}) = n\pi \qquad \ell'(E'_4 + \xi_c)/(D'_{tk}\beta'_{Ek}) = n\pi \qquad f'(E'_6 + \tau_c)/(D'_{tk}\beta'_{Ek}) = n\pi \quad (358)$$

$$\sigma'_I(P'_2 + \eta_c)/(D'_{sk}\beta'_{Pk}) = m\pi \qquad \nu'_I(P'_4 + \omega_c)/(D'_{sk}\beta'_{Pk}) = m\pi \qquad \eta'_I(P'_6 + \rho_c)/(D'_{sk}\beta'_{Pk}) = m\pi \quad (359)$$

Combining equations (356) through (359) gives the following eigenvalues

$$b'_n = n\pi D'_{tk}\beta'_{Ek}/(E'_2 - E'_1) \qquad \sigma'_{Im} = m\pi D'_{sk}\beta'_{Pk}/(P'_2 - P'_1) \quad (360)$$

$$\ell'_n = n\pi D'_{tk}\beta'_{Ek}/(E'_4 - E'_3) \qquad \nu'_{Im} = m\pi D'_{sk}\beta'_{Pk}/(P'_4 - P'_3) \quad (361)$$

$$f'_n = n\pi D'_{tk}\beta'_{Ek}/(E'_6 - E'_5) \qquad \eta'_{Im} = m\pi D'_{sk}\beta'_{Pk}/(P'_6 - P'_5) \quad (362)$$

The eigenfunctions for the time box, time dimension box and space dimension box can then be obtained from equations (353) through (355) and equations (360) through (362) to be

$$t'^{nm}_R = A_{nm} \exp \Phi_t \sin[n\pi(E' - E'_1)/(E'_2 - E'_1)]\sin[m\pi(P' - P'_1)/(P'_2 - P'_1)] \quad (363)$$

$$D'^{nm}_{tR} = B_{nm} \exp \Phi_{Dt} \sin[n\pi(E' - E'_3)/(E'_4 - E'_3)]\sin[m\pi(P' - P'_3)/(P'_4 - P'_3)] \quad (364)$$

$$D'^{nm}_{sR} = C_{nm} \exp \Phi_{Ds} \sin[n\pi(E' - E'_5)/(E'_6 - E'_5)]\sin[m\pi(P' - P'_5)/(P'_6 - P'_5)] \quad (365)$$

The limits of the energy-pressure box may coincide for time, time dimension and

space dimension in which case $E_2' = E_4' = E_6'$, $P_2' = P_4' = P_6'$, $E_1' = E_3' = E_5'$, and $P_1' = P_3' = P_5'$.

Within Case A of incoherent energy and partial coherence of spacetime and dimension there are two special types of solutions: type 1, a purely sinusoidal solution, and type 2, a real exponential solution. These two types of solutions will now be considered.

Type 1. Sinusoidal Solution.

For this type, $\Phi_t = 0$, $\Phi_{Dt} = 0$ and $\Phi_{Ds} = 0$ in equation (340), and therefore from equations (342), (344), (346), (317H) through (317J) and (334) through (339) it follows that

$$a' = 0 \qquad \sigma_R' = 0 \qquad \omega_R' = 1 \qquad \bar{\kappa}' = 0 \qquad \omega_I' = \sigma_I' \qquad b' = 0 \tag{366}$$

$$c' = 0 \qquad \nu_R' = 0 \qquad \lambda_R' = 1 \qquad \bar{\varepsilon}' = 0 \qquad \lambda_I' = \nu_I' \qquad \ell' = 0 \tag{367}$$

$$e' = 0 \qquad \eta_R' = 0 \qquad \delta_R' = 1 \qquad \bar{\tau}' = 0 \qquad \delta_I' = \eta_I' \qquad f' = 0 \tag{368}$$

For this type of solution $b' = 0$, $\ell' = 0$ and $f' = 0$ so that equations (340), (347), (349) and (351) give

$$\bar{t} = \bar{t}^b \exp(j\theta_t) \qquad \bar{D}_t = \bar{D}_t^b \exp(j\theta_{Dt}) \qquad \bar{D}_s = \bar{D}_s^b \exp(j\theta_{Ds}) \tag{369}$$

with

$$\theta_t = \sigma_I'/D_{sk}' \int dP'/\beta_P' \qquad \theta_{Dt} = -\nu_I'/D_{sk}' \int dP'/\beta_P' \qquad \theta_{Ds} = -\eta_I'/D_{sk}' \int dP'/\beta_P' \tag{370}$$

For constant values of β_P' the real parts of the solutions in equation (369) can always be written as

$$t_R = A \sin[\sigma_I'/(D_{sk}'\beta_{Pk}')(P' + \eta_c)] \tag{371}$$

$$D_{tR} = B \sin[\nu_I'/(D_{sk}'\beta_{Pk}')(P' + \omega_c)] \tag{372}$$

$$D_{sR} = C \sin[\eta_I'/(D_{sk}'\beta_{Pk}')(P' + \rho_c)] \tag{373}$$

For a time box, time dimension box and space dimension box located in pressure space and bounded respectively by (P_1, P_2), (P_3, P_4) and (P_5, P_6) it follows from equations (360) through (362) and (371) through (373) that eigenfunctions for a purely sinusoidal state are

$$t_R'^m = A_m \sin[m\pi(P' - P_1')/(P_2' - P_1')] \tag{374}$$

$$D_{tR}'^m = B_m \sin[m\pi(P' - P_3')/(P_4' - P_3')] \tag{375}$$

$$D_{sR}'^m = C_m \sin[m\pi(P' - P_5')/(P_6' - P_5')] \tag{376}$$

Type 2. Real Exponential Solutions.

For this type of solution, $\theta_t = 0$, $\theta_{Dt} = 0$ and $\theta_{Ds} = 0$ in equation (340). Then equations (341) through (346) give

$$b' = 0 \qquad \ell' = 0 \qquad f' = 0 \qquad \sigma_I' = 0 \qquad \nu_I' = 0 \qquad \eta_I' = 0 \tag{376A}$$

From equations (334) through (339) and equations (320) through (322) it follows that three possibilities exist within a Type 2 solution.

Possibility 1. $\theta_t = 0$, $\theta_{Dt} = 0$, $\theta_{Ds} = 0$ and the following conditions

$$\kappa_R' = 0 \qquad \omega_R' = 1 \qquad \sigma_R' = 0 \qquad \kappa_I' = -\omega_I' \qquad a' = -\kappa_I'/\omega_I' = 1 \tag{377}$$

$$\varepsilon_R' = 0 \qquad \lambda_R' = 1 \qquad \nu_R' = 0 \qquad \varepsilon_I' = -\lambda_I' \qquad c' = -\varepsilon_I'/\lambda_I' = 1 \tag{378}$$

$$\tau_R' = 0 \qquad \delta_R' = 1 \qquad \eta_R' = 0 \qquad \tau_I' = -\delta_I' \qquad e' = -\tau_I'/\delta_I' = 1 \tag{379}$$

Then equations (342), (344) and (346) give

$$\Phi_t = 1/D_{tk}' \int dE'/\beta_E' \tag{380}$$

$$\Phi_{Dt} = -1/D_{tk}' \int dE'/\beta_E' \tag{381}$$

$$\Phi_{Ds} = -1/D_{tk}' \int dE'/\beta_E' \tag{382}$$

Possibility 2. $\theta_t = 0$, $\theta_{Dt} = 0$, $\theta_{Ds} = 0$ and the following conditions

$$\kappa_R' = 0 \qquad \omega_I' = 0 \qquad \sigma_R' = \omega_R' - 1 \qquad \kappa_I' = 0 \qquad a' = 0 \tag{383}$$

$$\varepsilon_R' = 0 \qquad \lambda_I' = 0 \qquad \nu_R' = \lambda_R' - 1 \qquad \varepsilon_I' = 0 \qquad c' = 0 \tag{384}$$

$$\tau_R' = 0 \qquad \delta_I' = 0 \qquad \eta_R' = \delta_R' - 1 \qquad \tau_I' = 0 \qquad e' = 0 \tag{385}$$

Then from equations (342), (344) and (346) it follows that

$$\Phi_t = (\omega_R' - 1)/D_{sk}' \int dP'/\beta_P' \tag{386}$$

$$\Phi_{Dt} = -(\lambda_R' - 1)/D_{sk}' \int dP'/\beta_P' \tag{387}$$

$$\Phi_{Ds} = -(\delta_R' - 1)/D_{sk}' \int dP'/\beta_P' \tag{388}$$

Possibility 3. $\theta_t = 0$, $\theta_{Dt} = 0$, $\theta_{Ds} = 0$ and the following conditions

$$\omega_I' = 0 \qquad \sigma_R' = \kappa_R' + \omega_R' - 1 \qquad \kappa_I' = 0 \qquad a' = \kappa_R'/(1 - \omega_R') \tag{389}$$

$$\lambda_I' = 0 \qquad \nu_R' = \varepsilon_R' + \lambda_R' - 1 \qquad \varepsilon_I' = 0 \qquad c' = \varepsilon_R'/(1 - \lambda_R') \tag{390}$$

$$\delta_I' = 0 \qquad \eta_R' = \tau_R' + \delta_R' - 1 \qquad \tau_I' = 0 \qquad e' = \tau_R'/(1 - \delta_R') \tag{391}$$

and equations (342), (344) and (346) give

$$\Phi_t = b_t \int dE'/\beta_E' + b_t^o \int dP'/\beta_P' \tag{392}$$

$$\Phi_{Dt} = - b_{Dt} \int dE'/\beta_E' - b_{Dt}^o \int dP'/\beta_P' \tag{393}$$

$$\Phi_{Ds} = - b_{Ds} \int dE'/\beta_E' - b_{Ds}^o \int dP'/\beta_P' \tag{394}$$

where

$$b_t = \kappa_R'/[(1 - \omega_R')D_{tk}'] \qquad b_t^o = (\kappa_R' + \omega_R' - 1)/D_{sk}' \tag{395}$$

$$b_{Dt} = \epsilon_R'/[(1 - \lambda_R')D_{tk}'] \qquad b_{Dt}^o = (\epsilon_R' + \lambda_R' - 1)/D_{sk}' \tag{396}$$

$$b_{Ds} = \tau_R'/[(1 - \delta_R')D_{tk}'] \qquad b_{Ds}^o = (\tau_R' + \delta_R' - 1)/D_{sk}' \tag{397}$$

The sinusoidal and exponential solutions of this section may have application to quantum junctions such as Josephson junctions, superlattices, quantum wells, quantum wires and quantum dots.[21] In these electronic devices electrons are trapped in regions of space that are bounded by abrupt changes of energy density and pressure. The solutions may also have applications to cosmology because they can describe the dependence of time and dimension on the average energy density and pressure of the universe.[2,5]

Case B. Incoherent Energy and Coherent Spacetime and Dimension.

This is the case of a slow thermodynamic process occurring in the superconducting state of a high-T_c compound. Consider the decoupled form of equations (267) through (269) with external potentials set equal to zero

$$(1 - \bar{\mu}^{cs})(E^{cs} + \bar{D}_{tk}^{cs}\beta_E^{cs}) - \bar{D}_{sk}^{cs}\beta_P^{cs} = 0 \tag{398}$$

$$(1 - \bar{\omega}^{cs})(1 - j\bar{D}_{tk}^{cs}\beta_E^{cs}\ \partial\theta_t/\partial E^{cs}) + j\bar{D}_{sk}^{cs}\beta_P^{cs}\ \partial\theta_t/\partial P^{cs} = 0 \tag{399}$$

$$(1 - \bar{\lambda}^{cs})(1 + j\bar{D}_{tk}^{cs}\beta_E^{cs}\ \partial\theta_{Dt}/\partial E^{cs}) - j\bar{D}_{sk}^{cs}\beta_P^{cs}\ \partial\theta_{Dt}/\partial P^{cs} = 0 \tag{400}$$

$$(1 - \bar{\delta}^{cs})(1 + j\bar{D}_{tk}^{cs}\beta_E^{cs}\ \partial\theta_{Ds}/\partial E^{cs}) - j\bar{D}_{sk}^{cs}\beta_P^{cs}\ \partial\theta_{Ds}/\partial P^{cs} = 0 \tag{401}$$

where $E^{cs} = U^{cs}/(V\theta_V)$ and where t^{cs}, D_t^{cs}, D_s^{cs}, $\bar{D}_{tk}^{cs}$ and $\bar{D}_{sk}^{cs}$ = constants for the case of coherent spacetime and dimensions. Note that $\bar{t}^{cs}$, $\bar{D}_t^{cs}$ and $\bar{D}_s^{cs}$ have already been divided out of equations (399) through (401) respectively. The solutions to equations (399) through (401) are obtained in an analogous manner to the solutions presented in equations (317B) through (317D) with the result that

$$\theta_t = - a_t \int dE^{cs}/\beta_E^{cs} - a_t^o \int dP_{cs}/\beta_P^{cs} \tag{402}$$

$$\theta_{Dt} = a_{Dt} \int dE^{cs}/\beta_E^{cs} + a_{Dt}^o \int dP^{cs}/\beta_P^{cs} \tag{403}$$

$$\theta_{Ds} = a_{Ds} \int dE^{cs}/\beta_E^{cs} + a_{Ds}^o \int dP^{cs}/\beta_P^{cs} \tag{404}$$

For constant values of β_E^{cs} and β_P^{cs} the solutions are written as

$$\theta_t = - a_t E^{cs}/\beta_{Ek}^{cs} - a_t^o P^{cs}/\beta_{Pk}^{cs} \tag{405}$$

$$\theta_{Dt} = a_{Dt} E^{cs}/\beta_{Ek}^{cs} + a_{Dt}^o P^{cs}/\beta_{Pk}^{cs} \tag{406}$$

$$\theta_{Ds} = a_{Ds} E^{cs}/\beta_{Ek}^{cs} + a_{Ds}^o P^{cs}/\beta_{Pk}^{cs} \tag{407}$$

where the constants appearing in equations (402) through (407) are given by

$$a_t = j\bar{\kappa}^{cs}/[(1 - \bar{\omega}^{cs})\bar{D}_{tk}^{cs}] \qquad a_t^o = j\bar{\sigma}^{cs}/\bar{D}_{sk}^{cs} \tag{408}$$

$$a_{Dt} = j\bar{\varepsilon}^{cs}/[(1 - \bar{\lambda}^{cs})\bar{D}_{tk}^{cs}] \qquad a_{Dt}^o = j\bar{\nu}^{cs}/\bar{D}_{sk}^{cs} \tag{409}$$

$$a_{Ds} = j\bar{\tau}^{cs}/[(1 - \bar{\delta}^{cs})\bar{D}_{tk}^{cs}] \qquad a_{Ds}^o = j\bar{\eta}^{cs}/\bar{D}_{sk}^{cs} \tag{410}$$

where the right hand sides of equations (408) through (410) must be real numbers. The solutions given in equations (402) through (407) can be verified by direct substitution into equations (399) through (401) and taking account of the relationship between the eigenvalues and separation constants given in equations (317H) through (317J) which for the case at hand are written as

$$\bar{\kappa}^{cs} + \bar{\omega}^{cs} - 1 = \bar{\sigma}^{cs} \tag{411}$$

$$\bar{\varepsilon}^{cs} + \bar{\lambda}^{cs} - 1 = \bar{\nu}^{cs} \tag{412}$$

$$\bar{\tau}^{cs} + \bar{\delta}^{cs} - 1 = \bar{\eta}^{cs} \tag{413}$$

Additional equations result from the requirement that the imaginary parts of the right hand sides of equations (408) through (410) must vanish, which brings the components of the complex number time dimension and space dimension, $\bar{D}_{tk}^{cs}$ and $\bar{D}_{sk}^{cs}$ respectively, into the relationship between the eigenvalues and the separation constants. If the constant time dimension and space dimension coefficients that appear in equations (398) through (401) and in equations (408) through (410) are taken to be real numbers then the additional equations become

$$\kappa_R^{cs}(1 - \omega_R^{cs}) = \kappa_I^{cs}\omega_I^{cs} \qquad \sigma_R^{cs} = 0 \tag{414}$$

$$\varepsilon_R^{cs}(1 - \lambda_R^{cs}) = \varepsilon_I^{cs}\lambda_I^{cs} \qquad \nu_R^{cs} = 0 \tag{415}$$

$$\tau_R^{cs}(1 - \delta_R^{cs}) = \tau_I^{cs}\delta_I^{cs} \qquad \eta_R^{cs} = 0 \tag{416}$$

Case B also describes partially coherent energy.

Case C. Coherent Energy, Coherent Spacetime and Coherent Dimension.

This situation arises in an ultrafast thermodynamic process occurring in the superconducting state of a high-T_c compound. The decoupled equations cor-

responding to equations (282) through (285) are

$$(1 - \bar{\mu}^{tc})(E^{tc} + \bar{D}_{tk}^{tc}\beta_E^{tc}) - \bar{D}_{sk}^{tc}\beta_P^{tc} + \bar{W}_E E^{tc} = 0 \tag{417}$$

$$(1 - \bar{\omega}^{tc})(1 - j\bar{D}_{tk}^{tc}\beta_E^{tc}\,\partial\theta_t/\partial E^{tc}) + j\bar{D}_{sk}^{tc}\beta_P^{tc}\,\partial\theta_t/\partial P^{tc} + \bar{W}_t = 0 \tag{418}$$

$$(1 - \bar{\lambda}^{tc})(1 + j\bar{D}_{tk}^{tc}\beta_E^{tc}\,\partial\theta_{Dt}/\partial E^{tc}) - j\bar{D}_{sk}^{tc}\beta_P^{tc}\,\partial\theta_{Dt}/\partial P^{tc} + \bar{W}_{Dt} = 0 \tag{419}$$

$$(1 - \bar{\delta}^{tc})(1 + j\bar{D}_{tk}^{tc}\beta_E^{tc}\,\partial\theta_{Ds}/\partial E^{tc}) - j\bar{D}_{sk}^{tc}\beta_P^{tc}\,\partial\theta_{Ds}/\partial P^{tc} + \bar{W}_{Ds} = 0 \tag{420}$$

where U^{tc}, t^{tc}, D_t^{tc}, D_s^{tc}, $\bar{D}_{tk}^{tc}$ and $\bar{D}_{sk}^{tc}$ = constants. For coherent energy and coherent spacetime it is always possible to write

$$dE^{tc} = d[U^{tc}/(V\theta_V)] = -E^{tc}d\theta_V/\theta_V \tag{421}$$

where θ_V and T are not independent because U^{tc} = constant. The solutions to equations (418) through (420) with the external potentials set equal to zero can be written as

$$\theta_t = -c_t\int dE^{tc}/\beta_E^{tc} - c_t^o\int dP^{tc}/\beta_P^{tc} \tag{422}$$

$$\theta_{Dt} = c_{Dt}\int dE^{tc}/\beta_E^{tc} + c_{Dt}^o\int dP^{tc}/\beta_P^{tc} \tag{423}$$

$$\theta_{Ds} = c_{Ds}\int dE^{tc}/\beta_E^{tc} + c_{Ds}^o\int dP^{tc}/\beta_P^{tc} \tag{424}$$

where

$$c_t = j\bar{\kappa}^{tc}/[(1 - \bar{\omega}^{tc})\bar{D}_{tk}^{tc}] \qquad c_t^o = j\bar{\sigma}^{tc}/\bar{D}_{sk}^{tc} \tag{425}$$

$$c_{Dt} = j\bar{\epsilon}^{tc}/[(1 - \bar{\lambda}^{tc})\bar{D}_{tk}^{tc}] \qquad c_{Dt}^o = j\bar{\nu}^{cs}/\bar{D}_{sk}^{tc} \tag{426}$$

$$c_{Ds} = j\bar{\tau}^{tc}/[(1 - \bar{\delta}^{tc})\bar{D}_{tk}^{tc}] \qquad c_{Ds}^o = j\bar{\eta}^{cs}/\bar{D}_{sk}^{tc} \tag{427}$$

and, as before, the right hand sides of equations (425) through (427) must be real numbers. For a limited region of pressure and energy density in which β_E^{tc} and β_P^{tc} can be taken to be constants the solutions can be written as

$$\theta_t = -c_t E^{tc}/\beta_{Ek}^{tc} - c_t^o P^{tc}/\beta_{Pk}^{tc} \tag{428}$$

$$\theta_{Dt} = c_{Dt}E^{tc}/\beta_{Ek}^{tc} + c_{Dt}^o P^{tc}/\beta_{Pk}^{tc} \tag{429}$$

$$\theta_{Ds} = c_{Ds}E^{tc}/\beta_{Ek}^{tc} + c_{Ds}^o P^{tc}/\beta_{Pk}^{tc} \tag{430}$$

where the gauge functions are given by equations (101) and (102) respectively. The solutions in equations (422) through (424) and equations (428) through (430)

can be verified by direct substitution into equations (418) through (420). Also from equations (317H) through (317J) it follows that

$$\bar{\kappa}^{tc} + \bar{\omega}^{tc} - 1 = \bar{\sigma}^{tc} \tag{432}$$

$$\bar{\varepsilon}^{tc} + \bar{\lambda}^{tc} - 1 = \bar{\nu}^{tc} \tag{433}$$

$$\bar{\tau}^{tc} + \bar{\delta}^{tc} - 1 = \bar{\eta}^{tc} \tag{434}$$

The reality of the equations (425) through (427) gives additional equations that relate the real and imaginary components of the eigenvalues and separation constants to the real and imaginary parts of the constants $\bar{D}^{tc}_{tk}$ and $\bar{D}^{tc}_{sk}$ that introduce the time and space dimensions in equations (418) through (420). If the constant time and space dimension coefficients that appear in equations (417) through (420) and in equations (425) through (427) are assumed to be real numbers then the reality of equations (425) through (427) gives the following relations

$$\kappa^{tc}_R(1 - \omega^{tc}_R) = \kappa^{tc}_I\omega^{tc}_I \qquad \sigma^{tc}_R = 0 \tag{435}$$

$$\varepsilon^{tc}_R(1 - \lambda^{tc}_R) = \varepsilon^{tc}_I\lambda^{tc}_I \qquad \nu^{tc}_R = 0 \tag{436}$$

$$\tau^{tc}_R(1 - \delta^{tc}_R) = \tau^{tc}_I\delta^{tc}_I \qquad \eta^{tc}_R = 0 \tag{437}$$

7. SUBSTRUCTURE OF TIME AND DIMENSION. The first order differential eigenvalue equations of time, time dimension and space dimension that were presented in Sections 5 and 6 are the bulk matter analogs of the stationary state Dirac equation of microscopic physics. This section generalizes these equations and develops analogs of the time dependent Dirac equation. This can be done by making the following replacements for the eigenvalues

$$\bar{\omega} \to j\gamma\partial/\partial\mathcal{E} \qquad \bar{\lambda} \to j\gamma\partial/\partial\mathcal{E} \qquad \bar{\delta} \to j\gamma\partial/\partial\mathcal{E} \tag{438}$$

in all of the eigenvalue equations of Sections 5 and 6, where $\mathcal{E}$ = energy density of the particles (chronons) that constitute the physical basis of the substructure of coherent time and coherent dimensions, and γ = fundamental constant having the dimensions energy density which for atomic and molecular structure must be given by

$$\gamma \sim \hbar/(4\pi/3\ a_B^3 t_B) \tag{439}$$

where h = Planck's constant, $\hbar = h/(2\pi)$, a_B = Bohr radius and t_B = Bohr time, where the latter two quantities are given by[16]

$$a_B = \hbar^2/(m_e e^2) \qquad t_B = \hbar^3/(m_e e^4) \tag{440}$$

where m_e = electron mass. Then it follows that for atomic and molecular structure

$$\gamma \sim 3/(4\pi)(m_e^4 e^{10}/\hbar^8) \tag{441}$$

The constant γ sets the scale of the quantum partial differential equations

derived in this section. For instance, at the level of elementary particles

$$\gamma \sim \hbar/(4\pi/3\ a_c^3 t_c) \sim 3/(4\pi)(m^4c^5/\hbar^3) \tag{442}$$

where m = mass of the gauge boson mediating the interaction, and where $a_c = \hbar/(mc)$ = Compton wavelength and $t_c = R/c$ where R = range of weak interaction force or range of strong interaction force, and c = light speed. At the level of quantum gravity

$$\gamma = \hbar/(4\pi/3\ a_p^3 t_p) \tag{443}$$

where a_p = Planck length and t_p = Planck time.[5,7]

Chronons are time coherent bosons. For the weak interactions the bosons $W^{\pm}$ and Z^{o} in a time coherent state are chronons. The gluons are the eight gauge bosons of the strong interactions and when they occur in a coherent time state they can be represented as chronons. However, at the scale atomic and molecular structure the chronons generally are phonons, photons and electron pairs. Phonons are the quanta of lattice vibrations in solids, while photons are the gauge bosons of the electromagnetic interaction. The energy density of the chronons $\mathcal{E}$ can always be written as

$$\mathcal{E} = n\varepsilon \tag{444}$$

where n = average chronon number density and ε = average energy per chronon. This is to be distinguished from the ordinary matter energy density E, particle number density n and average energy per particle ε which are related by

$$E = n\varepsilon \tag{445}$$

It is assumed that E and $\mathcal{E}$ are independent quantities, and that the time and dimension variables are functions of both types of energy density. For high-T_c superconductors the ordinary matter energy density refers to the binding energy of the crystal lattice. On the other hand, the average energy per chronon is associated with the coherent phonons that represent the lattice vibrations of a high-T_c material in its superconducting state, and with the coherent time electron pairs that form bound states due to their interaction with the coherent time phonons of the lattice vibrations. The theory of coherent time photons that are associated with thermal states of high-T_c superconductors has already appeared in the literature.[22] These are the chronons of blackbody electromagnetic radiation in a coherent time state. The coherent spacetime state of the electrons that form Cooper pairs in the superconducting state of a high-T_c material has internal phase angles of the time and space coordinates given by $\theta_t = \pi/6$ and $\theta_r = \pi/3$.[16,19] For the coherent blackbody radiation associated with these electrons the conservation of momentum in electron-photon collisions $h\bar{\nu}/c = m\bar{v}$ gives the internal phase angles of the frequency as $\theta_\nu = \pi/3 - \pi/6 = \pi/6$ and therefore the internal phase angles of spacetime for coherent blackbody radiation are $\theta_{tR} = \theta_{rR} = -\pi/6$.[22] The spacetime internal phase angles for phonons in the superconducting state of a high-T_c material are approximately the same as those of coherent blackbody photons.

The quantum equations for time and dimension expressed in terms of the

chronon energy density $\bar{E}$ will now be developed for three cases of interest to high-T_c superconductivity.

Case A. Incoherent Energy and Partial Coherence of Spacetime and Dimension.

This case describes a slow thermodynamic process in the normal state of a high-T_c superconducting material. From equations (251) through (253) and equation (438) it follows that the decoupled substructure dependent equations of time, time dimension and space dimension are respectively

$$(1 - j\gamma\partial/\partial E)(\bar{t}' - D'_{tk}\beta'_{Ek}\,\partial\bar{t}'/\partial E') + D'_{sk}\beta'_{Pk}\,\partial\bar{t}'/\partial P' + \bar{W}'_{t}\bar{t}' = 0 \qquad (446)$$

$$(1 - j\gamma\partial/\partial E)(\bar{D}'_{t} + D'_{tk}\beta'_{Ek}\,\partial\bar{D}'_{t}/\partial E') - D'_{sk}\beta'_{Pk}\,\partial\bar{D}'_{t}/\partial P' + \bar{W}'_{Dt}\bar{D}'_{t} = 0 \qquad (447)$$

$$(1 - j\gamma\partial/\partial E)(\bar{D}'_{s} + D'_{tk}\beta'_{Ek}\,\partial\bar{D}'_{s}/\partial E') - D'_{sk}\beta'_{Pk}\,\partial\bar{D}'_{s}/\partial P' + \bar{W}'_{Ds}\bar{D}'_{s} = 0 \qquad (448)$$

In this case both the magnitudes and phase angles of the time and dimensions are functions of E', P' and $\bar{E}$. For simplicity the gauge parameters $\beta'_{\bar{E}}$ and $\beta'_{\bar{P}}$ are assumed to have constant values β'_{Ek} and β'_{Pk} respectively. Case A also describes partially coherent energy.

Case B. Incoherent Energy and Coherent Spacetime and Dimension.

This is the case of a slow thermodynamic process that occurs in the superconducting state of a high-T_c superconductor substance. Combining equations (251) through (253) with equation (438) and using the following conditions for the coherence of time and dimension

$$d\bar{t}^{cs} = j\bar{t}^{cs}d\theta_{t} \qquad d\bar{D}_{t}^{cs} = j\bar{D}_{t}^{cs}d\theta_{Dt} \qquad d\bar{D}_{s}^{cs} = j\bar{D}_{s}^{cs}d\theta_{Ds} \qquad (449)$$

gives the following decoupled equations

$$(1 - j\gamma\partial/\partial E)(\bar{t}^{cs} - jD_{tk}^{cs}\beta_{Ek}^{cs}\bar{t}^{cs}\,\partial\theta_{t}/\partial E^{cs}) + jD_{sk}^{cs}\beta_{Pk}^{cs}\bar{t}^{cs}\,\partial\theta_{t}/\partial P^{cs} + \bar{W}_{t}^{cs}\bar{t}^{cs} = 0 \qquad (450)$$

$$(1 - j\gamma\partial/\partial E)(\bar{D}_{t}^{cs} + jD_{tk}^{cs}\beta_{Ek}^{cs}\bar{D}_{t}^{cs}\,\partial\theta_{Dt}/\partial E^{cs}) - jD_{sk}^{cs}\beta_{Pk}^{cs}\bar{D}_{t}^{cs}\,\partial\theta_{Dt}/\partial P^{cs} + \bar{W}_{Dt}^{cs}\bar{D}_{t}^{cs} = 0 \qquad (451)$$

$$(1 - j\gamma\partial/\partial E)(\bar{D}_{s}^{cs} + jD_{tk}^{cs}\beta_{Ek}^{cs}\bar{D}_{s}^{cs}\,\partial\theta_{Ds}/\partial E^{cs}) - jD_{sk}^{cs}\beta_{Pk}^{cs}\bar{D}_{s}^{cs}\,\partial\theta_{Ds}/\partial P^{cs} + \bar{W}_{Ds}^{cs}\bar{D}_{s}^{cs} = 0 \qquad (452)$$

where it is assumed that $\bar{E}$ and E^{cs} are independent variables and that

$$\theta_{t} = \theta_{t}(E^{cs}, P^{cs}, \bar{E}) \qquad \theta_{Dt} = \theta_{Dt}(E^{cs}, P^{cs}, \bar{E}) \qquad \theta_{Ds} = \theta_{Ds}(E^{cs}, P^{cs}, \bar{E}) \qquad (453)$$

For this case the magnitudes of the time and dimensions are constants.

Taking the real and imaginary parts of equations (450) through (452) gives the following sets of equations that describe the superconducting state of a high-T_c compound

$$1 + \gamma\partial\theta_t/\partial E - \gamma D^{cs}_{tk}\beta^{cs}_{Ek}\,\partial^2\theta_t/\partial E\partial E^{cs} + W^{cs}_t \cos\theta_{Wt} = 0 \tag{454}$$

$$- D^{cs}_{tk}\beta^{cs}_{Ek}\,\partial\theta_t/\partial E^{cs}(1 + \gamma\partial\theta_t/\partial E) + D^{cs}_{sk}\beta^{cs}_{Pk}\,\partial\theta_t/\partial P^{cs} + W^{cs}_t \sin\theta_{Wt} = 0 \tag{455}$$

$$1 + \gamma\partial\theta_{Dt}/\partial E + \gamma D^{cs}_{tk}\beta^{cs}_{Ek}\,\partial^2\theta_{Dt}/\partial E\partial E^{cs} + W^{cs}_{Dt}\cos\theta_{WDt} = 0 \tag{456}$$

$$D^{cs}_{tk}\beta^{cs}_{Ek}\,\partial\theta_{Dt}/\partial E^{cs}(1 + \gamma\partial\theta_{Dt}/\partial E) - D^{cs}_{sk}\beta^{cs}_{Pk}\,\partial\theta_{Dt}/\partial P^{cs} + W^{cs}_{Dt}\sin\theta_{WDt} = 0 \tag{457}$$

$$1 + \gamma\partial\theta_{Ds}/\partial E + \gamma D^{cs}_{tk}\beta^{cs}_{Ek}\,\partial^2\theta_{Ds}/\partial E\partial E^{cs} + W^{cs}_{Ds}\cos\theta_{WDs} = 0 \tag{458}$$

$$D^{cs}_{tk}\beta^{cs}_{Ek}\,\partial\theta_{Ds}/\partial E^{cs}(1 + \gamma\partial\theta_{Ds}/\partial E) - D^{cs}_{sk}\beta^{cs}_{Pk}\,\partial\theta_{Ds}/\partial P^{cs} + W^{cs}_{Ds}\sin\theta_{WDs} = 0 \tag{459}$$

The derivative of the internal phase angle of time with respect to the temperature is then calculated as

$$\partial\theta_t/\partial T = \partial\theta_t/\partial E\;\partial E/\partial T + \partial\theta_t/\partial E^{cs}\;\partial E^{cs}/\partial T + \partial\theta_t/\partial P^{cs}\;\partial P^{cs}/\partial T \tag{460}$$

$$\sim \partial\theta_t/\partial E\;\partial E/\partial T$$

where the approximation in equation (460) is valid if E^{cs} and P^{cs} are slowly changing functions of temperature.

Case C. Coherent Energy, Coherent Spacetime and Coherent Dimensions.

This case corresponds to an ultrafast thermodynamic process in the superconducting phase of a high-T_c material. Then equation (438) and equations (251) through (253) give the decoupled equations for total coherence as

$$(1 - j\gamma\partial/\partial E)(\bar{t}^{tc} - jD^{tc}_{tk}\beta^{tc\text{-}tc}_{Ek}\bar{t}^{tc}\,\partial\theta_t/\partial E^{tc}) + jD^{tc}_{sk}\beta^{tc\text{-}tc}_{Pk}\bar{t}^{tc}\,\partial\theta_t/\partial P^{tc} + \bar{W}_t\bar{t}^{tc} = 0 \tag{461}$$

$$(1 - j\gamma\partial/\partial E)(\bar{D}^{tc}_t + jD^{tc}_{tk}\beta^{tc\text{-}tc}_{Ek}\bar{D}^{tc}_t\,\partial\theta_{Dt}/\partial E^{tc}) - jD^{tc}_{sk}\beta^{tc\text{-}tc}_{Pk}\bar{D}^{tc}_t\,\partial\theta_{Dt}/\partial P^{tc} + \bar{W}_{Dt}\bar{D}^{tc}_t = 0 \tag{462}$$

$$(1 - j\gamma\partial/\partial E)(\bar{D}^{tc}_s + jD^{tc}_{tk}\beta^{tc\text{-}tc}_{Ek}\bar{D}^{tc}_s\,\partial\theta_{Ds}/\partial E^{tc} - jD^{tc}_{sk}\beta^{tc\text{-}tc}_{Pk}\bar{D}^{tc}_s\,\partial\theta_{Ds}/\partial P^{tc} + \bar{W}_{Ds}\bar{D}^{tc}_s = 0 \tag{463}$$

These equations lead to relations analogous to equations (454) through (459).

Finally it should be stated that for a noninteracting system with $\beta_E = 0$ and $\beta_P = 0$ it follows from any of the basic substructure dependent eigenvalue equations such as equations (446) through (448) that

$$j\gamma\partial\bar{t}'/\partial E = (1 + \bar{W}'_t)\bar{t}' \tag{463A}$$

$$j\gamma\partial\bar{D}'_t/\partial E = (1 + \bar{W}'_{Dt})\bar{D}'_t \tag{463B}$$

$$j\gamma\partial\bar{D}'_s/\partial E = (1 + \bar{W}'_{Ds})\bar{D}_s \tag{463C}$$

whose solutions for constant external potentials are

$$\bar{t}' = \bar{A} \exp[-j(1 + \bar{W}_t')/\gamma \bar{E}] \tag{463D}$$

$$\bar{D}_t' = \bar{B} \exp[-j(1 + \bar{W}_{Dt}')/\gamma \bar{E}] \tag{463E}$$

$$\bar{D}_s' = \bar{C} \exp[-j(1 + \bar{W}_{Ds}')/\gamma \bar{E}] \tag{463F}$$

Similar solutions hold for the case of coherent spacetime (Case B) and for the coherence of both thermodynamics and spacetime (Case C).

8. QUANTIZED TIME AND DIMENSION STRUCTURES OF ENERGY AND PRESSURE. This section considers structured energy and pressure and develops the eigenvalues and eigenfunctions that describe the time and dimension structures that can exist in a pressure and energy density space in which a Coulomb-like form of external potential is present. A set of second order Schrödinger-like equations is developed which determines the spectrum and eigenfunctions for time and dimension structures. For a limited region of energy density-pressure space where $\beta_{\bar{E}}$ and β_P are approximately constants, the solution of the decoupled first order time, time dimension and space dimension equations (324) through (326) with zero external potentials can according to equations (340) through (352) be written as

$$\bar{t} = \bar{A}\bar{t}^h \exp \Phi_t \qquad \bar{D}_t = \bar{B}\bar{D}_t^h \exp \Phi_{Dt} \qquad \bar{D}_s = \bar{C}\bar{D}_s^h \exp \Phi_{Ds} \tag{464}$$

where the harmonic solutions are obtained from equations (347), (349) and (351), after dropping the primes for convenience, as

$$\bar{t}^h = A \exp\{j[b\bar{E}/(D_{tk}\beta_{\bar{E}k}) + \sigma_I P/(D_{sk}\beta_{Pk})]\} \tag{465}$$

$$\bar{D}_t^h = B \exp\{j[-\ell\bar{E}/(D_{tk}\beta_{\bar{E}k}) - \nu_I P/(D_{sk}\beta_{Pk})]\} \tag{466}$$

$$\bar{D}_s^h = C \exp\{j[-f\bar{E}/(D_{tk}\beta_{\bar{E}k}) - \eta_I P/(D_{sk}\beta_{Pk})]\} \tag{467}$$

where b, ℓ, f, σ_I, ν_I, η_I, D_{tk}, D_{sk}, $\beta_{\bar{E}k}$ and β_{Pk} are all constants. Equations (465) through (467) are only approximate solutions because they assume $\beta_{\bar{E}}$ and β_P to be constants.

A. Schrödinger Form of the Time and Dimension Equations.

The time and dimension equations that were developed in Section 5 are first order differential eigenvalue equations that describe the spectrum and eigenfunctions of time and dimension in an energy density-pressure field. They are the bulk matter analogs of the Dirac equation for microscopic particle systems. However, the approximate solutions in equations (465) through (467) suggest the definition of the following differential operators

$$\bar{F} = -j\gamma\partial/\partial\bar{E} \qquad \bar{G} = -j\gamma\partial/\partial P \tag{468}$$

$$H = a(\bar{F}^2 + \bar{G}^2) = -a\gamma^2(\partial^2/\partial\bar{E}^2 + \partial^2/\partial P^2) \tag{469}$$

where γ and a = fundamental constants that define the quantum structure of bulk matter. Combining equations (465) through (467) with equation (431) gives

$$\bar{\bar{F}}\bar{t}^h = [\gamma b/(D_{tk}\beta_{Ek})]\bar{t}^h \qquad \bar{\bar{G}}\bar{t}^h = [\gamma\sigma_I/(D_{sk}\beta_{Pk})]\bar{t}^h \tag{470}$$

$$\bar{\bar{F}}\bar{D}_t^h = -[\gamma\ell/(D_{tk}\beta_{Ek})]\bar{D}_t^h \qquad \bar{\bar{G}}\bar{D}_t^h = -[\gamma\nu_I/(D_{sk}\beta_{Pk})]\bar{D}_t^h \tag{471}$$

$$\bar{\bar{F}}\bar{D}_s^h = -[\gamma f/(D_{tk}\beta_{Ek})]\bar{D}_s^h \qquad \bar{\bar{G}}\bar{D}_s^h = -[\gamma\eta_I/(D_{sk}\beta_{Pk})]\bar{D}_s^h \tag{472}$$

Now consider the following eigenvalue equations

$$H\bar{t}^h = E_t\bar{t}^h \qquad H\bar{D}_t^h = E_{Dt}\bar{D}_t^h \qquad H\bar{D}_s^h = E_{Ds}\bar{D}_s^h \tag{473}$$

where E_t, E_{Dt} and E_{Ds} are eigenvalues to be determined. Combining equations (469) and (473) with equations (465) through (467) gives

$$H\bar{t}^h = a\gamma^2[b^2/(D_{tk}\beta_{Ek})^2 + \sigma_I^2/(D_{sk}\beta_{Pk})^2]\bar{t}^h = E_t\bar{t}^h \tag{474}$$

$$H\bar{D}_t^h = a\gamma^2[\ell^2/(D_{tk}\beta_{Ek})^2 + \nu_I^2/(D_{sk}\beta_{Pk})^2]\bar{D}_t^h = E_{Dt}\bar{D}_t^h \tag{475}$$

$$H\bar{D}_s^h = a\gamma^2[f^2/(D_{tk}\beta_{Ek})^2 + \eta_I^2/(D_{sk}\beta_{Pk})^2]\bar{D}_s^h = E_{Ds}\bar{D}_s^h \tag{476}$$

so that equations (465) through (467) are solutions of the eigenvalue equations (473) provided that

$$E_t = a\gamma^2[b^2/(D_{tk}\beta_{Ek})^2 + \sigma_I^2/(D_{sk}\beta_{Pk})^2] \tag{477}$$

$$E_{Dt} = a\gamma^2[\ell^2/(D_{tk}\beta_{Ek})^2 + \nu_I^2/(D_{sk}\beta_{Pk})^2] \tag{478}$$

$$E_{Ds} = a\gamma^2[f^2/(D_{tk}\beta_{Ek})^2 + \eta_I^2/(D_{sk}\beta_{Pk})^2] \tag{479}$$

where γ is given by equations (439), (441), (442) or (443) and a = constant having the dimensions of energy.

For the special cases of the time box, time dimension box and space dimension box described by equations (360) through (362) it follows from equations (477) through (479) that

$$E_t = a\gamma^2\pi^2[n^2/(E_2 - E_1)^2 + m^2/(P_2 - P_1)^2] \tag{480}$$

$$E_{Dt} = a\gamma^2\pi^2[n^2/(E_4 - E_3)^2 + m^2/(P_4 - P_3)^2] \tag{481}$$

$$E_{Ds} = a\gamma^2\pi^2[n^2/(E_6 - E_5)^2 + m^2/(P_6 - P_5)^2] \tag{482}$$

which are associated respectively with the following wave functions

$$\bar{t}^h = A \exp\{j\pi[n(E - E_1)/(E_2 - E_1) + m(P - P_1)/(P_2 - P_1)]\} \quad (483)$$

$$\bar{D}_t^h = B \exp\{j\pi[n(E - E_3)/(E_4 - E_3) + m(P - P_3)/(P_4 - P_3)]\} \quad (484)$$

$$\bar{D}_s^h = C \exp\{j\pi[n(E - E_5)/(E_6 - E_5) + m(P - P_5)/(P_6 - P_5)]\} \quad (485)$$

which are eigenfunctions of the operator H given in equations (469). Therefore for zero values of the external potential, the approximate solutions of the first order bulk matter eigenvalue equations as given by equations (465) through (467) are also solutions to the second order Schrödinger-like equations (473) provided that the eigenvalues of the Schrödinger-like equations are given by equations (477) through (479) for the general case and by equations (480) through (482) for boxes in energy-pressure space.

This suggests that for an external potential operating in energy-pressure space, the second order Schrödinger-like equations for time and dimension that approximate the first order decoupled Dirac-like bulk matter eigenvalue equations (315) through (317) are written as

$$H\bar{t} = - a\gamma^2(\partial^2/\partial E^2 + \partial^2/\partial P^2)\bar{t} + \bar{V}_t\bar{t} = \bar{E}_t\bar{t} \quad (486)$$

$$H\bar{D}_t = - a\gamma^2(\partial^2/\partial E^2 + \partial^2/\partial P^2)\bar{D}_t + \bar{V}_{Dt}\bar{D}_t = \bar{E}_{Dt}\bar{D}_t \quad (487)$$

$$H\bar{D}_s = - a\gamma^2(\partial^2/\partial E^2 + \partial^2/\partial P^2)\bar{D}_s + \bar{V}_{Ds}\bar{D}_s = \bar{E}_{Ds}\bar{D}_s \quad (488)$$

These equations give the stationary states of time and dimension with respect to a substructure parameter π = particle number density of the substructure particles (chronons) of time and dimension. For stationary states the time and dimensions have a dependence on the substructure parameter π that is described by

$$H\bar{t} = j\gamma\partial\bar{t}/\partial\pi \qquad \bar{t} = \bar{t}(E,P)\exp(- j\bar{E}_t\pi/\gamma) \quad (489)$$

$$H\bar{D}_t = j\gamma\partial\bar{D}_t/\partial\pi \qquad \bar{D}_t = \bar{D}_t(E,P)\exp(- j\bar{E}_{Dt}\pi/\gamma) \quad (490)$$

$$H\bar{D}_s = j\gamma\partial\bar{D}_s/\partial\pi \qquad \bar{D}_s = \bar{D}_s(E,P)\exp(- j\bar{E}_{Ds}\pi/\gamma) \quad (491)$$

Equations (489) through (491) are equivalent to equations (486) through (488). This suggests that the general second order quantum equations for bulk matter with arbitrary dependence on the time and dimension substructure particle (chronon) number density are given by

$$[- a\gamma^2(\partial^2/\partial E^2 + \partial^2/\partial P^2) + \bar{V}_t]\bar{t} = j\gamma\partial\bar{t}/\partial\pi \quad (492)$$

$$[- a\gamma^2(\partial^2/\partial E^2 + \partial^2/\partial P^2) + \bar{V}_{Dt}]\bar{D}_t = j\gamma\partial\bar{D}_t/\partial\pi \quad (493)$$

$$[- a\gamma^2(\partial^2/\partial E^2 + \partial^2/\partial P^2) + \bar{V}_{Ds}]\bar{D}_s = j\gamma\partial\bar{D}_s/\partial\pi \quad (494)$$

For stationary states equations (492) through (494) reduce to equations (486)

through (488).

It should be remembered that the second order bulk matter eigenvalue equations treated in this section were developed from the decoupled first order quantum bulk matter equations for time and dimension (Section 6) under the approximation that the gauge parameters $\beta_{\bar{E}}$ and $\beta_{\bar{P}}$ are constants. Thus the second order eigenvalue equations are only approximations that are valid in limited ranges of the pressure and energy density and only when the decoupling (linearization) procedure of Section 6 is valid. The nonlinear coupled first order eigenvalue equations of Section 5 are valid for the full range of pressure and energy density. Equations (486) through (488) and (492) through (494) can be generalized to the case of complex number energy density and pressure as follows

$$\bar{H}\bar{t} = -a\gamma^2(\partial^2/\partial\bar{E}^2 + \partial^2/\partial\bar{P}^2)\bar{t} + \bar{V}_t\bar{t} = \bar{E}_t\bar{t} \tag{495}$$

$$\bar{H}\bar{D}_t = -a\gamma^2(\partial^2/\partial\bar{E}^2 + \partial^2/\partial\bar{P}^2)\bar{D}_t + \bar{V}_{Dt}\bar{D}_t = \bar{E}_{Dt}\bar{D}_t \tag{496}$$

$$\bar{H}\bar{D}_s = -a\gamma^2(\partial^2/\partial\bar{E}^2 + \partial^2/\partial\bar{P}^2)\bar{D}_s + \bar{V}_{Ds}\bar{D}_s = \bar{E}_{Ds}\bar{D}_s \tag{497}$$

and for the case of dependence on a substructure particle (chronon) number density $\bar{\pi}$

$$\bar{H}\bar{t} = [-a\gamma^2(\partial^2/\partial\bar{E}^2 + \partial^2/\partial\bar{P}^2) + \bar{V}_t]\bar{t} = j\gamma\partial\bar{t}/\partial\bar{\pi} \tag{498}$$

$$\bar{H}\bar{D}_t = [-a\gamma^2(\partial^2/\partial\bar{E}^2 + \partial^2/\partial\bar{P}^2) + \bar{V}_{Dt}]\bar{D}_t = j\gamma\partial\bar{D}_t/\partial\bar{\pi} \tag{499}$$

$$\bar{H}\bar{D}_s = [-a\gamma^2(\partial^2/\partial\bar{E}^2 + \partial^2/\partial\bar{P}^2) + \bar{V}_{Ds}]\bar{D}_s = j\gamma\partial\bar{D}_s/\partial\bar{\pi} \tag{500}$$

Equation (444) relates chronon energy density to chronon particle number density.

The constant γ has the units of energy density and the constant a has the units of energy as designated in the following way

$$[\gamma] = [\bar{E}] = [P] \tag{501}$$

$$[a] = [E_t] = [E_{Dt}] = [E_{Ds}] = [V_t] = [V_{Dt}] = [V_{Ds}] \tag{502}$$

The substructure particle number density π is different from and unrelated to the particle number density of matter $n = \bar{E}/\varepsilon$ where ε = average single particle energy. The particles constituting the substructure of time and dimension are not known experimentally, but their existence can be conjectured from the second order time and dimension equations given in equations (492) through (494) and equations (498) through (500). The constants γ and a are fundamental constants of the quantum theory of time and dimension. In order to regain the standard Schrödinger equation for particles from equations (492) through (494) the following connections have to be made

$$(t\ ,\ D_t\ ,\ D_s) \rightarrow \psi \qquad \gamma \rightarrow \hbar \qquad a \rightarrow 1/(2\mu) \tag{503}$$

$$\pi \rightarrow t \qquad E \rightarrow x \qquad P \rightarrow y \tag{504}$$

However, the quantum bulk matter equations (492) through (494) are fundamentally different from the Schrödinger equation for particles because equation (492), for instance, has time in the numerator and treats time as a wave function in energy-pressure space. Values of γ for various energy scales are given in equations (439), (442) and (443).

For coherent time that is associated with the superconducting state of high-T_c materials the change in time is given by $d\bar{t} = j\bar{t}d\theta_t$ with t = constant, and equation (492) becomes

$$-a\gamma^2[j(\partial^2\theta_t/\partial E^2 + \partial^2\theta_t/\partial P^2) - (\partial\theta_t/\partial E)^2 - (\partial\theta_t/\partial P)^2] + \bar{V}_t = \bar{E}_t = -\gamma\partial\theta_t/\partial\pi \tag{505}$$

where $\bar{t}$ has been divided out in order to obtain equation (505). Taking the real and imaginary parts of equation (505) gives

$$a\gamma^2[(\partial\theta_t/\partial E)^2 + (\partial\theta_t/\partial P)^2] + V_t \cos\theta_{Vt} = -\gamma\partial\theta_t/\partial\pi = E_t \tag{506}$$

$$-a\gamma^2[\partial^2\theta_t/\partial E^2 + \partial^2\theta_t/\partial P^2] + V_t \sin\theta_{Vt} = 0 \tag{507}$$

and also $\theta_{Et} = 0$. If E_t = constant then

$$\theta_t = \theta_t^o - (E_t/\gamma)\pi \tag{508}$$

Because in general π is an increasing function of temperature for $T > T_c$ it follows that θ_t is a decreasing function of temperature above the critical temperature. For $T < T_c$ the internal phase angle of time is given by $\theta_t = \pi/6$ so that π is a constant given by

$$\pi_c = (\gamma/E_t)(\theta_t^o - \pi/6) \tag{509}$$

The value of θ_t^o depends on the atomic structure of the high-T_c compound. For a zero value of the external potential the internal phase angle of coherent time is determined by

$$(\partial\theta_t/\partial E)^2 + (\partial\theta_t/\partial P)^2 = E_t/(a\gamma^2) \tag{510}$$

$$\partial^2\theta_t/\partial E^2 + \partial^2\theta_t/\partial P^2 = 0 \tag{511}$$

where E_t = constant. A simple solution is

$$\theta_t = \alpha E + \delta P \tag{512}$$

where α and δ are constants that satisfy

$$\alpha^2 + \delta^2 = E_t/(a\gamma^2) \tag{513}$$

This suggests that for z = constant

$$\alpha = [E_t/(a\gamma^2)]^{1/2} \sin z \qquad \delta = [E_t/(a\gamma^2)]^{1/2} \cos z \tag{514}$$

and

$$\dot{\theta_t} = [E_t/(a\gamma^2)]^{1/2}(E \sin z + P \cos z) \tag{515}$$

for the case of zero external potential. Equation (511) is Laplace's equation in energy density-pressure space, and equation (510) is the eikonal equation for the internal phase angle of time.

The general solutions to equations (492) through (494) can be written as a sum over eigenfunctions as follows

$$\bar{t}(E,P,n) = \sum_{\nu=1}^{\infty} \bar{e}_\nu \bar{t}_\nu(E,P)e^{-jE_{t\nu}n/\gamma} \tag{516}$$

$$\bar{D}_t(E,P,n) = \sum_{\nu=1}^{\infty} \bar{f}_\nu \bar{D}_{t\nu}(E,P)e^{-jE_{Dt\nu}n/\gamma} \tag{517}$$

$$\bar{D}_s(E,P,n) = \sum_{\nu=1}^{\infty} \bar{g}_\nu \bar{D}_{s\nu}(E,P)e^{-jE_{Ds\nu}n/\gamma} \tag{518}$$

Time, time dimension and space dimension can be interpreted to be wave functions, and they are subject to normalization conditions of the form

$$\int|\bar{t}|^2 dE\, dP = 1 \qquad \int|\bar{D}_t|^2 dE\, dP = 1 \qquad \int|\bar{D}_s|^2 dE\, dP = 1 \tag{519}$$

B. Coulomb Form of External Potential in Energy-Pressure Space.

Consider now the bound states associated with a Coulomb form of external potential in two-dimensional energy-pressure space given by

$$\bar{V} = -\bar{g}/\bar{r} \qquad \bar{r} = (\bar{E}^2 + \bar{P}^2)^{1/2} \tag{520}$$

where

$$\bar{g} = g \exp(j\theta_g) = \text{constant} \tag{521}$$

Then any of the equations (495) through (497) can be written in terms of a two-dimensional Laplacian in polar coordinates of energy density-pressure space as

$$-a\gamma^2(\partial^2\bar{\psi}/\partial\bar{r}^2 + 1/\bar{r}\,\partial\bar{\psi}/\partial\bar{r} + 1/\bar{r}^2\,\partial^2\bar{\psi}/\partial\bar{\phi}^2) - \bar{g}/\bar{r}\,\bar{\psi} = \bar{E}\bar{\psi} \tag{522}$$

where

$$\tan\bar{\phi} = \bar{P}/\bar{E} \qquad \sin\bar{\phi} = \bar{P}/\bar{r} \qquad \cos\bar{\phi} = \bar{E}/\bar{r} \tag{523}$$

where

$$\bar{\phi} = \phi \exp(j\theta_\phi) \tag{524}$$

and where collectively $\bar{\psi} = \bar{t}$, $\bar{D}_t$ or $\bar{D}_s$; $\bar{V} = \bar{V}_t$, $\bar{V}_{Dt}$ or $\bar{V}_{Ds}$; and $\bar{E} = \bar{E}_t$, $\bar{E}_{Dt}$, or $\bar{E}_{Ds}$.

Assuming separation of variables in the form

$$\bar{\psi} = \bar{R}(\bar{r})\bar{\Phi}(\bar{\phi}) \tag{525}$$

allows equation (522) to be written as[17]

$$d^2\bar{\Phi}/d\bar{\phi}^2 + \bar{M}^2\bar{\Phi} = 0 \tag{526}$$

$$\bar{r}^2 d^2\bar{R}/d\bar{r}^2 + \bar{r}d\bar{R}/d\bar{r} + (\bar{k}^2\bar{r}^2 - \bar{M}^2)\bar{R} = 0 \tag{527}$$

where[17]

$$\bar{M} = m \cos\theta_\phi \exp(-j\theta_\phi) \qquad m = 0,\pm1,\pm2,\pm3, \tag{528}$$

$$\bar{k}^2 = (\bar{E} + \bar{g}/\bar{r})/(a\gamma^2) = \bar{k}_o^2 + (\bar{g}/\bar{r})/(a\gamma^2) \tag{529}$$

where

$$\bar{k}_o^2 = \bar{E}/(a\gamma^2) = -|\bar{E}|/(a\gamma^2) \exp(j\theta_E) \tag{530}$$

$$\bar{k}_o = k_o \exp(j\theta_E/2) = -i\bar{k}_o' = -ik_o' \exp(j\theta_E/2) \tag{531}$$

$$\bar{k}_o = -i[|\bar{E}|/(a\gamma^2)]^{1/2} \exp(j\theta_E/2) \qquad \bar{k}_o^2 = \bar{E}/(a\gamma^2) \tag{532}$$

$$\bar{k}_o' = [|\bar{E}|/(a\gamma^2)]^{1/2} \exp(j\theta_E/2) \qquad \bar{k}_o'^2 = -\bar{E}/(a\gamma^2) \tag{533}$$

$$k_o = -i[|\bar{E}|/(a\gamma^2)]^{1/2} \qquad k_o' = [|\bar{E}|/(a\gamma^2)]^{1/2} \tag{534}$$

and finally $k_o = -ik_o'$.

The solution of equation (469) is[17]

$$\bar{\Phi} = \bar{A}e^{i\bar{M}\phi} + Be^{-i\bar{M}\phi} \tag{535}$$

The solution to equation (527) for bound states described by equation (530) can be obtained by making the following substitutions

$$\bar{R} = \bar{y}\bar{r}^{\bar{M}'}e^{-i\bar{k}_o\bar{r}} = \bar{y}\bar{r}^{\bar{M}'}e^{-\bar{k}_o'\bar{r}} \tag{536}$$

$$\bar{x} = 2i\bar{k}_o\bar{r} = 2\bar{k}_o'\bar{r} \tag{537}$$

where[17]

$$\bar{M}' = |m| \cos\theta_\phi \, e^{-j\theta\phi} \tag{538}$$

Equation (536) can also be written as

$$\bar{R}(\bar{E},\bar{P}) = (\bar{E}^2 + \bar{P}^2)^{\bar{M}'/2}\,\bar{y}(\bar{E},\bar{P})\exp[-\bar{k}_o'(\bar{E}^2 + \bar{P}^2)^{1/2}] \tag{539}$$

Then equation (527) becomes

$$\bar{x}d^2\bar{y}/d\bar{x}^2 + (\bar{\beta} - \bar{x})d\bar{y}/d\bar{x} - \bar{\nu}\bar{y} = 0 \tag{540}$$

where

$$\bar{\beta} = 2\bar{M}' + 1 \qquad \bar{\nu} = \bar{M}' + 1/2 - \bar{g}/(2\bar{k}_o'a\gamma^2) \tag{541}$$

Equation (540) is the confluent hypergeometric equation but with complex number dependent and independent variables.[23] The two solutions to equation (540) are written as[23]

$$\bar{y} = A\ {}_1\bar{F}_1(\bar{\nu};\bar{\beta};\bar{x}) + B\,\bar{x}^{1-\bar{\beta}}\ {}_1\bar{F}_1(\bar{\nu} - \bar{\beta} + 1;2 - \bar{\beta};\bar{x}) \tag{542}$$

Only the first solution is finite at $\bar{x} = 0$. This can be seen by noting that equation (541) gives

$$1 - \bar{\beta} = -2\bar{M}' \tag{543}$$

and the real part of $\bar{M}'$ is positive because[17]

$$M_R' = |m|\cos^2\theta_\phi > 0 \tag{544}$$

and therefore $1 - \beta_R < 0$ and the second solution diverges at $\bar{x} = 0$. A generalization to complex numbers of a well known power series expansion gives[23]

$${}_1\bar{F}_1 = \sum_{\sigma=o}^{\infty} \bar{A}_\sigma \bar{x}^\sigma \tag{545}$$

where

$$\bar{A}_\sigma = [\bar{\nu}(\bar{\nu} + 1)\cdots(\bar{\nu} + \sigma - 1)]/[\bar{\beta}(\bar{\beta} + 1)\cdots(\bar{\beta} + \sigma - 1)\sigma!] \tag{546}$$

The only way the solution in equation (545) can be finite is if the series breaks off, and therefore from equation (546) $\bar{\nu}$ must be zero or a negative integer.[23] Therefore equation (541) is written as

$$-n' = \bar{M}' + 1/2 - \bar{g}/(2\bar{k}_o'a\gamma^2) \tag{547}$$

where n' is a positive integer or zero, n' = 0,1,2,3,$\cdots$. From equation (547) it follows that

$$\bar{k}_o' = \bar{g}/[2a\gamma^2(\bar{\eta} + 1/2)] \tag{548}$$

where

$$\bar{\eta} = \bar{M}' + n' \tag{549}$$

Equations (538) and (549) give

$$\eta \cos \theta_\eta = |m| \cos^2 \theta_\phi + n' \tag{550}$$

$$\eta \sin \theta_\eta = - |m| \cos \theta_\phi \sin \theta_\phi \tag{551}$$

from which it follows that

$$\tan \theta_\eta = - (|m| \cos \theta_\phi \sin \theta_\phi)/(|m| \cos^2 \theta_\phi + n') \tag{552}$$

$$\eta^2 = |m|(|m| + 2n')\cos^2 \theta_\phi + n'^2 \tag{553}$$

$$= n^2 - |m|(|m| + 2n')\sin^2 \theta_\phi$$

where

$$n = |m| + n' \tag{554}$$

so that n is a positive integer or zero, n = 0 , 1 , 2 , 3, ··· .

The energy eigenvalues are obtained from equations (533) and (548) to be

$$\bar{E}_\eta = - a\gamma^2 \bar{k}_o'^2 = - [\bar{g}^2/(4a\gamma^2)]/(\bar{\eta} + 1/2)^2 \tag{555}$$

The right hand side of equation (555) can be simplified by writing

$$\bar{N} = \bar{\eta} + 1/2 \tag{556}$$

where

$$\bar{N} = N \exp(j\theta_N) \qquad \bar{\eta} = \eta \exp(j\theta_\eta) \tag{557}$$

Equations (556) and (557) give

$$N \cos \theta_N = \eta \cos \theta_\eta + 1/2 \tag{558}$$

$$N \sin \theta_N = \eta \sin \theta_\eta \tag{559}$$

where θ_η and η are given by equations (552) and (553) respectively. From equations (558) and (559) it follows that

$$\tan \theta_N = (\eta \sin \theta_\eta)/(\eta \cos \theta_\eta + 1/2) \tag{560}$$

$$N^2 = \eta^2 + \eta \cos \theta_\eta + 1/4 \tag{561}$$

For $\theta_\phi = 0$ equations (553) and (561) give $\eta = n$ and $N = n + 1/2$. With these definitions the energy levels of equation (555) can be written as

$$\bar{E}_\eta = E_\eta e^{j\theta_{E\eta}} = - [\bar{g}^2/(4a\gamma^2)]/\bar{N}^2 \tag{562}$$

$$E_\eta = - [g^2/(4a\gamma^2)]/N^2 \tag{563}$$

$$\theta_{E\eta} = 2(\theta_g - \theta_N) \tag{564}$$

and the measured energy is given by

$$E_{\eta m} = E_{\eta} \cos \theta_{E\eta} \tag{565}$$

$$= - [g^2/(4a\gamma^2)]/N^2 \cos[2(\theta_g - \theta_N)]$$

where θ_N and N are given by equations (560) and (561) respectively. The ground state obtained from equation (555) when $\bar{\eta} = 0$ (or m = 0 and n' = 0) is given by

$$\bar{E}_o = - \bar{g}^2/(a\gamma^2) \qquad E_{om} = - g^2/(a\gamma^2) \cos(2\theta_g) \tag{566}$$

For real values of the energy density and pressure, an identical analysis gives

$$E_n = - [g^2/(4a\gamma^2)]/(n + 1/2)^2 \tag{567}$$

where n = 0 , 1 , 2 , 3 , $\cdots$ is a positive integer or zero given by equation (554). The corresponding ground state is

$$E_o = - g^2/(a\gamma^2) \tag{568}$$

Note that the energy eigenvalues $\bar{E}_{\eta}$ calculated in this section can be either $\bar{E}_{t\eta}$, $\bar{E}_{Dt\eta}$ or $\bar{E}_{Ds\eta}$ corresponding to equations (495) through (497) respectively.

The eigenfunctions corresponding to the eigenvalues given in equation (562) are obtained from equations (537), (541), (542) and (547) to be

$$\bar{y} = {}_1\bar{F}_1(\bar{M}' + 1/2 - \bar{g}/(2\bar{k}_o' a\gamma^2) \; ; \; 2\bar{M}' + 1 \; ; \; 2\bar{k}_o'\bar{r}) \tag{569}$$

then equation (539) gives $\bar{R}(\bar{E},\bar{P})$. Combining equations (541) and (548) gives

$$\bar{v} = \bar{M}' + 1/2 - (\bar{\eta} + 1/2) \tag{570}$$

$$= \bar{M}' - \bar{\eta}$$

$$= - n'$$

and therefore equation (569) becomes

$$\bar{y} = {}_1\bar{F}_1[\bar{M}' - \bar{\eta} \; ; \; 2\bar{M}' + 1 \; ; \; \bar{g}/[a\gamma^2(\bar{\eta} + 1/2)](\bar{E}^2 + \bar{P}^2)^{1/2}] \tag{571}$$

$$= {}_1\bar{F}_1[- n' \; ; \; 2\bar{M}' + 1 \; ; \; \bar{g}/[a\gamma^2(\bar{\eta} + 1/2)](\bar{E}^2 + \bar{P}^2)^{1/2}]$$

This gives the eigenfunctions for the bound states of a particle trapped in a Coulomb-form of potential in energy-pressure space.

Combining equations (520) and (555) suggests that the energy density and pressure in a macroscopic system with a Coulomb-like attractive potential in energy-pressure space will have quantized values determined by

$$\langle \bar{r}_{\eta} \rangle = \langle (\bar{E}_{\eta}^2 + \bar{P}_{\eta}^2)^{1/2} \rangle = 4a\gamma^2/\bar{g} \; (\bar{\eta} + 1/2)^2 \tag{572}$$

or if only real values are considered

$$\langle r_n \rangle = \langle (E_n^2 + P_n^2)^{1/2} \rangle = 4a\gamma^2/g \ (n + 1/2)^2 \tag{573}$$

The minimum values of pressure and energy density occur in the ground state which has

$$\langle r_o \rangle = \langle (E_o^2 + P_o^2)^{1/2} \rangle = a\gamma^2/g \tag{574}$$

The energy density and pressure in this special bulk matter system can exist only with quantized values of $\langle r_n \rangle$ because time and dimension behave like wave functions in bulk matter. The bound states of time and dimensions form structures in energy-pressure space that have quantized energies given by equations (555) through (568) and quantized extension in energy density-pressure given by equations (572) through (574). These structures are the bulk matter analogs of the atomic structures of electrons in atoms. The quantized structures of time and dimension in energy density-pressure space may exist in bulk matter at high energy densities and pressures associated with the interiors of stars, planets and atomic nuclei.

The internal structures of stars, planets and atomic nuclei may be more complicated than conventional theory predicts because the energy density and pressure may under some conditions be quantized variables associated with the wave functions of time and dimension in bulk matter. The quantized structures may exist in the cores of ordinary stars as well as in compact objects like neutron stars and white dwarfs. These structures may also exist in the interiors of atomic nuclei where the density of nuclear matter varies rapidly with radial distance from the center. The calculation of thermonuclear energy generation rates will be affected by the nature of the time and dimension states of bulk matter in stellar interiors. Stellar interiors are composed of real gases with non-zero gauge parameters $\beta_\mathcal{E}$ and β_P .[17] The calculation of the nuclear reaction rates will be affected by the gauge parameters and the nature of the time and dimension states of the bulk matter in stellar interiors. The predicted rates will depend on whether time and dimension are coherent or incoherent and whether quantized structures of time and dimension exist within the energy density-pressure space of the interior of stars. Stars, planets and atomic nuclei may exhibit complex time and dimension structures.

9. CONCLUSION. A previously developed gauge theory of thermodynamics is extended to consider coherent as well as incoherent spacetime. The relativistic thermodynamic trace equation of the gauge theory of thermodynamics is then converted to an eigenvalue problem thereby producing the basic equation of quantum thermodynamics. From a previously developed gauge theory of time and dimension in bulk matter, a quantum theory of time and dimension is created in terms of first and second order differential eigenvalue equations in energy density-pressure space. The substructure of time and dimension is considered by introducing the concept of a time coherent boson called the chronon. Solutions to the time and dimension eigenvalue equations are considered and applied to a particle trapped in a time box and a dimension box in energy density-pressure space. The special case of a Coulomb-like potential in energy density-pressure space is examined. This form of potential is suggested because at high pressure and energy density the potential vanishes and the system exhibits

asymptotic freedom. Quantized time and dimension structures may exist in the interiors of stars, planets and atomic nuclei, and the reaction rates and geometrical structures of the nuclear and chemical processes in these objects may be affected by these time and dimension structures. The existence of quantum time and dimension structures in bulk matter implies that there are regions in energy density-pressure space where chemical and nuclear processes can be enhanced or depressed. This may have applications to the explanation of the formation of order and structure in non-equilibrium situations such as in the Belousov-Zhabotinskii reaction, and for the Turing structures.[24-26] Time and dimension structures may also occur in electronic devices that utilize high-T_c superconductors because the superconducting state of a high-T_c material is a coherent time state. In this case the structure is associated with the internal phase angles of time and dimension for a gas of coherent time Cooper electron pairs interacting with coherent time phonons of a crystal lattice.

ACKNOWLEDGEMENT

The author wishes to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. Lindsay, R. B. and Margenau, H., Foundations of Physics, John Wiley, New York, 1936.

2. Berry, M., Principles of Cosmology and Gravitation, Cambridge Univ. Press, New York, 1976.

3. Born, M., Einstein's Theory of Relativity, Dover, New York, 1962.

4. Pauli, W., Theory of Relativity, Pergamon, New York, 1958.

5. Misner, C. W., Thorne, K. S. and Wheeler, J. A., Gravitation, W. H. Freeman, San Francisco, 1973.

6. Yilmaz, H., Theory of Relativity and the Principles of Modern Physics, Blaisdell, New York, 1965.

7. Isham, C., "Quantum Gravity," article in, The New Physics, edited by Davies, P., Cambridge University Press, New York, 1989.

8. Shallis, M., On Time, Schocken Books, New York, 1983.

9. Fraser, J. T., Time and Mind, International Universities Press, Madison, Connecticut, 1989.

10. Priestley, J. B., Man & Time, Crescent Books, New York, 1989.

11. Szamosi, G., The Twin Dimensions, McGraw-Hill, New York, 1986.

12. Penrose, R. and Isham, C. J., editors, Quantum Concepts in Space and Time, Clarendon Press, Oxford, 1986.

13. Zeh, H. D., The Physical Basis of the Direction of Time, Springer-Verlag, New York, 1989.

14. Carlip, S., "Observables, Gauge Invariance, and Time in (2 + 1)-Dimensional Quantum Gravity," Phys. Rev. D, Vol. 42, 15 Oct. 1990.

15. Rovelli, C., "Quantum Mechanics Without Time: A Model," Phys. Rev. D, Vol. 42, 15 Oct. 1990.

16. Weiss, R. A., "Gauge Theory of Time," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990, p. 367.

17. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.

18. Weiss, R. A., Relativistic Thermodynamics, Exposition Press, New York, 1976.

19. Weiss, R. A., "Electromagnetism and Gravity," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990, p. 265.

20. Collatz, L., Eigenwertprobleme und ihre Numerische Behandlung, Chelsea, New York, 1948.

21. Corcoran, E., "Diminishing Dimensions," Scientific American , p.122, Nov. 1990.

22. Weiss, R. A., "Thermal Radiation of High-T_c Superconductors," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990, p. 399.

23. Morse, P. M. and Feshbach, H., Methods of Theoretical Physics, Vols 1 & 2, McGraw-Hill, New York, 1953.

24. Prigogine, I., From Being to Becoming, H. W. Freeman, New York, 1980.

25. Winfree, A. T., The Geometry of Biological Time, Springer-Verlag, New York, 1980.

26. Lengyel, I. and Epstein, I. R., "Modeling of Turing Structures in the Chlorite-Iodide-Malonic Acid-Starch Reaction System," Science, Vol. 251, p. 650, 8 Feb. 1991.

ULTRAFAST COHERENT HEAT ENGINES

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. This paper considers the possibility of developing ultrafast thermodynamic engine cycles that operate by the exchange of internal phase heat with the environment. These engines operate on the basis of the first and second laws of thermodynamics which are written in a form where the entropy and internal energy are complex numbers which rotate in an internal space during an ultrafast process. Several types of cyclic engines are considered in which the magnitudes of both the entropy and internal energy remain fixed during each portion of the cycle. The efficiencies are calculated for internal phase engine cycles of the type: Carnot, Otto, Diesel, Stirling, Ericsson and Brayton. These efficiencies are complex numbers whose real parts represent measured efficiencies that must always be less than unity. A brief discussion is given of the application of broken symmetry internal phase engine cycles to practical power sources and to the thermodynamic processes that occur in high speed computer memories. The possibility of developing high-T_c superconducting electron-pair engines is considered.

1. INTRODUCTION. Man has always searched for new power sources. The development of heat engines predated the science of thermodynamics. In fact the earliest steam turbine was developed in Alexandria by Hero in about 120 B.C., while the next development came only after the dark ages when Branca developed an impulse steam turbine in 1629.[1,2] During the period 1663-1700 the Marquis of Worcester and Savery developed a steam powered pumping machine. In 1690 Newcomen developed a steam powered walking beam engine. Around 1763 Watt developed the first modern steam engine. Although early steam turbines were developed in 1859 by Laval and by Parsons in 1884 it was not until the twentieth century that the steam turbine had commercial development. The internal combustion engine originated about 1690 when gunpowder was used as the fuel to drive a piston in a cylinder. The first patent for a gas engine was awarded in England in 1794, but the first practical gas engine was produced by Lenoir in 1860. In 1876 Otto developed a gas engine that had some commercial success. The gas turbine engine was developed in 1791 by Barber, and hot air engines were developed by Stirling and Ericsson in 1827. Since the beginning of the twentieth century liquid fuels such as gasoline and petroleum have replaced the gas engine. In 1892 Diesel used compressed air to make a practical engine that burned heavy oils. Leduc developed the jet engine in 1913.[2] All of these engines burn wood or fossil fuels and their operation pollutes the atmosphere. New power sources must be developed in order to reduce the consumption of fossil fuels. This paper considers the possibility of generating power by creating ultrafast heat engines that operate with coherent heat. Also the possibility of operating an engine in coherent spacetime is considered by using high-T_c superconducting Cooper electron pairs as a working substance for the engine.

The first law of thermodynamics was discovered by Mayer in 1842 and by

Joule in 1843 nearly two millenia after the first heat engine was conceived. This law states that mechanical work can be changed into an equivalent amount of heat and vice versa.[1-9] Combined with the second law of thermodynamics, which states that for a closed system the amount of entropy (disorder) increases or remains constant during a process, the first law of thermodynamics can be written as[2-10]

$$TdS = dU + PdV + Md\alpha \tag{1}$$

where T = absolute temperature, S = entropy, U = internal energy, P = pressure and V = volume of a fixed amount of material, M = generalized force and α = generalized coordinate. The combined first and second laws can also be written as the Gibbs-Helmholtz-Maxwell relations[2-10]

$$(\partial U/\partial V)_{T,\alpha} = T(\partial P/\partial T)_{V,\alpha} - P \qquad (\partial U/\partial \alpha)_{T,V} = T(\partial M/\partial T)_{V,\alpha} - M \tag{2}$$

Essentially all of thermodynamics can be derived from equations (1) and (2).

Recently it has been suggested that the thermodynamic functions such as pressure, internal energy and entropy have internal phase angles and must be represented as complex numbers.[11,12] Within this formalism equations (1) and (2) are written as[12]

$$Td\bar{S} = d\bar{U} + \bar{P}dV + \bar{M}d\alpha \tag{3}$$

$$(\partial \bar{U}/\partial V)_{T,\alpha} = T(\partial \bar{P}/\partial T)_{V,\alpha} - \bar{P} \qquad (\partial \bar{U}/\partial \alpha)_{T,V} = T(\partial \bar{M}/\partial T)_{V,\alpha} - \bar{M} \tag{4}$$

where α and V are taken to be independent variables and where $\bar{S}$, $\bar{U}$, $\bar{P}$ and $\bar{M}$ = complex number entropy, internal energy, pressure and generalized force respectively, and α = generalized extensive variable. Equations (3) and (4) lead to complex number renormalization group equations which can be used to determine thermodynamic functions such as the Grüneisen function.[12] The entropy, internal energy, pressure and generalized force that appear in equation (3) can be written as[12]

$$\bar{S} = Se^{j\theta_S} \qquad \bar{U} = Ue^{j\theta_U} \tag{5}$$

$$\bar{P} = Pe^{j\theta_P} \qquad \bar{M} = Me^{j\theta_M} \tag{6}$$

$$|\bar{S}| = S \qquad |\bar{U}| = U \qquad |\bar{P}| = P \qquad |\bar{M}| = M \tag{7}$$

where S, U, P and M = magnitudes of the entropy, internal energy, pressure and generalized force respectively, and θ_S, θ_U, θ_P and θ_M = internal phase angles of the entropy, internal energy, pressure and generalized force respectively. The measured thermodynamic functions are

$$S_m = S\cos\theta_S \qquad U_m = U\cos\theta_U \tag{8}$$

$$P_m = P\cos\theta_P \qquad M_m = M\cos\theta_M \tag{9}$$

For the special case of coherent heat engine cycles where the entropy and

internal energy vectors are rotated in an internal space with the magnitudes S and U held fixed during a thermodynamic process, equation (3) becomes for incoherent space[12]

$$j\bar{T}\bar{S}d\theta_S = j\bar{U}d\theta_U + \bar{P}dV + \bar{M}d\alpha \tag{10}$$

Equation (10) represents the first and second laws of thermodynamics for the special case of an ultrafast process.[12] From equation (10) it is easy to show that the pressure associated with a transfer of internal phase of entropy and internal energy at constant S and U is given by[12]

$$\bar{P} = j[T\bar{S}(d\theta_S/dV)_{U,S} - \bar{U}(d\theta_U/dV)_{U,S}] - \bar{M}(d\alpha/dV)_{U,S} \tag{11}$$

$$= j[T\bar{\mathcal{S}}(Vd\theta_S/dV)_{U,S} - \bar{E}(Vd\theta_U/dV)_{U,S}] - \bar{\mathbb{M}}(Vd\alpha/dV)_{U,S}$$

where $\alpha = \alpha(V,T)$ and where

$$\bar{\mathcal{S}} = \mathcal{S}e^{j\theta S} = \bar{S}/V \qquad \bar{E} = Ee^{j\theta U} = \bar{U}/V \qquad \bar{\mathbb{M}} = \mathbb{M}e^{j\theta M} = \bar{M}/V \tag{12}$$

and where $\bar{\mathcal{S}}$ = incoherent average entropy density, $\bar{E}$ = incoherent average energy density, and $\bar{\mathbb{M}}$ = incoherent average generalized force density. From equations (7) and (11) it follows that approximately[12]

$$P \sim TS(d\theta_S/dV)_{U,S} - U(d\theta_U/dV)_{U,S} - M(d\alpha/dV)_{U,S} \tag{13}$$

$$\sim T\mathcal{S}(Vd\theta_S/dV)_{U,S} - E(Vd\theta_U/dV)_{U,S} - \mathbb{M}(Vd\alpha/dV)_{U,S}$$

$$\theta_P \sim \theta_M \sim \theta_U + \pi/2 \sim \theta_S + \pi/2 \tag{14}$$

For this case equation (4) becomes with S and U fixed

$$j\bar{U}(\partial\theta_U/\partial V)_{T,\alpha} = T(\partial\bar{P}/\partial T)_{V,\alpha} - \bar{P} \tag{14A}$$

$$j\bar{U}(\partial\theta_U/\partial\alpha)_{T,V} = T(\partial\bar{M}/\partial T)_{V,\alpha} - \bar{M} \tag{14B}$$

if α and V are taken to be independent variables. There is a pressure associated with internal phase changing thermodynamic processes and according to equation (14) the pressure is perpendicular to the internal energy and entropy in internal space. This is analogous to the classical Magnus effect in hydrodynamics.[13-16]

For the adiabatic case where $d\bar{S} = 0$ (or $dS = 0$ and $d\theta_S = 0$) as well as $dU = 0$ it follows from equations (11), (13) and (14) that[12]

$$\bar{P} = -j\bar{U}(d\theta_U/dV)_{\bar{S},U} - \bar{M}(d\alpha/dV)_{\bar{S},U} \tag{15}$$

$$P \sim -U|(d\theta_U/dV)_{\bar{S},U}| - M|(d\alpha/dV)_{\bar{S},U}| \tag{16}$$

$$\theta_P \sim \theta_M \sim \theta_U + \pi/2 \tag{17}$$

It should be mentioned that it is possible to have thermodynamic processes that have $\bar{U}$ and S fixed. For this case equation (11) gives

$$\bar{P} = j\bar{T}\bar{S}(d\theta_S/dV)_{\bar{U},S} - \bar{M}(d\alpha/dV)_{\bar{U},S} \tag{18A}$$

$$P \sim TS\left|(d\theta_S/dV)_{\bar{U},S}\right| - M\left|(d\alpha/dV)_{\bar{U},S}\right| \tag{18B}$$

$$\theta_P \sim \theta_M \sim \theta_S + \pi/2 \tag{18C}$$

where in general $\alpha = \alpha(V,T)$. The exact values of P and θ_P can only be obtained by obtaining the real and imaginary parts of equations (15) and (18A).

As pointed out in Reference 12, the inclusion of the generalized force terms in equation (10) is a logical necessity for internal phase processes. In some calculations it is possible as a first approximation to ignore the generalized forces but in other cases, such as the constant volume process which occurs in the cases of the Otto, Diesel and Stirling internal phase cycles, the generalized forces must be included as a logical necessity to have the possibility of thermodynamic cycles with U and S fixed.

Each path segment of every coherent engine cycle considered in this paper has S and U as constants. Each path segment has the same values of S and U, in other words S and U are fixed for the entire internal phase cycle. Thus if $\bar{S}_{bc}$ and $\bar{S}_{da}$ are two constant values of the complex number entropy along the path segments bc and da respectively, then (see Figure 1)

$$\bar{S}_{bc} = S\exp(j\theta_S^{bc}) \tag{19}$$

$$\bar{S}_{da} = S\exp(j\theta_S^{da}) \tag{20}$$

and only the internal phases differ on the two path segments. For the special case of coherent heat engines with coherent spacetime in the working chamber (as in the case when the working substance is a gas of high-T_c superconducting electron pairs) each segment of the engine cycle has the magnitude of the volume held fixed, and the fixed volume magnitude is the same for the entire internal phase cycle. Therefore if $\bar{V}_{ab}$ and $\bar{V}_{cd}$ are two constant values of the complex number volume along the path segments ab and cd respectively (see Figure 2) then

$$\bar{V}_{ab} = V\exp(j\theta_V^{ab}) \qquad \bar{V}_{cd} = V\exp(j\theta_V^{cd}) \tag{21}$$

and again only the internal phases are different on the path segments ab and cd. This is not the case with pressure. Thus if two path segments bc and da have constant pressures $\bar{P}_{bc}$ and $\bar{P}_{da}$ respectively (see Figure 4) then

$$\bar{P}_{bc} = P_{bc}\exp(j\theta_P^{bc}) \tag{22}$$

$$\bar{P}_{da} = P_{da}\exp(j\theta_P^{da}) \tag{23}$$

and the magnitudes and the internal phase angles are different for each path

segment. Note that the pressure does not undergo a pure rotation during an ultrafast process as can be seen from equation (14A).

This paper calculates the thermodynamic efficiencies of several ultrafast coherent heat engine cycles that involve the transfer of internal phase heat at constant U and S. Each engine cycle described in this paper operates on heat energy that is introduced into the engine in the form of internal phase heat. Each cycle converts a portion of this internal phase heat into a net usable work and deposits the remaining internal phase heat into the environment in accordance with the first and second laws of thermodynamics. For this reason the efficiency of an ultrafast internal phase heat engine must be less than unity as in the case of standard engine cycles. During the cycle the internal energy of the working substance rotates and changes its internal phase angle, but being a state function the internal energy must return to its initial value after a complete cycle. The efficiencies of the engine cycles are evaluated by calculating the ratio of the net complex number work to the value of the complex number heat introduced into the engine during each cycle. In general the efficiencies are complex numbers whose real parts are the measured efficiencies. The efficiencies are evaluated for several practical and historical engine cycles. Only closed thermodynamic cycles of a working substance are considered, and only changes of the internal phase angles of entropy and energy are considered in this paper. Sections 3 through 8 deal respectively with the internal phase cycles of the Carnot, Otto, Diesel, Ericsson, Stirling, and Brayton engines. By considering the case of coherent spacetime, the corresponding high-T_c superconducting electron pair engine for each of the above mentioned cycles is treated.

2. BROKEN SYMMETRY THERMODYNAMICS. This section summarizes the calculation of pressure, heat exchanged, and work done for partially coherent and totally coherent states of thermodynamic systems and for incoherent and coherent states of the spacetime in which the working substance of an engine is located. Engine cycle calculations done in this paper are only for the case of coherent thermodynamics (ultrafast processes) combined with incoherent spacetime (ordinary substances) and coherent spacetime (high-T_c superconducting electron pairs). Incoherent spacetime is associated with ordinary matter, coherent spacetime is associated with the superconducting state of high-T_c substances, and partially coherent spacetime is associated with the normal state of high-T_c materials.

For the general case of a thermodynamic system with broken symmetry thermodynamic functions and broken symmetry spacetime the pressure is given by[12]

$$T d\bar{S} = d\bar{U} + \bar{\bar{P}} d\bar{V} + \bar{\bar{M}} d\bar{\alpha} \tag{24}$$

$$= d\bar{U} + \bar{P} |d\bar{V}| + \bar{M} |d\bar{\alpha}|$$

where

$$\bar{V} = V \exp(j\theta_V) \qquad \bar{\alpha} = \alpha \exp(j\theta_\alpha) \tag{25}$$

$$d\bar{V} = \sec \beta_{VV} \, dV \exp[j(\theta_V + \beta_{VV})] \tag{26}$$

$$= \csc \beta_{VV} \, V d\theta_V \exp[j(\theta_V + \beta_{VV})]$$

$$|d\bar{V}| = \sec \beta_{VV}\, dV = \csc \beta_{VV}\, V d\theta_V \tag{27}$$

$$d\bar{\alpha} = \sec \beta_{\alpha\alpha}\, d\alpha \exp[j(\theta_\alpha + \beta_{\alpha\alpha})] \tag{28}$$

$$= \csc \beta_{\alpha\alpha}\, \alpha d\theta_\alpha \exp[j(\theta_\alpha + \beta_{\alpha\alpha})]$$

$$|d\bar{\alpha}| = \sec \beta_{\alpha\alpha}\, d\alpha = \csc \beta_{\alpha\alpha}\, \alpha d\theta_\alpha \tag{29}$$

$$\tan \beta_{VV} = V\partial\theta_V/\partial V \tag{30}$$

$$\tan \beta_{\alpha\alpha} = \alpha\partial\theta_\alpha/\partial\alpha \tag{31}$$

From equation (5) it follows that

$$d\bar{U} = \sec \beta_{UU}\, dU \exp[j(\theta_U + \beta_{UU})] \tag{32}$$

$$= \csc \beta_{UU}\, U d\theta_U \exp[j(\theta_U + \beta_{UU})]$$

$$d\bar{S} = \sec \beta_{SS}\, dS \exp[j(\theta_S + \beta_{SS})] \tag{33}$$

$$= \csc \beta_{SS}\, S d\theta_S \exp[j(\theta_S + \beta_{SS})]$$

where

$$\tan \beta_{UU} = U\partial\theta_U/\partial U \tag{34}$$

$$\tan \beta_{SS} = S\partial\theta_S/\partial S \tag{35}$$

The measured thermodynamic functions are given by equations (8) and (9) while the measured extensive variables are obtained from equation (25) as

$$V_m = V \cos \theta_V \qquad \alpha_m = \alpha \cos \theta_\alpha \tag{36}$$

From equation (36) it is clear that variation of the measured volume occurs in a coherent heat engine cycle for both incoherent spacetime where $\theta_V = 0$ and V is a variable, and for coherent spacetime where V = constant and θ_V is a variable. From equation (24) it follows that for α and V independent of each other

$$\cos \beta_{VV}\, \partial\bar{U}/\partial V = T\partial\bar{P}/\partial T - \bar{P} \qquad \cos \beta_{\alpha\alpha}\, \partial\bar{U}/\partial\alpha = T\partial\bar{M}/\partial T - \bar{M} \tag{36A}$$

$$\sin \beta_{VV}\, 1/V\partial\bar{U}/\partial\theta_V = T\partial\bar{P}/\partial T - \bar{P} \qquad \sin \beta_{\alpha\alpha}\, 1/\alpha\partial\bar{U}/\partial\theta_\alpha = T\partial\bar{M}/\partial T - \bar{M} \tag{36B}$$

which are the Gibbs-Helmholtz-Maxwell equations for broken spacetime symmetry. For coherent spacetime with $\beta_{VV} = \pi/2$ and $\beta_{\alpha\alpha} = \pi/2$ equation (36B) becomes

$$1/V\ \partial\bar{U}/\partial\theta_V = T\partial\bar{P}/\partial T - \bar{P} \qquad 1/\alpha\ \partial\bar{U}/\partial\theta_\alpha = T\partial\bar{M}/\partial T - \bar{M} \tag{36C}$$

and for coherent internal energy with $\beta_{UU} = \pi/2$ it follows from equations (32) and (36C) that

$$j\bar{U}/V\ \partial\theta_U/\partial\theta_V = T\partial\bar{P}/\partial T - \bar{P} \qquad j\bar{U}/\alpha\ \partial\theta_U/\partial\theta_\alpha = T\partial\bar{M}/\partial T - \bar{M} \tag{36D}$$

or

$$\bar{U}/\bar{V}\ \partial\theta_U/\partial\theta_V = T\partial\bar{\bar{P}}/\partial T - \bar{\bar{P}} \qquad \bar{U}/\bar{\alpha}\ \partial\theta_U/\partial\theta_\alpha = T\partial\bar{\bar{M}}/\partial T - \bar{\bar{M}} \tag{36E}$$

The first integral of the differential form of the first and second laws of thermodynamics given in equation (24) is the following path dependent equation

$$\bar{W}_{ab} = \bar{Q}_{ab} - (\bar{U}_b - \bar{U}_a) - \bar{\omega}_{ab} \tag{37}$$

where

$$\bar{W}_{ab} = \int_a^b \bar{P}|d\bar{V}| \qquad \bar{\omega}_{ab} = \int_a^b \bar{M}|d\bar{\alpha}| \tag{38}$$

$$\bar{U}_b - \bar{U}_a = \int_a^b d\bar{U} \qquad \bar{Q}_{ab} = \int_a^b T d\bar{S} \tag{39}$$

where ab refers to a path for a specified thermodynamic process, $\bar{W}_{ab}$ = work done by pressure, $\bar{\omega}_{ab}$ = work done by the generalized force, $\bar{Q}_{ab}$ = heat transferred during the process, and $\bar{U}_b - \bar{U}_a$ = change in the internal energy. For a cycle a b c d a three equations in addition to equation (37) are required for a description of the thermodynamic cycle

$$\bar{W}_{bc} = \bar{Q}_{bc} - (\bar{U}_c - \bar{U}_b) - \bar{\omega}_{bc} \tag{40}$$

$$\bar{W}_{cd} = \bar{Q}_{cd} - (\bar{U}_d - \bar{U}_c) - \bar{\omega}_{cd} \tag{41}$$

$$\bar{W}_{da} = \bar{Q}_{da} - (\bar{U}_a - \bar{U}_d) - \bar{\omega}_{da} \tag{42}$$

Now three special cases of broken symmetry thermodynamics will be considered.

A. Partially Coherent Thermodynamic State and Partially Coherent Spacetime.

This case corresponds to a moderately fast process in the normal state of a high-T_c superconductor. For this situation the pressure is given by

$$T d\bar{S} = d\bar{U} + \bar{\bar{P}} d\bar{V} + \bar{\bar{M}} d\bar{\alpha} \tag{43}$$

$$= d\bar{U} + \bar{P}|d\bar{V}| + \bar{M}|d\bar{\alpha}|$$

For a thermodynamic path segment $\mu\nu$, where $\mu\nu = ab, bc, cd, da$, the heat transferred, change in internal energy, and the work done by the pressure and the generalized force are respectively given by

$$\bar{Q}_{\mu\nu} = \int_{\mu}^{\nu} T d\bar{S} = \int_{\mu}^{\nu} T \sec \beta_{SS} \exp[j(\theta_S + \beta_{SS})] dS \tag{44}$$

$$= \int_{\mu}^{\nu} T \csc \beta_{SS} \exp[j(\theta_S + \beta_{SS})] S d\theta_S$$

$$\bar{U}_{\mu\nu} = \bar{U}_{\nu} - \bar{U}_{\mu} = \int_{\mu}^{\nu} d\bar{U} = \int_{\mu}^{\nu} \sec \beta_{UU} \exp[j(\theta_U + \beta_{UU})] dU \tag{45}$$

$$= \int_{\mu}^{\nu} \csc \beta_{UU} \exp[j(\theta_U + \beta_{UU})] U d\theta_U$$

$$\bar{W}_{\mu\nu} = \int_{\mu}^{\nu} \bar{P} |d\bar{V}| = \int_{\mu}^{\nu} \bar{P} \sec \beta_{VV} \, dV = \int_{\mu}^{\nu} \bar{P} \csc \beta_{VV} \, V d\theta_V \tag{46}$$

$$\bar{\omega}_{\mu\nu} = \int_{\mu}^{\nu} \bar{M} |d\bar{\alpha}| = \int_{\mu}^{\nu} \bar{M} \sec \beta_{\alpha\alpha} \, d\alpha = \int_{\mu}^{\nu} \bar{M} \csc \beta_{\alpha\alpha} \, \alpha d\theta_{\alpha} \tag{47}$$

Case A is included only for completeness, and no engine cycles with the partial coherence of spacetime and the partial coherence of thermodynamic functions are considered in this paper.

B. Coherent Thermodynamics and Incoherent Spacetime.

This case corresponds to an ultrafast process in an ordinary material. For this case S = constant, U = constant, θ_V = constant, θ_α = constant and

$$\beta_{UU} = \pi/2 \qquad \beta_{SS} = \pi/2 \qquad \beta_{VV} = 0 \qquad \beta_{\alpha\alpha} = 0 \tag{48}$$

Then the combined first and second laws of thermodynamics is written as

$$jT\bar{S} d\theta_S = j\bar{U} d\theta_U + \bar{\bar{P}} d\bar{V} + \bar{\bar{M}} d\bar{\alpha} \tag{49}$$

$$= j\bar{U} d\theta_U + \bar{P} dV + \bar{M} d\alpha$$

and the pressure is given as follows

$$\bar{P} = Ts_{cth} \, e^{j(\theta_S + \pi/2)} - e_{cth} \, e^{j(\theta_U + \pi/2)} - Ma_{inc} \, e^{j\theta_M} \tag{50}$$

where cth refers to coherent thermodynamics, inc = refers to incoherent space and incoherent generalized coordinate, and where s_{cth}, e_{cth} and a_{inc} are defined by

$$s_{cth} = S\ \partial\theta_S/\partial V \tag{51}$$

$$e_{cth} = U\ \partial\theta_U/\partial V \tag{52}$$

$$a_{inc} = d\alpha/dV \tag{53}$$

From equation (50) the following approximations are valid

$$P \sim Ts_{cth} - e_{cth} - Ma_{inc} \tag{53A}$$

$$\theta_P \sim \theta_M \sim \theta_S + \pi/2 \sim \theta_U + \pi/2 \tag{53B}$$

Equation (50) is just the pressure derived in equation (11).

The heat transferred, change in internal energy, and the work done by the pressure and generalized force on a thermodynamic path segment $\mu\nu$ = ab , bc , cd , da is given repsectively by

$$\bar{Q}_{\mu\nu} = \int_{\mu}^{\nu} T d\bar{S} = jS \int_{\mu}^{\nu} T \exp(j\theta_S) d\theta_S \tag{54}$$

$$\bar{U}_{\mu\nu} = \bar{U}_{\nu} - \bar{U}_{\mu} = \int_{\mu}^{\nu} d\bar{U} = jU \int_{\mu}^{\nu} \exp(j\theta_U) d\theta_U \tag{55}$$

$$= U[\exp(j\theta_{U\nu}) - \exp(j\theta_{U\mu})]$$

$$\bar{W}_{\mu\nu} = W_{\mu\nu} \exp(j\theta_W^{\mu\nu}) = \int_{\mu}^{\nu} \bar{P} dV \tag{56A}$$

$$\bar{\omega}_{\mu\nu} = \omega_{\mu\nu} \exp(j\theta_\omega^{\mu\nu}) = \int_{\mu}^{\nu} \bar{M} d\alpha \tag{56B}$$

From equations (56A) and (56B) it follows that

$$W_{\mu\nu} \cos \theta_W^{\mu\nu} = \int_{\mu}^{\nu} P \cos \theta_P\ dV \tag{56C}$$

$$W_{\mu\nu} \sin \theta_W^{\mu\nu} = \int_{\mu}^{\nu} P \sin \theta_P\ dV \tag{56D}$$

$$\omega_{\mu\nu} \cos \theta_\omega^{\mu\nu} = \int_{\mu}^{\nu} M \cos \theta_M\ d\alpha \tag{56E}$$

$$\omega_{\mu\nu} \sin \theta_\omega^{\mu\nu} = \int_{\mu}^{\nu} M \sin \theta_M\ d\alpha \tag{56F}$$

The real and imaginary parts of the work elements enter into the calculation of engine efficiency.

Using the mean value theorem for integrals and some appropriately defined constant values for the pressure and generalized force in the interval $\mu\nu$ allows equations (56C) through (56F) to be written as

$$W_{\mu\nu} \cos \theta_W^{\mu\nu} = P_{\mu\nu}(V_\nu - V_\mu)\langle \cos \theta_P \rangle_{\mu\nu} \tag{56G}$$

$$W_{\mu\nu} \sin \theta_W^{\mu\nu} = P_{\mu\nu}(V_\nu - V_\mu)\langle \sin \theta_P \rangle_{\mu\nu} \tag{56H}$$

$$\omega_{\mu\nu} \cos \theta_\omega^{\mu\nu} = M_{\mu\nu}(\alpha_\nu - \alpha_\mu)\langle \cos \theta_M \rangle_{\mu\nu} \tag{56I}$$

$$\omega_{\mu\nu} \sin \theta_\omega^{\mu\nu} = M_{\mu\nu}(\alpha_\nu - \alpha_\mu)\langle \sin \theta_M \rangle_{\mu\nu} \tag{56J}$$

where $P_{\mu\nu}$ and $M_{\mu\nu}$ are constants defined for the path segment $\mu\nu$ by equations (56G) through (56J). From equation (53B) it follows that

$$\langle \cos \theta_P \rangle_{\mu\nu} \sim -\langle \sin \theta_U \rangle_{\mu\nu} = (\theta_{U\nu} - \theta_{U\mu})^{-1}(\cos \theta_{U\nu} - \cos \theta_{U\mu}) \tag{56K}$$

$$\sim -\langle \sin \theta_S \rangle_{\mu\nu} = (\theta_{S\nu} - \theta_{S\mu})^{-1}(\cos \theta_{S\nu} - \cos \theta_{S\mu})$$

$$\langle \sin \theta_P \rangle_{\mu\nu} \sim \langle \cos \theta_U \rangle_{\mu\nu} = (\theta_{U\nu} - \theta_{U\mu})^{-1}(\sin \theta_{U\nu} - \sin \theta_{U\mu}) \tag{56L}$$

$$\sim \langle \cos \theta_S \rangle_{\mu\nu} = (\theta_{S\nu} - \theta_{S\mu})^{-1}(\sin \theta_{S\nu} - \sin \theta_{S\mu})$$

Because $\theta_M \sim \theta_P$ the quantities $\langle \cos \theta_M \rangle_{\mu\nu}$ and $\langle \sin \theta_M \rangle_{\mu\nu}$ are given by equations (56K) and (56L) respectively.

For small values of θ_U and θ_S it follows from equations (56G) through (56L) that

$$W_{\mu\nu} \cos \theta_W^{\mu\nu} \sim -P_{\mu\nu}(V_\nu - V_\mu)(\theta_{U\nu} + \theta_{U\mu})/2 \tag{56M}$$

$$\sim -P_{\mu\nu}(V_\nu - V_\mu)(\theta_{S\nu} + \theta_{S\mu})/2$$

$$W_{\mu\nu} \sin \theta_W^{\mu\nu} \sim P_{\mu\nu}(V_\nu - V_\mu) \tag{56N}$$

$$\omega_{\mu\nu} \cos \theta_\omega^{\mu\nu} \sim -M_{\mu\nu}(\alpha_\nu - \alpha_\mu)(\theta_{U\nu} + \theta_{U\mu})/2 \tag{56O}$$

$$\sim -M_{\mu\nu}(\alpha_\nu - \alpha_\mu)(\theta_{S\nu} + \theta_{S\mu})/2$$

$$\omega_{\mu\nu} \sin \theta_\omega^{\mu\nu} \sim M_{\mu\nu}(\alpha_\nu - \alpha_\mu) \tag{56P}$$

These expressions will be used in Sections 3 through 8 to evaluate the efficiencies of ultrafast engine cycles in the incoherent spacetime of an ordinary working substance.

C. Coherent Thermodynamic Processes in Coherent Spacetime.

This case corresponds to an ultrafast process in the superconducting phase of a high-T_c compound, and is described by S = constant, U = constant, V = constant, α = constant and

$$\beta_{UU} = \pi/2 \qquad \beta_{SS} = \pi/2 \qquad \beta_{VV} = \pi/2 \qquad \beta_{\alpha\alpha} = \pi/2 \tag{57}$$

This gives the following form for the combined first and second laws of thermodynamics

$$j\bar{T}\bar{S}d\theta_S = j\bar{U}d\theta_U + j\bar{\bar{P}}\bar{V}d\theta_V + j\bar{\bar{M}}\bar{\alpha}d\theta_\alpha \tag{58}$$

$$= j\bar{U}d\theta_U + \bar{P}Vd\theta_V + \bar{M}\alpha d\theta_\alpha \tag{59}$$

The pressure is then given by

$$\bar{P} = Ts_{tc}\, e^{j(\theta_S+\pi/2)} - e_{tc}\, e^{j(\theta_U+\pi/2)} - Ma_{tc}\, e^{j\theta_M} \tag{60}$$

where

$$s_{tc} = S/V\ \partial\theta_S/\partial\theta_V \tag{61}$$

$$e_{tc} = U/V\ \partial\theta_U/\partial\theta_V \tag{62}$$

$$a_{tc} = \alpha/V\ \partial\theta_\alpha/\partial\theta_V \tag{63}$$

From equation (60) the following approximations are obtained

$$P = Ts_{tc} - e_{tc} - Ma_{tc} \tag{63A}$$

$$\theta_P \sim \theta_M \sim \theta_S + \pi/2 \sim \theta_U + \pi/2 \tag{63B}$$

For the thermodynamic path $\mu\nu$ = ab , bc , cd , da it follows that for coherent spacetime and coherent thermodynamics

$$\bar{Q}_{\mu\nu} = jS \int_\mu^\nu T \exp(j\theta_S)d\theta_S \tag{64}$$

$$\bar{U}_{\mu\nu} = \bar{U}_\nu - \bar{U}_\mu = \int_\mu^\nu d\bar{U} = jU \int_\mu^\nu \exp(j\theta_U)d\theta_U \tag{65}$$

$$= U[\exp(j\theta_{U_\nu}) - \exp(j\theta_{U_\mu})]$$

$$\bar{W}_{\mu\nu} = W_{\mu\nu} \exp(j\theta_W^{\mu\nu}) = \int_\mu^\nu \bar{\bar{P}} d\bar{V} = \int_\mu^\nu \bar{P} |d\bar{V}| \tag{66}$$

$$= V \int_\mu^\nu \bar{P} d\theta_V = V \int_\mu^\nu P \exp(j\theta_P) d\theta_V$$

$$\omega_{\mu\nu} = \omega_{\mu\nu} \exp(j\theta_\omega^{\mu\nu}) = \int_\mu^\nu \bar{\bar{M}} d\bar{\alpha} = \int_\mu^\nu \bar{M} |d\bar{\alpha}| \tag{67}$$

$$= \alpha \int_\mu^\nu \bar{M} d\theta_\alpha = \alpha \int_\mu^\nu M \exp(j\theta_M) d\theta_\alpha$$

These are the basic elements of coherent thermodynamic processes in coherent spacetime.

From equations (66), (67) and (63B) it follows for coherent thermodynamics and coherent spacetime that

$$W_{\mu\nu} \cos \theta_W^{\mu\nu} = V \int_\mu^\nu P \cos \theta_P \, d\theta_V \tag{67A}$$

$$= P_{\mu\nu} V(\theta_{V\nu} - \theta_{V\mu}) \langle \cos \theta_P \rangle_{\mu\nu}$$

$$\sim - P_{\mu\nu} V(\theta_{V\nu} - \theta_{V\mu}) \langle \sin \theta_U \rangle_{\mu\nu}$$

$$\sim - P_{\mu\nu} V(\theta_{V\nu} - \theta_{V\mu}) \langle \sin \theta_S \rangle_{\mu\nu}$$

$$W_{\mu\nu} \sin \theta_W^{\mu\nu} = V \int_\mu^\nu P \sin \theta_P \, d\theta_V \tag{67B}$$

$$= P_{\mu\nu} V(\theta_{V\nu} - \theta_{V\mu}) \langle \sin \theta_P \rangle_{\mu\nu}$$

$$\sim P_{\mu\nu} V(\theta_{V\nu} - \theta_{V\mu}) \langle \cos \theta_U \rangle_{\mu\nu}$$

$$\sim P_{\mu\nu} V(\theta_{V\nu} - \theta_{V\mu}) \langle \cos \theta_S \rangle_{\mu\nu}$$

$$\omega_{\mu\nu} \cos \theta_\omega^{\mu\nu} = \alpha \int_\mu^\nu M \cos \theta_M \, d\theta_\alpha \tag{67C}$$

$$= M_{\mu\nu} \alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu}) \langle \cos \theta_M \rangle_{\mu\nu}$$

$$\sim - M_{\mu\nu} \alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu}) \langle \sin \theta_U \rangle_{\mu\nu}$$

$$\sim - M_{\mu\nu} \alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu}) \langle \sin \theta_S \rangle_{\mu\nu}$$

$$\omega_{\mu\nu} \sin \theta_{\omega}^{\mu\nu} = \alpha \int_{\mu}^{\nu} M \sin \theta_M \, d\theta_\alpha \tag{67D}$$

$$= M_{\mu\nu}\alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu})\langle \sin \theta_M \rangle_{\mu\nu}$$

$$\sim M_{\mu\nu}\alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu})\langle \cos \theta_U \rangle_{\mu\nu}$$

$$\sim M_{\mu\nu}\alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu})\langle \cos \theta_S \rangle_{\mu\nu}$$

For small values of θ_U and θ_S it follows from equations (67A) through (67D) for coherent spacetime that

$$W_{\mu\nu} \cos \theta_W^{\mu\nu} \sim - P_{\mu\nu}V(\theta_{V\nu} - \theta_{V\mu})(\theta_{U\nu} + \theta_{U\mu})/2 \tag{67E}$$

$$\sim - P_{\mu\nu}V(\theta_{V\nu} - \theta_{V\mu})(\theta_{S\nu} + \theta_{S\mu})/2$$

$$W_{\mu\nu} \sin \theta_W^{\mu\nu} \sim P_{\mu\nu}V(\theta_{V\nu} - \theta_{V\mu}) \tag{67F}$$

$$\omega_{\mu\nu} \cos \theta_{\omega}^{\mu\nu} \sim - M_{\mu\nu}\alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu})(\theta_{U\nu} + \theta_{U\mu})/2 \tag{67G}$$

$$\sim - M_{\mu\nu}\alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu})(\theta_{S\nu} + \theta_{S\mu})/2$$

$$\omega_{\mu\nu} \sin \theta_{\omega}^{\mu\nu} \sim M_{\mu\nu}\alpha(\theta_{\alpha\nu} - \theta_{\alpha\mu}) \tag{67H}$$

These equations will be used in Sections 3 through 8 to calculate the efficiencies of ultrafast high-T_c superconducting electron pair engines.

3. ULTRAFAST CARNOT ENGINE. Carnot realized that the operation of any thermodynamic engine required the transfer of heat from a hot reservoir to a cold reservoir, and that the amount of heat transformed into work was proportional to the difference in the temperatures of the source and sink of heat.[2-9] The Carnot cycle, as given by Kelvin, is represented in Figure 1a and consists of four distinct processes, a → b isothermal expansion, b → c adiabatic expansion, c → d isothermal compression, and d → a adiabatic compression.[2-9] The well known expression for the efficiency of the conventional Carnot cycle is given by[1-8]

$$\eta = (Q_{ab} - Q_{cd})/Q_{ab} = (T_{ab} - T_{cd})/T_{ab} = 1 - T_{cd}/T_{ab} \tag{68}$$

where η = efficiency, T_{ab} = temperature of hot reservoir and T_{cd} = temperature of cold reservoir. The result in equation (68) is most easily derived using the ideal gas as a working substance, but in fact equation (68) is universally true for all working substances.[2-9] Although the Carnot engine has maximum efficiency, it is not practical because the mean operating pressure is low and the cycle cannot be applied to vapors. This section considers the ultrafast internal phase Carnot engine cycle.

A. Internal Phase Carnot Engine for Incoherent Spacetime.

Consider now the ultrafast internal phase cycle for the Carnot engine with ordinary matter as a working substance (for incoherent spacetime Case B of Section 2) that is shown in Figure 1b. The variables and fixed quantities for the various path segments of Figure 1b are

	Variables	Constants	
path ab	θ_S , θ_U , θ_P , P , V , α	S , U , T_{ab}	(69)
path bc	θ_U , θ_P , P , V , T , α	S , U , θ_S^{bc}	(70)
path cd	θ_S , θ_U , θ_P , P , V , α	S , U , T_{cd}	(71)
path da	θ_U , θ_P , P , V , T , α	S , U , θ_S^{da}	(72)

From equation (54) it follows that the heat transfers for each path segment are written as

$$\bar{Q}_{ab} = jT_{ab}S\int_a^b \exp(j\theta_S)d\theta_S = T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] \tag{73}$$

$$\bar{Q}_{bc} = 0 \tag{74}$$

$$\bar{Q}_{cd} = jT_{cd}S\int_c^d \exp(j\theta_S)d\theta_S = T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] \tag{75}$$

$$\bar{Q}_{da} = 0 \tag{76}$$

where $\bar{Q}_{ab} = -\bar{Q}_{ba}$, $\bar{Q}_{cd} = -\bar{Q}_{dc}$ and where because $\bar{S}_{bc}$ = constant and $\bar{S}_{da}$ = constant it follows that

$$\theta_{Sb} = \theta_{Sc} = \theta_S^{bc} \tag{77}$$

$$\theta_{Sa} = \theta_{Sd} = \theta_S^{da} \tag{78}$$

The work elements are given by equations (49) through (56) as

$$\bar{W}_{ab} = \int_a^b \bar{P}dV = \bar{Q}_{ab} - jU\int_a^b \exp(j\theta_U)d\theta_U - \int_a^b \bar{M}d\alpha \tag{79}$$

$$= T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] - U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] - \bar{\omega}_{ab}$$

$$\bar{W}_{bc} = \int_b^c \bar{P}dV = -jU\int_b^c \exp(j\theta_U)d\theta_U - \int_b^c \bar{M}d\alpha \tag{80}$$

$$= -U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ub})] - \bar{\omega}_{bc}$$

$$\bar{W}_{cd} = \int_c^d \bar{P}dV = \bar{Q}_{cd} - jU \int_c^d \exp(j\theta_U)d\theta_U - \int_c^d \bar{M}d\alpha \tag{81}$$

$$= T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] - U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] - \bar{\omega}_{cd}$$

$$\bar{W}_{da} = \int_d^a \bar{P}dV = - jU \int_d^a \exp(j\theta_U)d\theta_U - \int_d^a \bar{M}d\alpha \tag{82}$$

$$= - U[\exp(j\theta_{Ua}) - \exp(j\theta_{Ud})] - \bar{\omega}_{da}$$

where the works $\bar{W}_{\mu\nu}$ and the generalized works $\bar{\omega}_{\mu\nu}$ are given in equation (56) for incoherent spacetime. The pressure associated with an ultrafast thermodynamic process in incoherent space is given by equation (50). The net work for the closed path a b c d a is given by

$$\bar{W} = \bar{W}_{ab} + \bar{W}_{bc} + \bar{W}_{cd} + \bar{W}_{da} \tag{83}$$

$$= \bar{Q}_{ab} + \bar{Q}_{bc} + \bar{Q}_{cd} + \bar{Q}_{da} - \bar{\omega}_{ab} - \bar{\omega}_{bc} - \bar{\omega}_{cd} - \bar{\omega}_{da}$$

Substituting equations (73) through (82) into equation (83) gives, after the cancellation of the internal energy terms, the net work as

$$\bar{W} = T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] + T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] \tag{84}$$

$$- \bar{\omega}_{ab} - \bar{\omega}_{bc} - \bar{\omega}_{cd} - \bar{\omega}_{da}$$

The change in internal energy for each path segment is obtained from equation (55) to be

$$\bar{U}_{ab} = U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] \tag{85}$$

$$\bar{U}_{bc} = U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ub})] \tag{86}$$

$$\bar{U}_{cd} = U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] \tag{87}$$

$$\bar{U}_{da} = U[\exp(j\theta_{Ua}) - \exp(j\theta_{Ud})] \tag{88}$$

so that for a closed cycle

$$\bar{U}_{ab} + \bar{U}_{bc} + \bar{U}_{cd} + \bar{U}_{da} = 0 \tag{89}$$

which gives the cancellation of the internal energy terms in obtaining equation (84). Equation (89) is valid for all of the ultrafast engine cycles considered in this paper because $\bar{U}$ is a complex number state function and

$$\oint d\bar{U} = 0 \tag{90}$$

where the integral refers to any closed thermodynamic cycle.

The complex number engine efficiency is given by

$$\bar{\eta} = \bar{W}/\bar{Q}_{ab} = 1 - \bar{A}/\bar{B} \tag{91}$$

where equations (73) and (83) give

$$\begin{aligned}\bar{A} &= -\bar{Q}_{bc} - \bar{Q}_{cd} - \bar{Q}_{da} + \bar{\omega}_{ab} + \bar{\omega}_{bc} + \bar{\omega}_{cd} + \bar{\omega}_{da} \\ &= T_{cd}S[\exp(j\theta_{Sc}) - \exp(j\theta_{Sd})] + \bar{\omega}_{ab} + \bar{\omega}_{bc} + \bar{\omega}_{cd} + \bar{\omega}_{da} \\ &= G + jH\end{aligned} \tag{92}$$

$$\begin{aligned}\bar{B} &= T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] \\ &= E + jF\end{aligned} \tag{93}$$

Then the efficiency can be written as

$$\begin{aligned}\bar{\eta} &= 1 - (G + jH)/(E + jF) \\ &= 1 - C - jD\end{aligned} \tag{94}$$

and the measured efficiency is given by the real part of equation (94) so that

$$\eta_m = \eta_R = 1 - C \tag{95}$$

where

$$C = (GE + HF)/(E^2 + F^2) \tag{96}$$

$$D = (HE - GF)/(E^2 + F^2) \tag{97}$$

and where

$$\begin{aligned}G = T_{cd}S(\cos\theta_{Sc} - \cos\theta_{Sd}) + \omega_{ab}\cos\theta_{\omega}^{ab} + \omega_{bc}\cos\theta_{\omega}^{bc} \\ + \omega_{cd}\cos\theta_{\omega}^{cd} + \omega_{da}\cos\theta_{\omega}^{da}\end{aligned} \tag{98}$$

$$\begin{aligned}H = T_{cd}S(\sin\theta_{Sc} - \sin\theta_{Sd}) + \omega_{ab}\sin\theta_{\omega}^{ab} + \omega_{bc}\sin\theta_{\omega}^{bc} \\ + \omega_{cd}\sin\theta_{\omega}^{cd} + \omega_{da}\sin\theta_{\omega}^{da}\end{aligned} \tag{99}$$

$$E = T_{ab}S(\cos\theta_{Sb} - \cos\theta_{Sa}) \tag{100}$$

$$F = T_{ab}S(\sin\theta_{Sb} - \sin\theta_{Sa}) \tag{101}$$

Note also that

$$E^2 + F^2 = 2[1 - \cos(\theta_{Sb} - \theta_{Sa})] \tag{102}$$

Equations (95) through (101) show that for the Carnot engine the generalized work elements $\bar{\omega}_{\mu\nu}$ enter directly into the calculation of efficiency, while the ordinary work elements $\bar{W}_{\mu\nu}$ do not enter directly.

If the generalized work elements are neglected in equations (98) and (99) it follows from equations (96) through (102) that

$$C = (T_{cd}/T_{ab})(I/L) \tag{103}$$

$$D = (T_{cd}/T_{ab})(J/L) \tag{104}$$

$$I = \cos(\theta_{Sc} - \theta_{Sb}) + \cos(\theta_{Sd} - \theta_{Sa}) - \cos(\theta_{Sc} - \theta_{Sa}) - \cos(\theta_{Sd} - \theta_{Sb}) \tag{105}$$

$$J = \sin(\theta_{Sc} - \theta_{Sb}) + \sin(\theta_{Sd} - \theta_{Sa}) - \sin(\theta_{Sc} - \theta_{Sa}) - \sin(\theta_{Sd} - \theta_{Sb}) \tag{106}$$

$$L = 2[1 - \cos(\theta_{Sb} - \theta_{Sa})] \tag{107}$$

Because θ_S varies inversely with S [as shown by equation (176) in the accompanying paper on the Quantum Theory of Time and Thermodynamics] it follows that $\theta_{Sb} < \theta_{Sa}$ and $\theta_{Sc} < \theta_{Sd}$ so that equations (98) through (101) give $G > 0$, $H < 0$, $E > 0$ and $F < 0$. Therefore the value of C given by equation (96) for the general case or approximately by equation (103) satisfies $C > 0$ and therefore from equation (95) $\eta_m < 1$ because $T_{cd} < T_{ab}$ as shown in Figure 1b. The measured efficiency of an ultrafast Carnot engine is always less than unity.

For small values of internal phase angles θ_U and θ_S equations (98) through (101) can be simplified by using equations (560) and (56P) as follows

$$G = T_{cd}S(\theta_{Sd}^2 - \theta_{Sc}^2)/2 - M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2 - M_{bc}(\alpha_c - \alpha_b)(\theta_{Ub} + \theta_{Uc})/2 - M_{cd}(\alpha_d - \alpha_c)(\theta_{Uc} + \theta_{Ud})/2 - M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2 \tag{107A}$$

$$H = T_{cd}S(\theta_{Sc} - \theta_{Sd}) + M_{ab}(\alpha_b - \alpha_a) + M_{bc}(\alpha_c - \alpha_b) + M_{cd}(\alpha_d - \alpha_c) + M_{da}(\alpha_a - \alpha_d) \tag{107B}$$

$$E = T_{ab}S(\theta_{Sa}^2 - \theta_{Sb}^2)/2 \tag{107C}$$

$$F = T_{ab}S(\theta_{Sb} - \theta_{Sa}) \tag{107D}$$

If in addition, the generalized work terms can be neglected (which is not generally the case) then equations (96), (97) and (107A) through (107D) give

$$C = (T_{cd}/T_{ab})(\theta_{Sd} - \theta_{Sc})/(\theta_{Sa} - \theta_{Sb}) \tag{108}$$

$$D = 1/2(T_{cd}/T_{ab})(\theta_{Sd} - \theta_{Sc})(\theta_{Sa} + \theta_{Sb} - \theta_{Sd} - \theta_{Sc})/(\theta_{Sb} - \theta_{Sa}) \tag{109}$$

and therefore because $\theta_{Sc} < \theta_{Sd}$, $\theta_{Sb} < \theta_{Sa}$ and $T_{cd}/T_{ab} < 1$ it follows that $0 < C < 1$ and there $\eta_m < 1$ as is required by the second law of thermodynamics. Finally, it follows from equations (107A) and (107B) that the terms $(\alpha_\mu - \alpha_\nu)$ must be first order homogeneous functions of the terms $(\theta_{Sd} \pm \theta_{Sc})$ and $(\theta_{U\mu} + \theta_{U\nu})$.

B. Internal Phase Carnot Engine in Coherent Spacetime.

This is the case of an ultrafast coherent heat Carnot engine whose working substance is a gas of high-T_c superconducting electron pairs which exist in a coherent spacetime (Case C of Section 2). The variables and constants for the path segments of the closed cycle shown in Figure 1c are as follows

	Variables	Constants	
path ab	$\theta_S , \theta_U , \theta_V , \theta_P , P , \theta_\alpha$	$S , U , V , T_{ab} , \alpha$	(110)
path bc	$\theta_U , \theta_P , \theta_V , P , T , \theta_\alpha$	$S , U , V , \theta_S^{bc} , \alpha$	(111)
path cd	$\theta_S , \theta_U , \theta_V , \theta_P , P , \theta_\alpha$	$S , U , V , T_{cd} , \alpha$	(112)
path da	$\theta_U , \theta_P , \theta_V , P , T , \theta_\alpha$	$S , U , V , \theta_S^{da} , \alpha$	(113)

For an ultrafast thermodynamic process occurring in coherent spacetime the pressure is given by equation (60). The thermodynamic functions correspond to Case C of Section 2. The same expressions for the efficiency that were developed in equations (91) through (109) for incoherent spacetime are also valid for coherent spacetime provided that the replacement

$$\alpha_\mu - \alpha_\nu \to \alpha(\theta_\mu - \theta_\nu) \tag{113A}$$

is made for all expressions for the generalized work elements as is done in equations (67C), (67D), (67G) and (67H). With these substitutions made in equations (107A) and (107B) it follows that $\theta_\mu - \theta_\nu$ must be first order homogeneous functions of $\theta_{Sd} \pm \theta_{Sc}$ and $\theta_{U\mu} + \theta_{U\nu}$.

4. ULTRAFAST OTTO ENGINE. The most common type of internal combustion engine is based on the Otto cycle (Figure 2a). In its simplest form the air-standard Otto engine cycle has four phases: a → b the air is heated at constant volume, b → c the air is expanded reversibly and adiabatically, c → d the air is cooled at constant volume, and d → a the air is compressed reversibly and adiabatically. The ignition phase a → b consists of a constant volume combus-

tion or a constant volume process of heat addition to the equivalent air cycle. In the air-standard cycle a constant volume heat addition from an external heat reservoir is substituted for the combustion process, and a constant volume cooling followed by an adiabatic compression ends the cycle. It is easy to show that the efficiency for this closed cycle conventional Otto cycle is given by[2-9]

$$\eta = 1 - (T_c - T_d)/(T_b - T_a) = 1 - T_d/T_a \tag{114}$$

This can be rewritten in terms of the compression ratio $r = V_c/V_b$ as follows[2-9]

$$\eta = 1 - r^{-(c-1)} \tag{115}$$

where c = adiabatic constant = 1.4 for ideal gas. Unlike the Carnot engine the Otto engine is not reversible and therefore its efficiency is lower than that of the Carnot engine. More complicated Otto cycles exist such as the Otto cycle with throttling.[2-9]

A. Internal Phase Otto Engine for Incoherent Spacetime.

An ultrafast internal phase cycle for the Otto engine with ordinary matter as a working substance is now considered. The variables and constants for the various thermodynamic path elements of the internal phase Otto engine for incoherent spacetime (Case B of Section 2) shown in Figure 2b are as follows

	Variables	Constants	
path ab	$\theta_S, \theta_U, \theta_P, P, T, \alpha$	S, U, V_{ab}	(116)
path bc	$\theta_U, \theta_P, P, V, T, \alpha$	S, U, θ_S^{bc}	(117)
path cd	$\theta_S, \theta_U, \theta_P, P, T, \alpha$	S, U, V_{cd}	(118)
path da	$\theta_U, \theta_P, P, V, T, \alpha$	S, U, θ_S^{da}	(119)

From Figure 2b and equations (49) through (56B) it follows that for the ultrafast Otto engine the heat generated and exhausted at constant volume (paths ab and cd) and adiabatically (paths bc and da) are given by

$$\bar{Q}_{ab} = jS \int_a^b T \exp(j\theta_S) d\theta_S = jU \int_a^b \exp(j\theta_U) d\theta_U + \int_a^b \bar{M} d\alpha \tag{120}$$

$$= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] + \bar{\omega}_{ab}$$

$$\bar{Q}_{bc} = 0 \tag{121}$$

$$\bar{Q}_{cd} = jS \int_c^d T \exp(j\theta_S) d\theta_S = jU \int_c^d \exp(j\theta_U) d\theta_U + \int_c^d \bar{M} d\alpha \tag{122}$$

$$= U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] + \bar{\omega}_{cd}$$

$$\bar{Q}_{da} = 0 \tag{123}$$

where $\bar{\omega}_{ab}$ and $\bar{\omega}_{cd}$ = incoherent work elements done by the generalized forces for paths ab and cd respectively and which are given by eqaution (56B). If these generalized force terms were not included then equation (10) shows that both paths ab and cd with constant volumes would have a common constant temperature $T = U/S$, with $\theta_S = \theta_U$, and a cycle would not be possible. Thus generalized forces must be included in the internal phase Otto cycle.

From equations (49) through (56B) it is easy to see that the work done along the path segments is given by

$$\bar{W}_{ab} = 0 \tag{124}$$

$$\bar{W}_{bc} = \int_b^c \bar{P}dV = -j \int_b^c \bar{U}(d\theta_U/dV)dV - \int_b^c \bar{M}d\alpha \tag{125}$$

$$= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Uc})] - \bar{\omega}_{bc}$$

$$\bar{W}_{cd} = 0 \tag{126}$$

$$\bar{W}_{da} = \int_d^a \bar{P}dV = -j \int_d^a \bar{U}(d\theta_U/dV)dV - \int_d^a \bar{M}d\alpha \tag{127}$$

$$= U[\exp(j\theta_{Ud}) - \exp(j\theta_{Ua})] - \bar{\omega}_{da}$$

For an ultrafast thermodynamic process in incoherent spacetime the pressure is given by equation (50). Then the total work done for the complete cycle is

$$\bar{W} = \bar{W}_{ab} + \bar{W}_{bc} + \bar{W}_{cd} + \bar{W}_{da} \tag{128}$$

$$= \bar{Q}_{ab} + \bar{Q}_{bc} + \bar{Q}_{cd} + \bar{Q}_{da} - \bar{\omega}_{ab} - \bar{\omega}_{bc} - \bar{\omega}_{cd} - \bar{\omega}_{da}$$

$$= \bar{Q}_{ab} + \bar{Q}_{cd} - \bar{\omega}_{ab} - \bar{\omega}_{bc} - \bar{\omega}_{cd} - \bar{\omega}_{da}$$

$$= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua}) + \exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] - \bar{\omega}_{bc} - \bar{\omega}_{da}$$

The complex number efficiency is given by

$$\bar{\eta} = \bar{W}/\bar{Q}_{ab} = 1 - \bar{A}/\bar{B} \tag{129}$$

where

$$\bar{A} = U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ud})] + \bar{\omega}_{ab} + \bar{\omega}_{bc} + \bar{\omega}_{da} \tag{130}$$

$$\bar{B} = U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] + \bar{\omega}_{ab} \tag{131}$$

Equations (129) through (131) can be rewritten as

$$\bar{\eta} = \eta \exp(j\theta_\eta) = 1 - (G + jH)/(E + jF) \tag{132}$$

$$= 1 - C - jD$$

where

$$C = (GE + HF)/(E^2 + F^2) \tag{133}$$

$$D = (HE - GF)/(E^2 + F^2) \tag{134}$$

$$G = U(\cos\theta_{Uc} - \cos\theta_{Ud}) + \omega_{ab}\cos\theta_\omega^{ab} + \omega_{bc}\cos\theta_\omega^{bc} + \omega_{da}\cos\theta_\omega^{da} \tag{135}$$

$$H = U(\sin\theta_{Uc} - \sin\theta_{Ud}) + \omega_{ab}\sin\theta_\omega^{ab} + \omega_{bc}\sin\theta_\omega^{bc} + \omega_{da}\sin\theta_\omega^{da} \tag{136}$$

$$E = U(\cos\theta_{Ub} - \cos\theta_{Ua}) + \omega_{ab}\cos\theta_\omega^{ab} \tag{137}$$

$$F = U(\sin\theta_{Ub} - \sin\theta_{Ua}) + \omega_{ab}\sin\theta_\omega^{ab} \tag{138}$$

where the complex number generalized works are written as in equations (56B), (56I) and (56J). Note that $C > 0$. For the Otto engine the work elements $\bar{W}_{\mu\nu}$ do not enter directly into the efficiency calculations given by equations (132) through (138).

For small values of the internal phase angles it follows from equations (56O) and (56P) and equations (135) through (138) that

$$G \sim U(\theta_{Ud}^2 - \theta_{Uc}^2)/2 - M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2 \tag{139}$$

$$- M_{bc}(\alpha_c - \alpha_b)(\theta_{Ub} + \theta_{Uc})/2 - M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2$$

$$H \sim U(\theta_{Uc} - \theta_{Ud}) + M_{ab}(\alpha_b - \alpha_a) + M_{bc}(\alpha_c - \alpha_b) + M_{da}(\alpha_a - \alpha_d) \tag{140}$$

$$E \sim U(\theta_{Ua}^2 - \theta_{Ub}^2)/2 - M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2 \tag{141}$$

$$= [U(\theta_{Ua} - \theta_{Ub}) - M_{ab}(\alpha_b - \alpha_a)](\theta_{Ua} + \theta_{Ub})/2$$

$$F \sim U(\theta_{Ub} - \theta_{Ua}) + M_{ab}(\alpha_b - \alpha_a) \tag{142}$$

The terms $(\alpha_\mu - \alpha_\nu)$ must be first order homogeneous functions of $(\theta_{U\mu} \pm \theta_{U\nu})$.

From equation (132) it follows that the measured efficiency is given by

$$\eta_m = \eta_R = 1 - C \tag{143}$$

$$\eta_I = - D \tag{144}$$

$$\eta = [(1 - C)^2 + D^2]^{1/2} \tag{145}$$

$$\tan \theta_\eta = - D/(1 - C) \tag{146}$$

where C is given by equation (133). Because $C > 0$ it follows that $\eta_m < 1$.

B. Internal Phase Otto Engine in Coherent Spacetime.

The ultrafast coherent heat Otto engine uses a gas of high-T_c superconducting electron pairs as a working substance which is located in coherent spacetime. The variables and fixed quantities for the thermodynamic path given in Figure 2c are

	Variables	Constants	
path ab	$\theta_S , \theta_U , \theta_P , P , T , \theta_\alpha$	$S , U , \theta_V^{ab} , V , \alpha$	(147)
path bc	$\theta_U , \theta_P , P , \theta_V , T , \theta_\alpha$	$S , U , \theta_S^{bc} , V , \alpha$	(148)
path cd	$\theta_S , \theta_U , \theta_P , P , T , \theta_\alpha$	$S , U , \theta_V^{cd} , V , \alpha$	(149)
path da	$\theta_U , \theta_P , P , \theta_V , T , \theta_\alpha$	$S , U , \theta_S^{da} , V , \alpha$	(150)

The work elements for coherent spacetime are given in Case C of Section 2. With these changes the preceding analysis can be used to calculate the efficiency for the case of coherent thermodynamic functions and coherent spacetime. In particular, equations (132) through (143) give the efficiency for this case provided the generalized works $\bar{\omega}_{ab}$, $\bar{\omega}_{bc}$ and $\bar{\omega}_{da}$ are calculated from equation (67) as follows

$$\bar{\omega}_{ab} = \alpha \int_a^b \bar{M} d\theta_\alpha \qquad \alpha_b - \alpha_a = \alpha(\theta_{\alpha b} - \theta_{\alpha a}) \tag{150A}$$

$$\bar{\omega}_{bc} = \alpha \int_b^c \bar{M} d\theta_\alpha \qquad \alpha_c - \alpha_b = \alpha(\theta_{\alpha c} - \theta_{\alpha b}) \tag{150B}$$

$$\bar{\omega}_{da} = \alpha \int_d^a \bar{M} d\theta_\alpha \qquad \alpha_a - \alpha_d = \alpha(\theta_{\alpha a} - \theta_{\alpha d}) \tag{150C}$$

which result in equations (67C) and (67D) generally and equations (67G) and (67H) specifically for the case of small internal phase angles. The components of the generalized work elements $\bar{\omega}_{ab}$, $\bar{\omega}_{bc}$ and $\bar{\omega}_{da}$ given in equations (150A) through (150C) are used in equations (135) through (138) to calculate the efficiency of a coherent spacetime Otto engine for the general case, while the quantities $\alpha_b - \alpha_a$, $\alpha_c - \alpha_b$ and $\alpha_a - \alpha_d$ that appear in equations (150A) through (150C) are used in equations (139) through (142) to calculate the coherent spacetime engine

efficiency for the case of small internal phase angles. The quantities $\theta_{\alpha\mu} - \theta_{\alpha\nu}$ must be first order homogeneous functions of $\theta_{U\mu} \pm \theta_{U\nu}$. Finally, for an ultrafast process in coherent spacetime the pressure can be represented by equation (60).

5. ULTRAFAST DIESEL ENGINE. In the conventional Diesel engine combustion is regulated to occur at a constant pressure. In order to accomplish this the air temperature in the cylinder must be higher than the self-ignition temperature of the fuel. Therefore a simplified four phase structure of the Diesel cycle is shown in Figure 3a and consists of: a → b addition of heat at constant pressure, b → c isentropic expansion, c → d heat rejection at constant volume, and d → a isentropic compression.[2-9] The ignition phase is a → b and consists of constant pressure combustion. The fuel is added later after the compression of air to achieve constant pressure combustion. The thermal efficiency of the conventional Diesel cycle is given by[2-9]

$$\eta = 1 - c^{-1}(T_c - T_d)/(T_b - T_a) \tag{151}$$

where c = adiabatic constant = 1.4 for ideal gases. For high compression ratios the Diesel cycle is more efficient than the Otto cycle. For equal compression ratios the Otto cycle is more efficient than the Diesel cycle.

A. Internal Phase Diesel Engine in Incoherent Spacetime.

This section describes an ultrafast Diesel cycle with ordinary matter as a working material. The variables and fixed quantities for the internal phase Diesel engine are discerned from Figure 3b for incoherent spacetime (Case B of Section 2) to be

	Variables	Constants	
path ab	$\theta_S, \theta_U, V, T, \alpha$	$S, U, P_{ab}, \theta_P^{ab}$	(153)
path bc	$\theta_U, \theta_P, P, V, T, \alpha$	S, U, θ_S^{bc}	(154)
path cd	$\theta_S, \theta_U, \theta_p, P, T, \alpha$	S, U, V_{cd}	(155)
path da	$\theta_U, \theta_P, P, V, T, \alpha$	S, U, θ_S^{da}	(156)

Referring to Figure 3b, the work done on the four path segments of the cycle can be deduced from equations (49), (56A) and (56B) to be

$$\bar{W}_{ab} = \int_a^b \bar{P}dV = \bar{P}_{ab}(V_b - V_a) \tag{157}$$

$$\bar{W}_{bc} = \int_b^c \bar{P}dV = -jU \int_b^c \exp(j\theta_U)d\theta_U - \int_b^c \bar{M}d\alpha \tag{158}$$

$$= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Uc})] - \bar{\omega}_{bc}$$

$$\bar{W}_{cd} = 0 \tag{159}$$

$$\bar{W}_{da} = \int_d^a \bar{P}dV = -jU \int_d^a \exp(j\theta_U)d\theta_U - \int_d^a \bar{M}d\alpha \tag{160}$$

$$= U[\exp(j\theta_{Ud}) - \exp(j\theta_{Ua})] - \bar{\omega}_{da}$$

where $\bar{P}$ is given by equation (50). The total work done around the path a b c d a is given by

$$\bar{W} = \bar{W}_{ab} + \bar{W}_{bc} + \bar{W}_{cd} + \bar{W}_{da} \tag{161}$$

$$= \bar{W}_{ab} + U[\exp(j\theta_{Ub}) - \exp(j\theta_{Uc}) + \exp(j\theta_{Ud}) - \exp(j\theta_{Ua})] - \bar{\omega}_{bc} - \bar{\omega}_{da}$$

The heat trnasferred during the ultrafast Diesel cycle is obtained from Figure 3b and equations (49) and (54) through (56B) to be

$$\bar{Q}_{ab} = jS \int_a^b T \exp(j\theta_S)d\theta_S = jU \int_a^b \exp(j\theta_U)d\theta_U + \bar{W}_{ab} + \int_a^b \bar{M}d\alpha \tag{162}$$

$$= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] + \bar{W}_{ab} + \bar{\omega}_{ab}$$

$$\bar{Q}_{bc} = 0 \tag{163}$$

$$\bar{Q}_{cd} = U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] + \bar{\omega}_{cd} \tag{164}$$

$$\bar{Q}_{da} = 0 \tag{165}$$

The net heat transferred is obtained from equations (162) through (165) to be

$$\bar{Q} = \bar{Q}_{ab} + \bar{Q}_{bc} + \bar{Q}_{cd} + \bar{Q}_{da} \tag{166}$$

$$= \bar{W} + \bar{\omega}_{ab} + \bar{\omega}_{bc} + \bar{\omega}_{cd} + \bar{\omega}_{da}$$

where $\bar{W}$ is given by equation (161).

The efficiency is given by the ratio of the net work to the input heat, so that equations (161) and (162) give

$$\bar{\eta} = \bar{W}/\bar{Q}_{ab} = 1 - \bar{A}/\bar{B} = 1 - (G + jH)/(E + jF) \tag{167}$$

$$= 1 - C - jD$$

where

$$\bar{A} = U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ud})] + \bar{\omega}_{ab} + \bar{\omega}_{bc} + \bar{\omega}_{da} \tag{168}$$

$$\bar{B} = U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] + \bar{W}_{ab} + \bar{\omega}_{ab} \tag{169}$$

$$C = (GE + HF)/(E^2 + F^2) \tag{170}$$

$$D = (HE - GF)/(E^2 + F^2) \tag{171}$$

$$G = U(\cos\theta_{Uc} - \cos\theta_{Ud}) + \omega_{ab}\cos\theta_{\omega}^{ab} + \omega_{bc}\cos\theta_{\omega}^{bc} + \omega_{da}\cos\theta_{\omega}^{da} \tag{172}$$

$$H = U(\sin\theta_{Uc} - \sin\theta_{Ud}) + \omega_{ab}\sin\theta_{\omega}^{ab} + \omega_{bc}\sin\theta_{\omega}^{bc} + \omega_{da}\sin\theta_{\omega}^{da} \tag{173}$$

$$E = U(\cos\theta_{Ub} - \cos\theta_{Ua}) + W_{ab}\cos\theta_{W}^{ab} + \omega_{ab}\cos\theta_{\omega}^{ab} \tag{174}$$

$$F = U(\sin\theta_{Ub} - \sin\theta_{Ua}) + W_{ab}\sin\theta_{W}^{ab} + \omega_{ab}\sin\theta_{\omega}^{ab} \tag{175}$$

with $C > 0$, and where the elements of work can be written as in equations (56G) through (56J). The measured efficiency is obtained from equations (167) to be

$$\eta_m = 1 - C \tag{176}$$

where $C > 0$ from equation (170). The internal phase angle and magnitude of the efficiency is obtained from equation (167) as

$$\tan\theta_\eta = -D/(1 - C) \tag{177}$$

$$\eta = [(1 - C)^2 + D^2]^{1/2} \tag{178}$$

Note that the work element $\bar{W}_{ab}$ enters the efficiency calculation through equations (169), (174) and (175). In this way the general expression for the efficiency of the ultrafast Diesel engine operating in incoherent spacetime is calculated.

For small internal phase angles it follows from equations (56M) through (56P) and (172) through (175) that

$$G = U(\theta_{Ud}^2 - \theta_{Uc}^2)/2 - M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2 \tag{179}$$

$$- M_{bc}(\alpha_c - \alpha_b)(\theta_{Ub} + \theta_{Uc})/2 - M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2$$

$$H = U(\theta_{Uc} - \theta_{Ud}) + M_{ab}(\alpha_b - \alpha_a) + M_{bc}(\alpha_c - \alpha_b) + M_{da}(\alpha_a - \alpha_d) \tag{180}$$

$$E = U(\theta_{Ua}^2 - \theta_{Ub}^2)/2 - P_{ab}(V_b - V_a)(\theta_{Ua} + \theta_{Ub})/2 \qquad (181)$$

$$- M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2$$

$$F = U(\theta_{Ub} - \theta_{Ua}) + P_{ab}(V_b - V_a) + M_{ab}(\alpha_b - \alpha_a) \qquad (182)$$

from which the efficiency is calculated by equations (167) through (171). The elements $V_\mu - V_\nu$ and $\alpha_\mu - \alpha_\nu$ must be first order homogeneous functions of $\theta_{U\mu} \pm \theta_{U\nu}$.

B. Internal Phase Diesel Engine in Coherent Spacetime.

This is the case of an ultrafast coherent heat Diesel engine operating with a gas of high-T_c superconducting electron pairs that exist in coherent spacetime. For coherent spacetime the variables and constant quantities are obtained from Figure 3c and Case C of Section 2 to be

	Variables	Constants	
path ab	$\theta_S , \theta_U , \theta_V , T , \theta_\alpha$	$S , U , P_{ab} , \theta_P^{ab} , V , \alpha$	(183)
path bc	$\theta_U , \theta_P , P , \theta_V , T , \theta_\alpha$	$S , U , \theta_S^{bc} , V , \alpha$	(184)
path cd	$\theta_S , \theta_U , \theta_P , P , T , \theta_\alpha$	$S , U , \theta_V^{cd} , V , \alpha$	(185)
path da	$\theta_U , \theta_P , P , \theta_V , T , \theta_\alpha$	$S , U , \theta_S^{da} , V , \alpha$	(186)

The analysis for this case gives the same results as in Case A for incoherent spacetime except that the following expressions for the work elements $\bar{W}_{\mu\nu}$ in coherent spacetime are obtained from equation (66)

$$\bar{W}_{ab} = \bar{P}_{ab} V(\theta_{Vb} - \theta_{Va}) \qquad V_b - V_a = V(\theta_{Vb} - \theta_{Va}) \qquad (187)$$

$$\bar{W}_{bc} = V \int_b^c \bar{P} d\theta_V \qquad V_c - V_b = V(\theta_{Vc} - \theta_{Vb}) \qquad (188)$$

$$\bar{W}_{cd} = 0 \qquad V_d - V_c = 0 \qquad (189)$$

$$\bar{W}_{da} = V \int_d^a \bar{P} d\theta_V \qquad V_a - V_d = V(\theta_{Va} - \theta_{Vd}) \qquad (190)$$

and similarly the generalized work elements $\bar{\omega}_{\mu\nu}$ are given by equation (67) for coherent spacetime. The efficiency calculations in equations (157) through (182) are therefore valid for the case of coherent spacetime if the work elements in equations (66) through (67H) are used, and in particular if the following substitutions are made

$$V_\mu - V_\nu \to V(\theta_{V\mu} - \theta_{V\nu}) \tag{191}$$

$$\alpha_\mu - \alpha_\nu \to \alpha(\theta_{\alpha\mu} - \theta_{\alpha\nu}) \tag{192}$$

The elements $\theta_{V\mu} - \theta_{V\nu}$ and $\theta_{\alpha\mu} - \theta_{\alpha\nu}$ must then be first order homogeneous functions of $\theta_{U\mu} \pm \theta_{U\nu}$ as can be from equations (179) through (182). For an ultrafast thermodynamic process occurring in coherent spacetime the pressure is given by equation (60).

6. ULTRAFAST ERICSSON ENGINE. The Ericsson cycle is important because it makes use of a regenerator in a hot air engine to heat the air. A regenerator is a chamber filled with brickwork or wire mesh which serves the purpose to transfer energy from a hot gas and store it at constant pressure.[2-9] The Ericsson cycle consists of four phases as shown in Figure 4a and which consists of the following elements: a → b constant temperature transfer of heat to the engine from an external source, b → c constant pressure transfer of heat from the engine to the regenerator, c → d constant temperature loss of heat (entropy) to an external sink, and d → a constant pressure energy retrieval from the regenerator. The thermal efficiency of the conventional Ericsson engine is given by[2-9]

$$\eta = 1 - T_{cd}/T_{ab} \tag{193}$$

which is the same maximum efficiency associated with the Carnot engine. Ericsson engines have low mean effective pressures and low temperatures of the working substance, and cannot compete with modern reciprocating engines which are based on the Otto or Diesel cycles. These engines are not used today but the cycle is of heuristic value.

A. Internal Phase Ericsson Engine with Incoherent Spacetime.

This section describes an ultrafast coherent heat Ericsson engine with ordinary matter used as a working substance. A glance at Figure 4b for the ultrafast Ericsson engine in incoherent spacetime shows that the variables and fixed quantities for the various path segments are (Case B of Section 2)

	Variables	Constants	
path ab	$\theta_S, \theta_U, \theta_P, P, V, \alpha$	S, U, T_{ab}	(194)
path bc	$\theta_S, \theta_U, V, T, \alpha$	$S, U, P_{bc}, \theta_P^{bc}$	(195)
path cd	$\theta_S, \theta_U, \theta_P, P, V, \alpha$	S, U, T_{cd}	(196)
path da	$\theta_S, \theta_U, V, T, \alpha$	$S, U, P_{da}, \theta_P^{da}$	(197)

Figure 4b and equations (49) through (56B) give the heat transfers for the path segments as follows

$$\bar{Q}_{ab} = jT_{ab}S \int_a^b \exp(j\theta_s)d\theta_S = T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] \tag{198}$$

$$\bar{Q}_{bc} = jS \int_b^c T \exp(j\theta_S)d\theta_S = jU \int_b^c \exp(j\theta_U)d\theta_U + \bar{P}_{bc}(V_c - V_b) + \int_b^c \bar{M}d\alpha \tag{199}$$

$$= U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ub})] + \bar{W}_{bc} + \bar{\omega}_{bc}$$

$$\bar{Q}_{cd} = jT_{cd}S \int_c^d \exp(j\theta_S)d\theta_S = T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] \tag{200}$$

$$\bar{Q}_{da} = jS \int_d^a T \exp(j\theta_S)d\theta_S = jU \int_d^a \exp(j\theta_U)d\theta_U + \bar{P}_{da}(V_a - V_d) + \int_d^a \bar{M}d\alpha \tag{201}$$

$$= U[\exp(j\theta_{Ua}) - \exp(j\theta_{Ud})] + \bar{W}_{da} + \bar{\omega}_{da}$$

where

$$\bar{W}_{bc} = \bar{P}_{bc}(V_c - V_b) \qquad W_{bc} = P_{bc}(V_c - V_b) \qquad \theta_W^{bc} = \theta_P^{bc} \tag{202}$$

$$\bar{W}_{da} = \bar{P}_{da}(V_a - V_d) \qquad W_{da} = P_{da}(V_a - V_d) \qquad \theta_W^{da} = \theta_P^{da} \tag{203}$$

and where

$$\bar{W}_{bc} = -\bar{W}_{cb} \qquad \bar{W}_{da} = -\bar{W}_{ad} \tag{204}$$

The work elements are obtained using equations (49) through (56) as

$$\bar{W}_{ab} = \int_a^b \bar{P}dV = j \int_a^b (T\bar{S}\ \partial\theta_S/\partial V - \bar{U}\ \partial\theta_U/\partial V - \bar{M}\ \partial\alpha/\partial V)dV \tag{205}$$

$$= T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] - U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] - \bar{\omega}_{ab}$$

$$\bar{W}_{bc} = \bar{P}_{bc}(V_c - V_b) \tag{206}$$

$$\bar{W}_{cd} = \int_c^d \bar{P}dV = j \int_c^d (T\bar{S}\ \partial\theta_S/\partial V - \bar{U}\ \partial\theta_U/\partial V - \bar{M}\ \partial\alpha/\partial V)dV \tag{207}$$

$$= T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] - U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] - \bar{\omega}_{cd}$$

$$\bar{W}_{da} = \bar{P}_{da}(V_a - V_d) \tag{208}$$

The pressure for this case is given by equation (50). The net work is given by

$$\bar{W} = \bar{W}_{ab} + \bar{W}_{bc} + \bar{W}_{cd} + \bar{W}_{da} \tag{209}$$

$$= \bar{Q}_{ab} + \bar{Q}_{bc} + \bar{Q}_{cd} + \bar{Q}_{da} - \bar{\omega}_{ab} - \bar{\omega}_{bc} - \bar{\omega}_{cd} - \bar{\omega}_{da}$$

$$= T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] + U[\exp(j\theta_{Ua}) - \exp(j\theta_{Ub})] + \bar{W}_{bc} - \bar{\omega}_{ab}$$

$$+ T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] + U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ud})] + \bar{W}_{da} - \bar{\omega}_{cd}$$

These work elements can then be used to calculate efficiency.

The Ericsson engine is unusual in the sense that heat is added to the engine twice during each cycle, once in the element ab when heat is obtained from an external source, and a second time during the path segment da when heat is retrieved from the regenerator. The efficiency is then written as

$$\bar{\eta} = \bar{W}/(\bar{Q}_{ab} + \bar{Q}_{da}) = 1 - \bar{A}/\bar{B} \tag{210}$$

$$= 1 - (G + jH)/(E + jF)$$

$$= 1 - C - jD \tag{211}$$

where

$$\bar{A} = T_{cd}S[\exp(j\theta_{Sc}) - \exp(j\theta_{Sd})] + U[\exp(j\theta_{Ub}) - \exp(j\theta_{Uc})] \tag{212}$$

$$- \bar{W}_{bc} + \bar{\omega}_{ab} + \bar{\omega}_{cd} + \bar{\omega}_{da}$$

$$\bar{B} = T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] + U[\exp(j\theta_{Ua}) - \exp(j\theta_{Ud})] \tag{213}$$

$$+ \bar{W}_{da} + \bar{\omega}_{da}$$

$$C = (GE + HF)/(E^2 + F^2) \tag{214}$$

$$D = (HE - GF)/(E^2 + F^2) \tag{215}$$

$$G = T_{cd}S(\cos\theta_{Sc} - \cos\theta_{Sd}) + U(\cos\theta_{Ub} - \cos\theta_{Uc}) - W_{bc}\cos\theta_W^{bc} \tag{216}$$

$$+ \omega_{ab}\cos\theta_\omega^{ab} + \omega_{cd}\cos\theta_\omega^{cd} + \omega_{da}\cos\theta_\omega^{da}$$

$$H = T_{cd}S(\sin\theta_{Sc} - \sin\theta_{Sd}) + U(\sin\theta_{Ub} - \sin\theta_{Uc}) - W_{bc}\sin\theta_W^{bc} \tag{217}$$

$$+ \omega_{ab}\sin\theta_\omega^{ab} + \omega_{cd}\sin\theta_\omega^{cd} + \omega_{da}\sin\theta_\omega^{da}$$

$$E = T_{ab}S(\cos\theta_{Sb} - \cos\theta_{Sa}) + U(\cos\theta_{Ua} - \cos\theta_{Ud}) \tag{218}$$
$$+ W_{da}\cos\theta_W^{da} + \omega_{da}\cos\theta_\omega^{da}$$

$$F = T_{ab}S(\sin\theta_{Sb} - \sin\theta_{Sa}) + U(\sin\theta_{Ua} - \sin\theta_{Ud}) \tag{219}$$
$$+ W_{da}\sin\theta_W^{da} + \omega_{da}\sin\theta_\omega^{da}$$

The work terms in equations (216) through (219) can be rewritten as in equations (56G) through (56L). The measured efficiency is given by the real part of equation (213) so that

$$\eta_m = 1 - C \tag{220}$$

The internal phase angle and magnitude of the complex number efficiency given in equation (213) are obtained from equations (177) and (178). Both $\bar{\omega}_{\mu\nu}$ and $\bar{W}_{\mu\nu}$ enter the calculation of the efficiency of the Ericsson engine.

For small internal phase angles it follows from equations (56M) through (56P) and equations (216) through (219) that

$$G = T_{cd}S(\theta_{Sd}^2 - \theta_{Sc}^2)/2 + U(\theta_{Uc}^2 - \theta_{Ub}^2)/2 + P_{bc}(V_c - V_b)(\theta_{Ub} + \theta_{Uc})/2 \tag{221}$$
$$- M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2 - M_{cd}(\alpha_d - \alpha_c)(\theta_{Uc} + \theta_{Ud})/2$$
$$- M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2$$

$$H = T_{cd}S(\theta_{Sc} - \theta_{Sd}) + U(\theta_{Ub} - \theta_{Uc}) - P_{bc}(V_c - V_b) \tag{222}$$
$$+ M_{ab}(\alpha_b - \alpha_a) + M_{cd}(\alpha_d - \alpha_c) + M_{da}(\alpha_a - \alpha_d)$$

$$E = T_{ab}S(\theta_{Sa}^2 - \theta_{Sb}^2) + U(\theta_{Ud}^2 - \theta_{Ua}^2) - P_{da}(V_a - V_d)(\theta_{Ud} + \theta_{Ua})/2 \tag{223}$$
$$- M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2$$

$$F = T_{ab}S(\theta_{Sb} - \theta_{Sa}) + U(\theta_{Ua} - \theta_{Ud}) + P_{da}(V_a - V_d) + M_{da}(\alpha_a - \alpha_d) \tag{224}$$

and the efficiency is calculated by inserting these equations into equations (214) and (220). The quantities $(V_\mu - V_\nu)$ and $(\alpha_\mu - \alpha_\nu)$ must be first order homogeneous functions of $(\theta_{S\mu} \pm \theta_{S\nu})$ and $(\theta_{U\mu} \pm \theta_{U\nu})$. Further simplification occurs if the

approximation $\theta_S = \theta_U$ is made in equations (221) through (224).

B. Internal Phase Ericsson Engine in Coherent Spacetime.

The working substance for the coherent spacetime coherent heat Ericsson engine is a coherent spacetime assembly of high-T_c superconducting electron pairs. Figure 4c shows that the variables and fixed quantities for the various path elements are given by

	Variables	Constants	
path ab	$\theta_S , \theta_U , \theta_P , P , \theta_V , \theta_\alpha$	$S , U , T_{ab} , V , \alpha$	(225)
path bc	$\theta_S , \theta_U , T , \theta_V , \theta_\alpha$	$S , U , P_{bc} , \theta_P^{bc} , V , \alpha$	(226)
path cd	$\theta_S , \theta_U , \theta_P , P , \theta_V , \theta_\alpha$	$S , U , T_{cd} , V , \alpha$	(227)
path da	$\theta_S , \theta_U , T , \theta_V , \theta_\alpha$	$S , U , P_{da} , \theta_P^{da} , V , \alpha$	(228)

For the coherent spacetime case the work elements $\bar{W}_{\mu\nu}$ are obtained from equation (66) and Figure 4c to be as follows

$$\bar{W}_{ab} = V \int_a^b \bar{P} d\theta_V \qquad V_b - V_a = V(\theta_{Vb} - \theta_{Va}) \tag{229}$$

$$\bar{W}_{bc} = V\bar{P}_{bc}(\theta_{Vc} - \theta_{Vb}) \qquad V_c - V_b = V(\theta_{Vc} - \theta_{Vb}) \tag{230}$$

$$\bar{W}_{cd} = V \int_c^d \bar{P} d\theta_V \qquad V_d - V_c = V(\theta_{Vd} - \theta_{Vc}) \tag{231}$$

$$\bar{W}_{da} = V\bar{P}_{ad}(\theta_{Va} - \theta_{Vd}) \qquad V_a - V_d = V(\theta_{Va} - \theta_{Vd}) \tag{232}$$

while the generalized work elements $\bar{\omega}_{\mu\nu}$ for coherent spacetime are given by equation (67). The pressure is given by equation (60). With the replacements of $\bar{W}_{\mu\nu}$ and $\bar{\omega}_{\mu\nu}$ with their coherent values as in equations (67A) through (67H), equations (198) through (224) can be used to calculate the efficiency of an internal phase Ericsson engine for coherent spacetime. The quantities $(\theta_{V\mu} - \theta_{V\nu})$ and $(\theta_{\alpha\mu} - \theta_{\alpha\nu})$ must be first order homogeneous functions of $(\theta_{U\mu} \pm \theta_{U\nu})$ and $(\theta_{S\mu} \pm \theta_{S\nu})$.

7. ULTRAFAST STIRLING ENGINE. The Stirling cycle was first introduced as the basis of a hot-air engine and uses a regenerator to heat the working substance. As shown in Figure 5a this cycle consists of four parts: a → b isothermal absorption of heat from an external reservoir at high temperature, b → c constant volume reversible rejection of heat to a regenerator, c → d isothermal rejection of heat to an external energy reservoir at low temperature, and d → a constant volume reversible absorption of heat from a regenerator. The efficiency of the conventional Stirling cycle is given by[2-9]

$$\eta = 1 - T_{cd}/T_{ab} \tag{233}$$

which, as in the Carnot engine, is the maximum possible value of the efficiency for conventional heat engines. Like the Ericsson engine, the Stirling engine has low mean effective pressure and a low working substance temperature and is not a practical source of power compared to the internal combustion engines. They are not used today for any commercial purposes.

A. Internal Phase Stirling Engine with Incoherent Spacetime.

Consider now an ultrafast Stirling engine whose working substance is ordinary matter. From 5b it is clear that the variables and constants for each path segment of the ultrafast Stirling engine in incoherent spacetime (Case B of Section 2) are given by

	Variables	Constants	
path ab	θ_S , θ_U , θ_P , P , V , α	S , U , T_{ab}	(234)
path bc	θ_S , θ_U , θ_P , P , T , α	S , U , V_{bc}	(235)
path cd	θ_S , θ_U , θ_P , P , V , α	S , U , T_{cd}	(236)
path da	θ_S , θ_U , θ_P , P , T , α	S , U , V_{da}	(237)

From equations (49) through (56B) and Figure 5b it follows that the heat transfers for the various path segments are

$$\bar{Q}_{ab} = jT_{ab}S \int_a^b \exp(j\theta_S)d\theta_S = T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] \tag{238}$$

$$\bar{Q}_{bc} = jS \int_b^c T \exp(j\theta_S)d\theta_S = jU \int_b^c \exp(j\theta_U)d\theta_U + \int_b^c \bar{M}d\alpha \tag{239}$$

$$= U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ub})] + \bar{\omega}_{bc}$$

$$\bar{Q}_{cd} = jT_{cd}S \int_c^d \exp(j\theta_S)d\theta_S = T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] \tag{240}$$

$$\bar{Q}_{da} = jS \int_d^a T \exp(j\theta_S)d\theta_S = jU \int_d^a \exp(j\theta_U)d\theta_U + \int_d^a \bar{M}d\alpha \tag{241}$$

$$= U[\exp(j\theta_{Ua}) - \exp(j\theta_{Ud})] + \bar{\omega}_{da}$$

Both $\bar{Q}_{ab}$ and $\bar{Q}_{da}$ correspond to the absorption of heat.

The work elements are obtained from equations (49) through (56) to be

$$\bar{W}_{ab} = T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] - U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] - \bar{\omega}_{ab} \quad (242)$$

$$\bar{W}_{bc} = 0 \quad (243)$$

$$\bar{W}_{cd} = T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})] - U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] - \bar{\omega}_{cd} \quad (244)$$

$$\bar{W}_{da} = 0 \quad (245)$$

The net work is given by

$$\bar{W} = \bar{W}_{ab} + \bar{W}_{bc} + \bar{W}_{cd} + \bar{W}_{da} \quad (246)$$

$$= \bar{Q}_{ab} + \bar{Q}_{bc} + \bar{Q}_{cd} + \bar{Q}_{da} - \bar{\omega}_{ab} - \bar{\omega}_{bc} - \bar{\omega}_{cd} - \bar{\omega}_{da}$$

$$= T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] + T_{cd}S[\exp(j\theta_{Sd}) - \exp(j\theta_{Sc})]$$

$$- U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] - U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] - \bar{\omega}_{ab} - \bar{\omega}_{cd}$$

Then the efficiency is given by

$$\bar{\eta} = \bar{W}/(\bar{Q}_{ab} + \bar{Q}_{da}) = 1 - \bar{A}/\bar{B} \quad (247)$$

$$= 1 - (G + jH)/(E + jF)$$

$$= 1 - C - jD$$

where

$$\bar{A} = T_{cd}S[\exp(j\theta_{Sc}) - \exp(j\theta_{Sd})] + U[\exp(j\theta_{Ub}) - \exp(j\theta_{Uc})] \quad (248)$$

$$+ \bar{\omega}_{ab} + \bar{\omega}_{cd} + \bar{\omega}_{da}$$

$$\bar{B} = T_{ab}S[\exp(j\theta_{Sb}) - \exp(j\theta_{Sa})] + U[\exp(j\theta_{Ua}) - \exp(j\theta_{Ud})] + \bar{\omega}_{da} \quad (249)$$

$$C = (GE + HF)/(E^2 + F^2) \quad (250)$$

$$D = (HE - GF)/(E^2 + F^2) \quad (251)$$

$$G = T_{cd}S(\cos\theta_{Sc} - \cos\theta_{Sd}) + U(\cos\theta_{Ub} - \cos\theta_{Uc}) \quad (252)$$

$$+ \omega_{ab}\cos\theta_{\omega}^{ab} + \omega_{cd}\cos\theta_{\omega}^{cd} + \omega_{da}\cos\theta_{\omega}^{da}$$

$$H = T_{cd}S(\sin\theta_{Sc} - \sin\theta_{Sd}) + U(\sin\theta_{Ub} - \sin\theta_{Uc}) \tag{253}$$
$$+ \omega_{ab}\sin\theta_{\omega}^{ab} + \omega_{cd}\sin\theta_{\omega}^{cd} + \omega_{da}\sin\theta_{\omega}^{da}$$

$$E = T_{ab}S(\cos\theta_{Sb} - \cos\theta_{Sa}) + U(\cos\theta_{Ua} - \cos\theta_{Ud}) + \omega_{da}\cos\theta_{\omega}^{da} \tag{254}$$

$$F = T_{ab}S(\sin\theta_{Sb} - \sin\theta_{Sa}) + U(\sin\theta_{Ua} - \sin\theta_{Ud}) + \omega_{da}\sin\theta_{\omega}^{da} \tag{255}$$

The generalized work elements can be written as in equations (56I) and (56J). The measured efficiency is given by

$$\eta_m = 1 - C \tag{255A}$$

The work elements $\bar{W}_{\mu\nu}$ do not directly enter the calculation of the efficiency of the Stirling engine as given by equations (247) through (255).

The small internal angle approximation for the efficiency can be calculated by using equations (56O) and (56P) and noting that for this case equations (252) through (255) become

$$G = T_{cd}S(\theta_{Sd}^2 - \theta_{Sc}^2)/2 + U(\theta_{Uc}^2 - \theta_{Ub}^2) - M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2 \tag{256}$$
$$- M_{cd}(\alpha_d - \alpha_c)(\theta_{Uc} + \theta_{Ud})/2 - M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2$$

$$H = T_{cd}S(\theta_{Sc} - \theta_{Sd}) + U(\theta_{Ub} - \theta_{Uc}) + M_{ab}(\alpha_b - \alpha_a) \tag{257}$$
$$+ M_{cd}(\alpha_d - \alpha_c) + M_{da}(\alpha_a - \alpha_d)$$

$$E = T_{ab}S(\theta_{Sa}^2 - \theta_{Sb}^2) + U(\theta_{Ud}^2 - \theta_{Ua}^2) - M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2 \tag{258}$$

$$F = T_{ab}S(\theta_{Sb} - \theta_{Sa}) + U(\theta_{Ua} - \theta_{Ud}) + M_{da}(\alpha_a - \alpha_d) \tag{259}$$

Then the efficiency is calculated using equations (247), (250), (251) and (255A). Again, the quantities $(\alpha_\mu - \alpha_\nu)$ must be first order homogeneous functions of $(\theta_{S\mu} \pm \theta_{S\nu})$ and $(\theta_{U\mu} \pm \theta_{U\nu})$.

B. Internal Phase Stirling Engine with Coherent Spacetime.

The coherent spacetime coherent heat Stirling engine uses a gas of high-T_c superconducting electron pairs for a working substance. Figure 5c shows that the variables and constants for each of the path segments for this type of engine (Case C of Section 2) are given by

	Variables	Constants	
path ab	$\theta_S , \theta_U , \theta_P , P , \theta_V , \theta_\alpha$	$S , U , T_{ab} , V , \alpha$	(260)
path bc	$\theta_S , \theta_U , \theta_P , P , T , \theta_\alpha$	$S , U , \theta_V^{bc} , V , \alpha$	(261)
path cd	$\theta_S , \theta_U , \theta_P , P , \theta_V , \theta_\alpha$	$S , U , T_{cd} , V , \alpha$	(262)
path da	$\theta_S , \theta_U , \theta_P , P , T , \theta_\alpha$	$S , U , \theta_V^{da} , V , \alpha$	(263)

The generalized work elements $\bar{\omega}_{\mu\nu}$ for the case of coherent spacetime are given by equations (67), (67C), (67D), (67G) and (67H) and when these results are substituted into equations (238) through (259) these equations give the efficiency for the internal phase Stirling engine for coherent spacetime. For this case $\alpha_\mu - \alpha_\nu = \alpha(\theta_{\alpha\mu} - \theta_{\alpha\nu})$, and the quantities $(\theta_{\alpha\mu} - \theta_{\alpha\nu})$ must be first order homogeneous functions of $(\theta_{U\mu} \pm \theta_{U\nu})$ and $(\theta_{S\mu} \pm \theta_{S\nu})$.

8. ULTRAFAST BRAYTON ENGINE. The Brayton (Joule) engine is a two cylinder engine, one used for compression and the other used for complete expansion of the products to atmospheric pressure.[2-9] The Brayton cycle consists of two constant pressure processes and two isentropic processes as shown in Figure 6a. More specifically the cycle consists of: a → b constant pressure addition of heat, b → c isentropic expansion, c → d constant pressure rejection of heat, and d → a isentropic compression. The efficiency of the conventional Brayton cycle is given as follows[2-9]

$$\eta = 1 - (h_c - h_d)/(h_b - h_a) = 1 - (T_c - T_d)/(T_b - T_a) \tag{264}$$

$$= 1 - T_d/ T_a = 1 - T_c/ T_b$$

where h = specific enthalpy given by

$$h = E + P \tag{265}$$

The Brayton cycle is used in gas turbines and jet engines with a compressor, combustion chamber and turbine, although the original Brayton engine was reciprocating. The mean effective pressure is low for the Brayton cycle and therefore the Brayton engine in impractical.

A. Internal Phase Brayton Engine for Incoherent Spacetime.

The ultrafast Brayton engine with ordinary matter for a working substance is treated in this section. Figure 6b shows that the ultrafast Brayton cycle for incoherent spacetime has the following variables and fixed quantities for each path segment of the cycle (Case B of Section 2)

	Variables	Constants	
path ab	$\theta_S , \theta_U , V , T , \alpha$	$S , U , P_{ab} , \theta_P^{ab}$	(266)
path bc	$\theta_U , \theta_P , P , V , T , \alpha$	S , U , θ_S^{bc}	(267)
path cd	$\theta_S , \theta_U , V , T , \alpha$	$S , U , P_{cd} , \theta_P^{cd}$	(268)
path da	$\theta_U , \theta_P , P , V , T , \alpha$	S , U , θ_S^{da}	(269)

The heat transfer for each path segment is obtained from Figure 6b and equations (49) through (56B) to be

$$\bar{Q}_{ab} = j \int_a^b T\bar{S}d\theta_S = jU \int_a^b \exp(j\theta_U)d\theta_U + \int_a^b \bar{P}dV + \int_a^b \bar{M}d\alpha \tag{270}$$

$$= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] + \bar{W}_{ab} + \bar{\omega}_{ab}$$

$$\bar{Q}_{bc} = 0 \tag{271}$$

$$\bar{Q}_{cd} = U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})] + \bar{W}_{cd} + \bar{\omega}_{cd} \tag{273}$$

$$\bar{Q}_{da} = 0 \tag{273}$$

where for constant pressure and incoherent spacetime the work elements are given by

$$\bar{W}_{ab} = \bar{P}_{ab}(V_b - V_a) \qquad W_{ab} = P_{ab}(V_b - V_a) \qquad \theta_W^{ab} = \theta_P^{ab} \tag{274}$$

$$\bar{W}_{cd} = \bar{P}_{cd}(V_d - V_c) \qquad W_{cd} = P_{cd}(V_d - V_c) \qquad \theta_W^{cd} = \theta_P^{cd} \tag{275}$$

The work done for each path segment is obtained from equations (49) through (56) to be

$$\bar{W}_{ab} = \bar{P}_{ab}(V_b - V_a) \tag{276}$$

$$\bar{W}_{bc} = - jU \int_b^c \exp(j\theta_U)d\theta_U - \int_b^c \bar{M}d\alpha \tag{277}$$

$$= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Uc})] - \bar{\omega}_{bc}$$

$$\bar{W}_{cd} = \bar{P}_{cd}(V_d - V_c) \tag{278}$$

$$\bar{W}_{da} = U[\exp(j\theta_{Ud}) - \exp(j\theta_{Ua})] - \bar{\omega}_{da} \tag{279}$$

The pressure for an ultrafast process in incoherent spacetime is given by equation (50). The net usable work done during the cycle is written as

$$\begin{aligned}\bar{W} &= \bar{W}_{ab} + \bar{W}_{bc} + \bar{W}_{cd} + \bar{W}_{da} \qquad (280)\\ &= \bar{Q}_{ab} + \bar{Q}_{bc} + \bar{Q}_{cd} + \bar{Q}_{da} - \bar{\omega}_{ab} - \bar{\omega}_{bc} - \bar{\omega}_{cd} - \bar{\omega}_{da}\\ &= U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] + U[\exp(j\theta_{Ud}) - \exp(j\theta_{Uc})]\\ &\quad + \bar{W}_{ab} + \bar{W}_{cd} - \bar{\omega}_{bc} - \bar{\omega}_{da}\end{aligned}$$

The net work can be used to calculate engine efficiency.

The efficiency is given by

$$\bar{\eta} = \bar{W}/\bar{Q}_{ab} = 1 - \bar{A}/\bar{B} \tag{281}$$

where

$$\bar{A} = U[\exp(j\theta_{Uc}) - \exp(j\theta_{Ud})] - \bar{W}_{cd} + \bar{\omega}_{ab} + \bar{\omega}_{bc} + \bar{\omega}_{da} \tag{282}$$

$$\bar{B} = U[\exp(j\theta_{Ub}) - \exp(j\theta_{Ua})] + \bar{W}_{ab} + \bar{\omega}_{ab} \tag{283}$$

Then the efficiency can be written as

$$\begin{aligned}\bar{\eta} &= \eta \exp(j\theta_{\eta}) = 1 - (G + jH)/(E + jF) \qquad (284)\\ &= 1 - C - jD\end{aligned}$$

where

$$C = (GE + HF)/(E^2 + F^2) \tag{285}$$

$$D = (HE - GF)/(E^2 + F^2) \tag{286}$$

$$G = U(\cos\theta_{Uc} - \cos\theta_{Ud}) - W_{cd}\cos\theta_W^{cd} + \omega_{ab}\cos\theta_{\omega}^{ab} + \omega_{bc}\cos\theta_{\omega}^{bc} + \omega_{da}\cos\theta_{\omega}^{da} \tag{287}$$

$$H = U(\sin\theta_{Uc} - \sin\theta_{Ud}) - W_{cd}\sin\theta_W^{cd} + \omega_{ab}\sin\theta_{\omega}^{ab} + \omega_{bc}\sin\theta_{\omega}^{bc} + \omega_{da}\sin\theta_{\omega}^{da} \tag{288}$$

$$E = U(\cos\theta_{Ub} - \cos\theta_{Ua}) + W_{ab}\cos\theta_{W}^{ab} + \omega_{ab}\cos\theta_{\omega}^{ab} \tag{289}$$

$$F = U(\sin\theta_{Ub} - \sin\theta_{Ua}) + W_{ab}\sin\theta_{W}^{ab} + \omega_{ab}\sin\theta_{\omega}^{ab} \tag{290}$$

The work elements for incoherent space are given by equations (56G) through (56L). The measured efficiency is given by

$$\eta_m = 1 - C \tag{291}$$

Note that $0 < C < 1$ for physical systems. As shown by equations (281) through (291) both work elements $\bar{W}_{\mu\nu}$ and generalized work elements $\bar{\omega}_{\mu\nu}$ enter directly into the calculation of the efficiency of the Brayton engine.

For small values of the internal phase angles of the thermodynamic functions the engine efficiency can be calculated by using equations (56M) through (56P) and equations (287) through (290) with the result that

$$G = U(\theta_{Ud}^2 - \theta_{Uc}^2)/2 + P_{cd}(V_d - V_c)(\theta_{Uc} + \theta_{Ud})/2 \tag{292}$$

$$- M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2 - M_{bc}(\alpha_c - \alpha_b)(\theta_{Ub} + \theta_{Uc})/2$$

$$- M_{da}(\alpha_a - \alpha_d)(\theta_{Ud} + \theta_{Ua})/2$$

$$H = U(\theta_{Uc} - \theta_{Ud}) - P_{cd}(V_d - V_c) + M_{ab}(\alpha_b - \alpha_a) \tag{293}$$

$$+ M_{bc}(\alpha_c - \alpha_b) + M_{da}(\alpha_a - \alpha_d)$$

$$E = U(\theta_{Ua}^2 - \theta_{Ub}^2)/2 - P_{ab}(V_b - V_a)(\theta_{Ua} + \theta_{Ub})/2 \tag{294}$$

$$- M_{ab}(\alpha_b - \alpha_a)(\theta_{Ua} + \theta_{Ub})/2$$

$$F = U(\theta_{Ub} - \theta_{Ua}) + P_{ab}(V_b - V_a) + M_{ab}(\alpha_b - \alpha_a) \tag{295}$$

where P_{ab} and P_{cd} are constants. The efficiency is then calculated using equations (284) through (291). As before the quantities $(V_\mu - V_\nu)$ and $(\alpha_\mu - \alpha_\nu)$ must be first order homogeneous functions of $(\theta_{U\mu} \pm \theta_{U\nu})$.

B. Internal Phase Brayton Engine for Coherent Spacetime.

The working substance for the coherent spacetime version of the ultrafast coherent heat Brayton engine is a gas of high-T_c superconducting electron pairs. From Figure 6c it follows that the variables and constants for each path segment of this type of engine cycle are given by (Case C of Section 2)

	Variables	Constants	
path ab	$\theta_S , \theta_U , T , \theta_V , \theta_\alpha$	$S , U , P_{ab} , \theta_P^{ab} , V , \alpha$	(296)
path bc	$\theta_U , \theta_P , P , T , \theta_V , \theta_\alpha$	$S , U , \theta_S^{bc} , V , \alpha$	(297)
path cd	$\theta_S , \theta_U , T , \theta_V , \theta_\alpha$	$S , U , P_{cd} , \theta_P^{cd} , V , \alpha$	(298)
path da	$\theta_U , \theta_P , P , T , \theta_V , \theta_\alpha$	$S , U , \theta_S^{da} , V , \alpha$	(299)

The work elements $\bar{W}_{\mu\nu}$ and the generalized work elements $\bar{\omega}_{\mu\nu}$ for coherent spacetime are obtained from equations (66) and (67) respectively. The work elements that enter directly into the calculation of the engine efficiency through equations (287) through (290) are for coherent spacetime now given by

$$\bar{W}_{ab} = V\bar{P}_{ab}(\theta_{Vb} - \theta_{Va}) \qquad W_{ab} = VP_{ab}(\theta_{Vb} - \theta_{Va}) \qquad \theta_W^{ab} = \theta_P^{ab} \tag{300}$$

$$\bar{W}_{cd} = V\bar{P}_{cd}(\theta_{Vd} - \theta_{Vc}) \qquad W_{cd} = VP_{cd}(\theta_{Vd} - \theta_{Vc}) \qquad \theta_W^{cd} = \theta_P^{cd} \tag{301}$$

where $\bar{P}_{ab}$ and $\bar{P}_{cd}$ are constants which may be obtained from equation (60). The elements of generalized work are written as

$$\bar{\omega}_{ab} = \alpha\bar{M}_{ab}(\theta_{\alpha b} - \theta_{\alpha a}) \qquad \omega_{ab} = \alpha M_{ab}(\theta_{\alpha b} - \theta_{\alpha a}) \qquad \theta_\omega^{ab} = \theta_M^{ab} \tag{302}$$

$$\bar{\omega}_{bc} = \alpha\bar{M}_{bc}(\theta_{\alpha c} - \theta_{\alpha b}) \qquad \omega_{bc} = \alpha M_{bc}(\theta_{\alpha c} - \theta_{\alpha b}) \qquad \theta_\omega^{bc} = \theta_M^{bc} \tag{303}$$

$$\bar{\omega}_{da} = \alpha\bar{M}_{da}(\theta_{\alpha a} - \theta_{\alpha d}) \qquad \omega_{da} = \alpha M_{da}(\theta_{\alpha a} - \theta_{\alpha d}) \qquad \theta_\omega^{da} = \theta_M^{da} \tag{304}$$

where $\bar{M}_{ab}$, $\bar{M}_{bc}$ and $\bar{M}_{da}$ are average values over the respective path segments. With these changes, as in equations (67A) through (67H), the set of equations (270) through (295) can be used to calculate the efficiency of the ultrafast coherent heat Brayton engine for coherent spacetime. The terms $\theta_{V\mu} - \theta_{V\nu}$ and $\theta_{\alpha\mu} - \theta_{\alpha\nu}$ must be first order homogeneous functions of $\theta_{U\mu} \pm \theta_{U\nu}$.

9. CONCLUSION. It is possible in theory to develop cyclic engines that convert heat in the form of internal phase (coherent heat) into useful external work. If ordinary matter is used as a working substance then the rotation of the entropy and internal energy vectors in internal space can produce a pressure and a change in the magnitude of the volume of space that contains the matter. A simple mechanical analogy of this effect is the volume change and pressure created by the shearing of a granular material.[17] If the working substance of the engine is a gas of high-T_C superconducting Cooper electron pairs that move in coherent spacetime, then the transfer of internal phase energy and internal phase entropy during an engine cycle must be accompanied by a change in the internal phase angles of the spacetime coordinates within the working chamber. In other words, the working chamber volume will be sheared (at constant volume magnitude) in space and time during the operation of the engine, and a pressure and external work will be developed. This coherent spacetime

coherent heat engine is in fact a practical example of vacuum engineering. The engineered vacuum has already been discussed in particle physics.[18] Ultrafast coherent heat engines operate within the limits of the first and second laws of thermodynamics and have measurable efficiencies that are always less than unity. The low work output per cycle of these engines may be compensated by their ultrafast nature which may produce high power outputs. The ultrafast cycles considered in this paper may be applicable to the dynamic processes that occur in the interaction of molecules with ultrafast light pulses.[19-22] These cycles may also have application to the study of the energetics of thermodynamic processes associated with the storage and retrieval of information in the memories of high speed supercomputers.

ACKNOWLEDGEMENT

The author would like to thank Elizabeth K. Klein for typing this paper.

REFERENCES

1. Duncan, J., Steam and Other Engines, MacMillan, New York, 1909.

2. Mackey, C. O., Barnard, W. N. and Ellenwood, F. O., Engineering Thermodynamics, John Wiley, New York, 1957.

3. Keenan, J. H., Thermodynamics, John Wiley, New York, 1941.

4. Schmidt, E., Thermodynamics, Oxford Univ. Press, New York, 1949.

5. Fernald, E. M., Elements of Thermodynamics, McGraw-Hill, New York, 1931.

6. Hayes, A. E. J., Applied Thermodynamics, Pergamon, New York, 1963.

7. Kiefer, P. J., Kinney, G. F. and Stuart, M. C., Principles of Engineering Thermodynamics, John Wiley, New York, 1954.

8. Faires, V. M., Thermodynamics, MacMillan, New York, 1962.

9. Doolittle, J. S. and Zerban, A. H., Engineering Thermodynamics, International Textbook Co., Scranton, 1962.

10. Cambell, A. S., Thermodynamic Analysis of Combustion Engines, Krieger Publishing Co., New York, 1985.

11. Weiss, R. A., Relativistic Thermodynamics, Exposition Press, New York, 1976.

12. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.

13. Prandtl, L. and Tietjens, O. G., Applied Hydro- and Aeromechanics, Dover, New York, 1934.

14. Birkhoff, G., Hydrodynamics, Dover, New York, 1950.

15. Salmelin, R. H., Salomaa, M. M. and Mineev, V. P., "Internal Magnus Effects in Superfluid ^{3}He-A," Phys. Rev. Lett., 63, 868, 21 Aug. 1989.

16. Davis, R. L., "Quantum Turbulence," Phys. Rev. Lett., 64, 2519, 21 May 1990.

17. Onoda, G. Y. and Liniger, E. G., "Random Loose Packings of Uniform Spheres and the Dilatancy Onset," Phys. Rev. Lett., 64, 2727, 28 May 1990.

18. Lee, T. D., Particle Physics and Introduction to Field Theory, Harwood Academic Publishers, New York, 1981.

19. Fork, R. L., Avramopoulos, H. and Valdmanis, J. A., "Ultrashort Light Pulses," American Scientist, 78, 216, May-June 1990.

20. Binder, R., Koch, S. W., Lindberg, M. and Peyghambarian, N., "Ultrafast Adiabatic Following in Semiconductors," Phys. Rev. Lett., 65, 899, 13 Aug 1990.

21. Zewail, A. H., "The Birth of Molecules," Scientific American, 76, Dec. 1990.

22. Grinberg, A. A., "Nonstationary Quasiperiodic Energy Distribution of an Electron Gas upon Ultrafast Thermal Excitation," Phys. Rev. Lett., 65, 1251, 3 Sept. 1990.

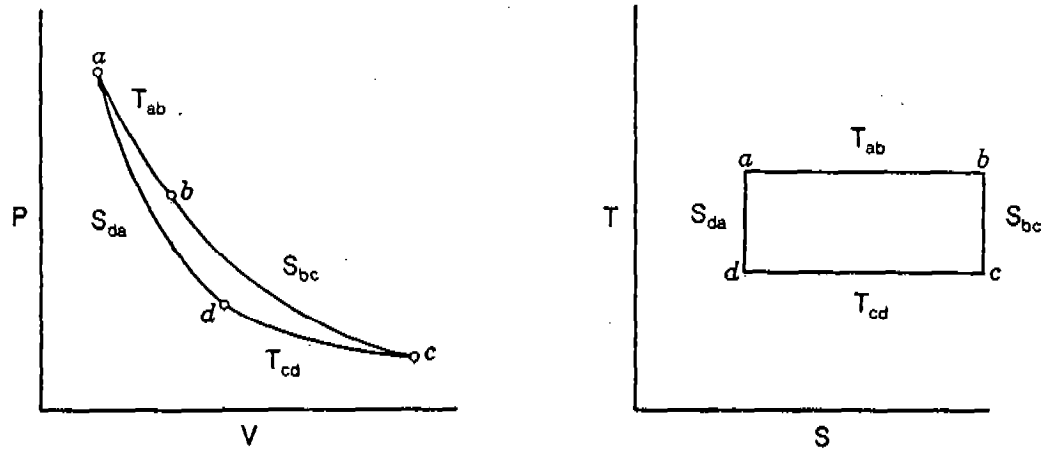


Figure 1a. Standard Carnot cycle.

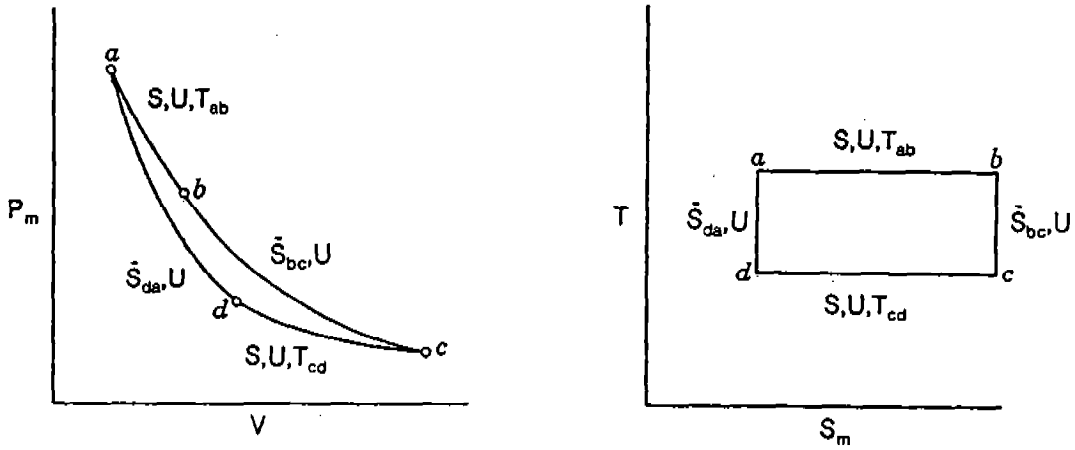


Figure 1b. Internal phase Carnot cycle for incoherent spacetime. Note, $\bar{S}_{da} = (S, \theta_S^{da})$, $\bar{S}_{bc} = (S, \theta_S^{bc})$, $P_m = P \cos \theta_P$, $S_m = S \cos \theta_S$.

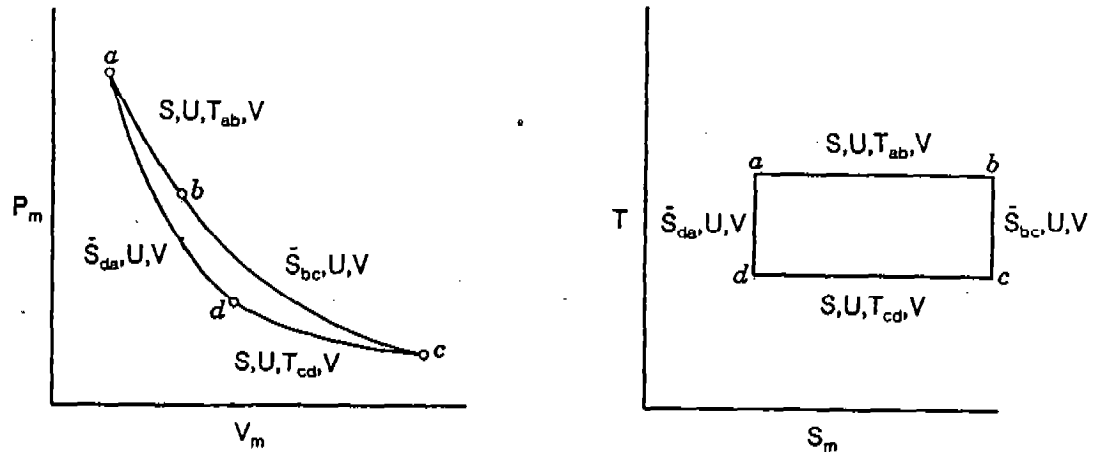


Figure 1c. Internal phase Carnot cycle for coherent spacetime. Note, $V_m = V \cos \theta_V$.

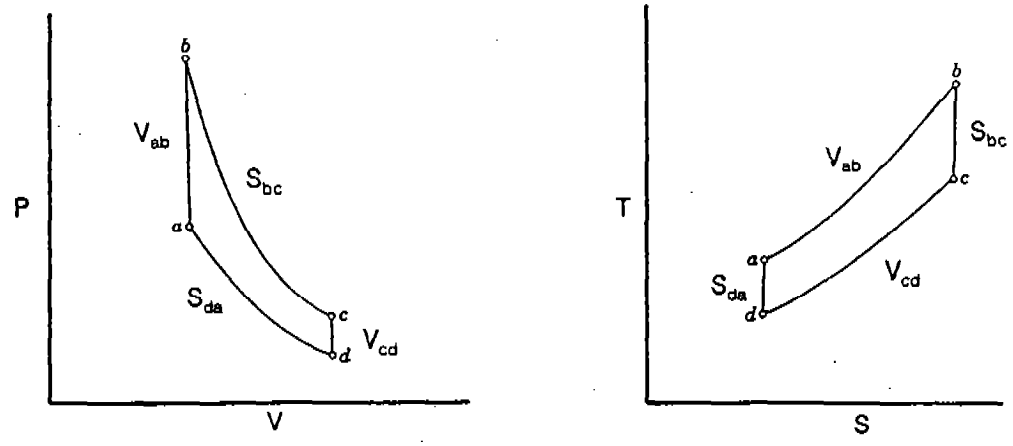


Figure 2a. Standard closed Otto cycle.

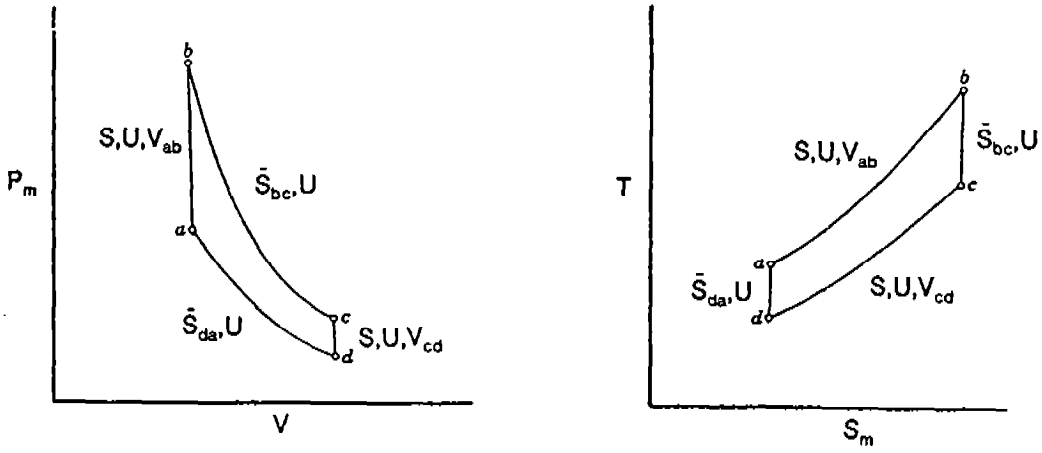


Figure 2b. Internal phase Otto cycle for incoherent spacetime. Note, $\overline{S}_{da} = (S, \theta_S^{da})$, $\overline{S}_{bc} = (S, \theta_S^{bc})$, $P_m = P \cos \theta_P$, $S_m = S \cos \theta_S$.

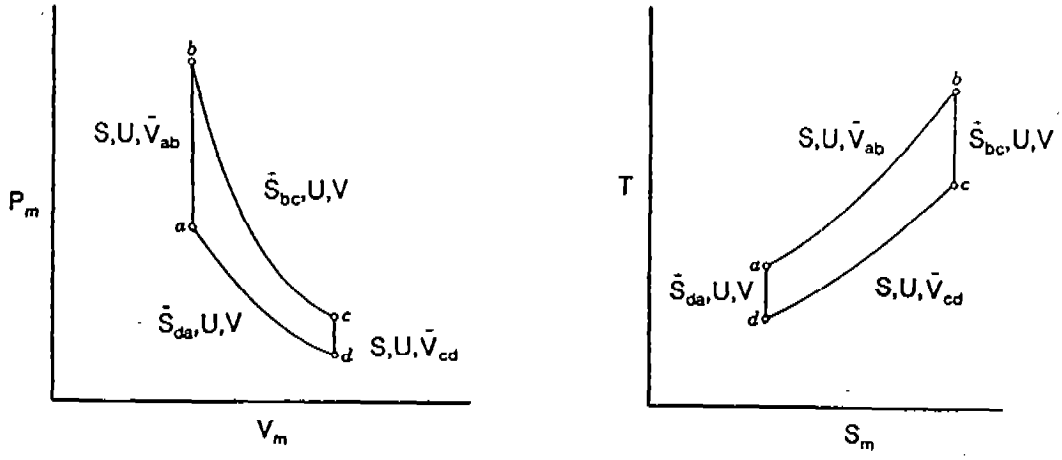


Figure 2c. Internal phase Otto cycle for coherent spacetime. Note, $\overline{V}_{ab} = (V, \theta_V^{ab})$, $\overline{V}_{cd} = (V, \theta_V^{cd})$, $V_m = V \cos \theta_V$.

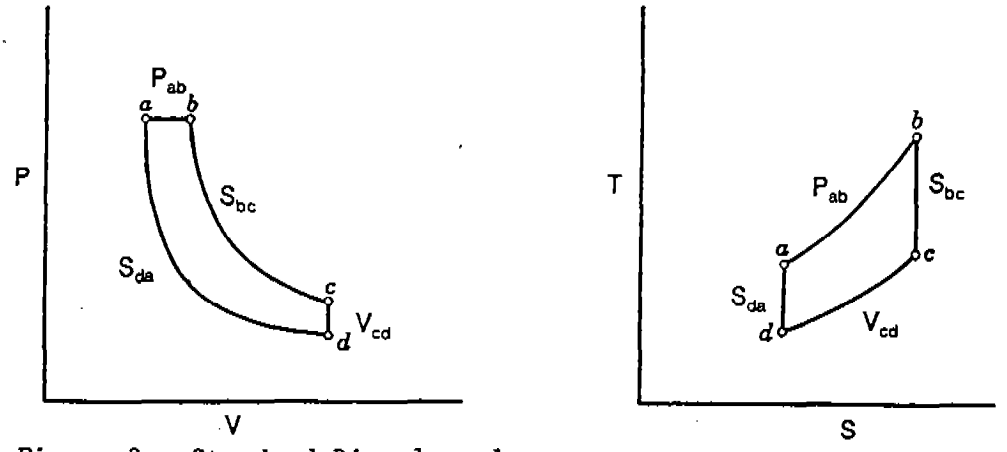


Figure 3a. Standard Diesel cycle.

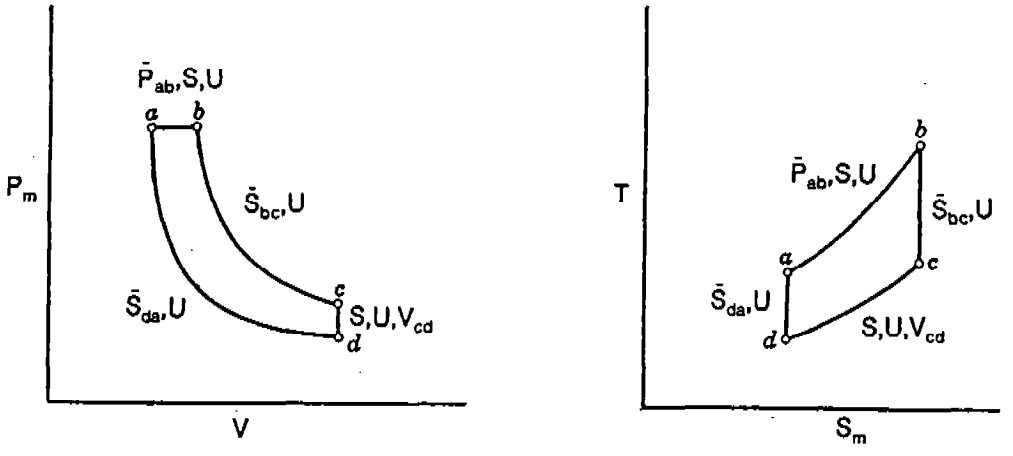


Figure 3b. Internal phase Diesel cycle for incoherent spacetime. Note, $\overline{S}_{da} = (S, \theta_S^{da})$, $\overline{S}_{bc}(S, \theta_S^{bc})$, $\overline{P}_{ab}(P_{ab}, \theta_P^{ab})$, $P_m = P \cos \theta_P$, $S_m = S \cos \theta_S$.

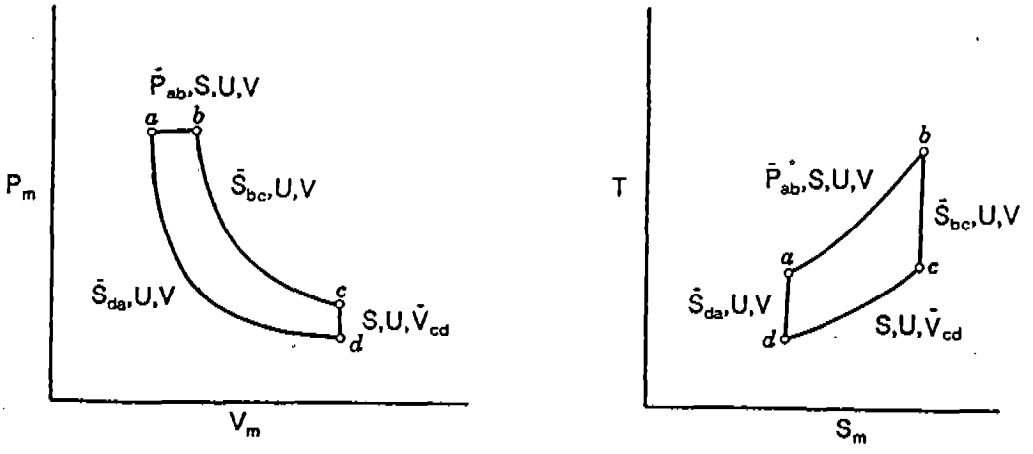


Figure 3c. Internal phase Diesel engine for coherent spacetime. Note, $\overline{V}_{cd} = (V, \theta_V^{cd})$, $V_m = V \cos \theta_V$.

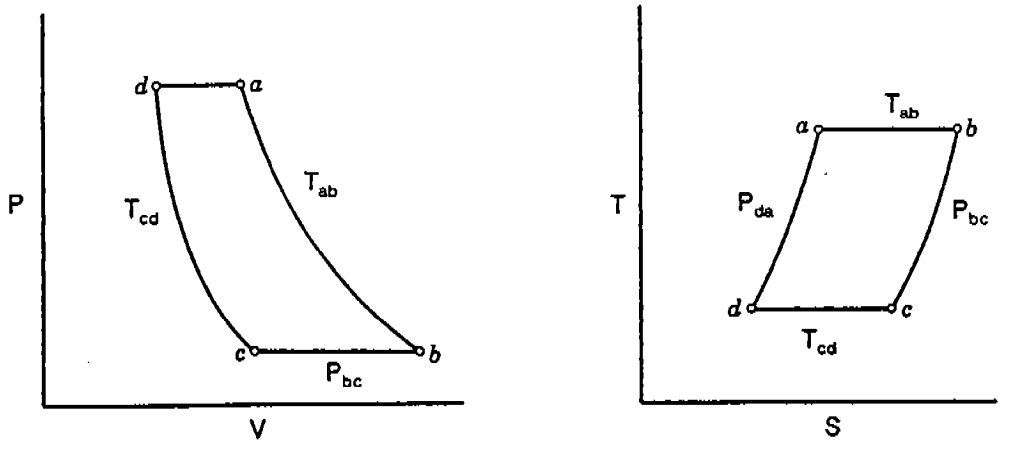


Figure 4a. Standard Ericsson cycle.

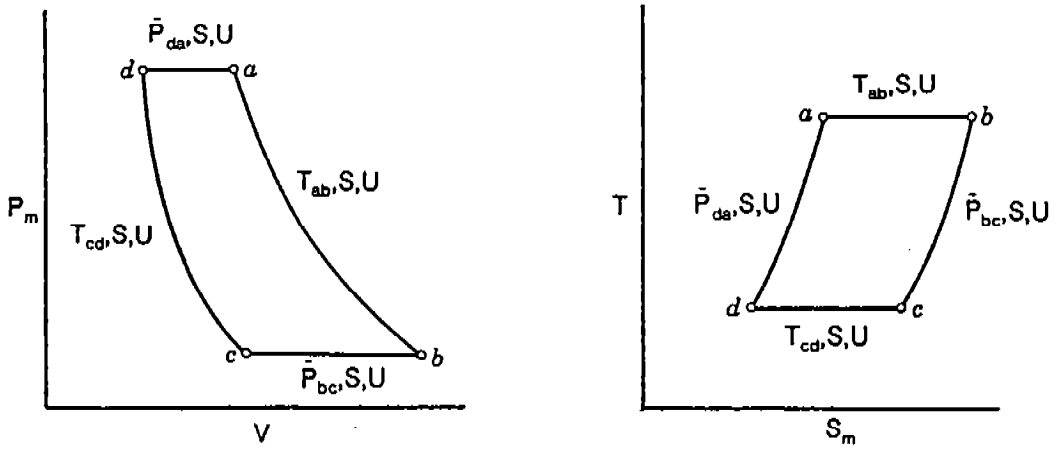


Figure 4b. Internal phase Ericsson cycle for incoherent spacetime. Note, $\bar{P}_{da} = (P_{da}, \theta_P^{da})$, $\bar{P}_{bc} = (P_{bc}, \theta_P^{bc})$, $P_m = P \cos \theta_P$, $S_m = S \cos \theta_S$.

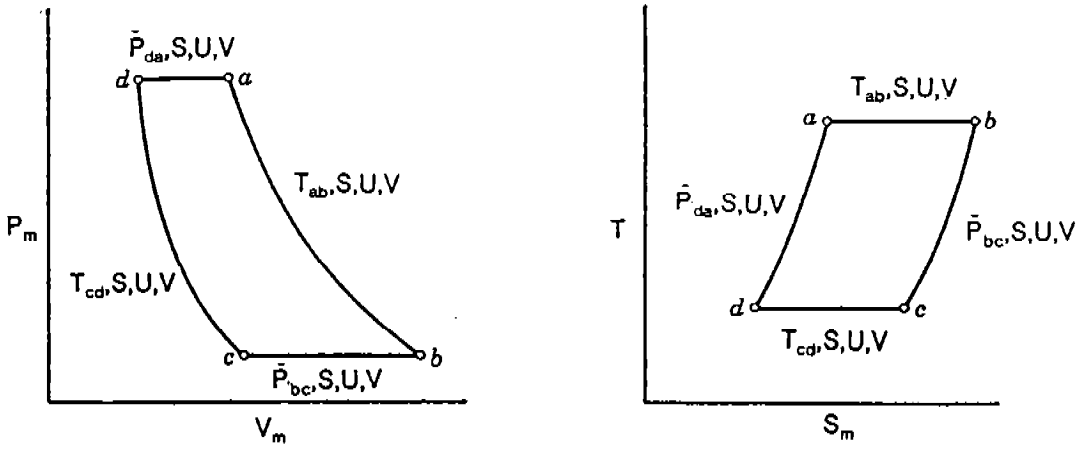


Figure 4c. Internal phase Ericsson cycle for coherent spacetime. Note, $V_m = V \cos \theta_V$.

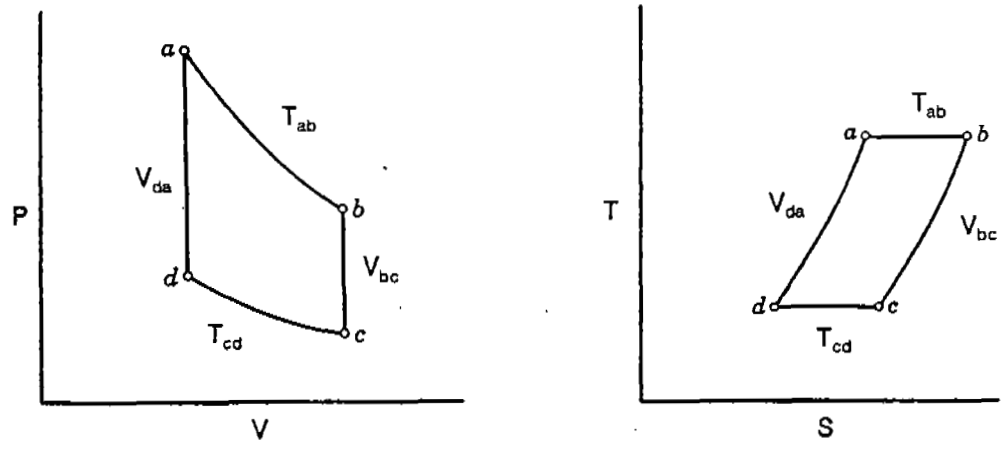


Figure 5a. Conventional Stirling engine cycle.

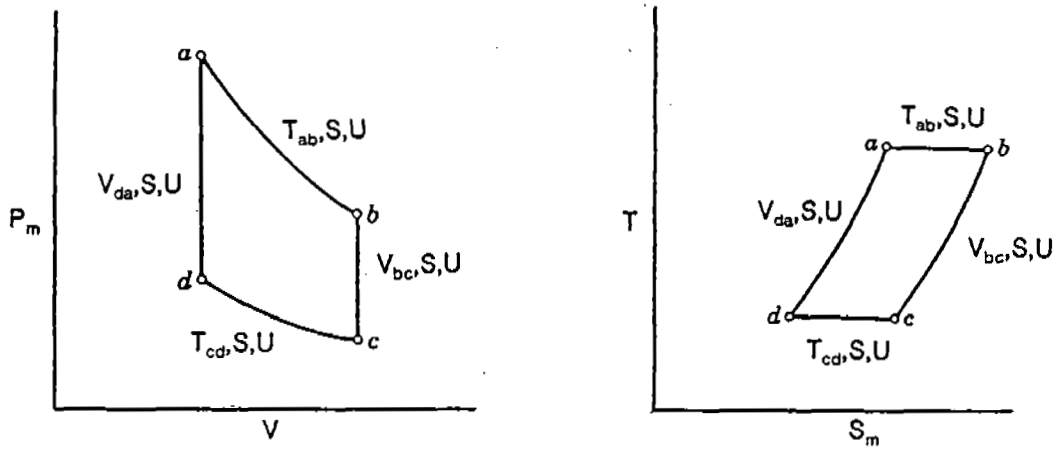


Figure 5b. Internal phase Stirling cycle for incoherent spacetime.
Note, $P_m = P \cos \theta_P$, $S_m = S \cos \theta_S$.

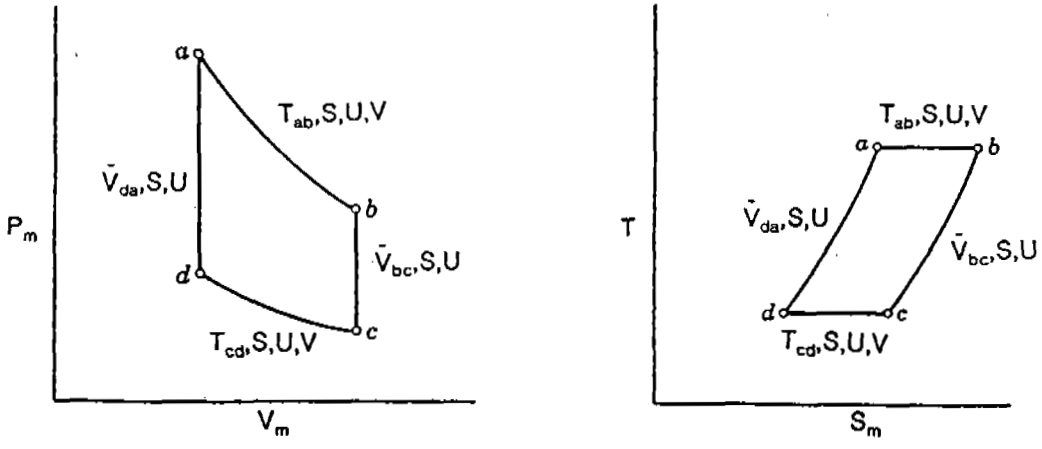


Figure 5c. Internal phase Stirling cycle for coherent spacetime.
Note, $\bar{V}_{da} = (V, \theta_V^{da})$, $\bar{V}_{bc} = (V, \theta_V^{bc})$, $V_m = V \cos \theta_V$.

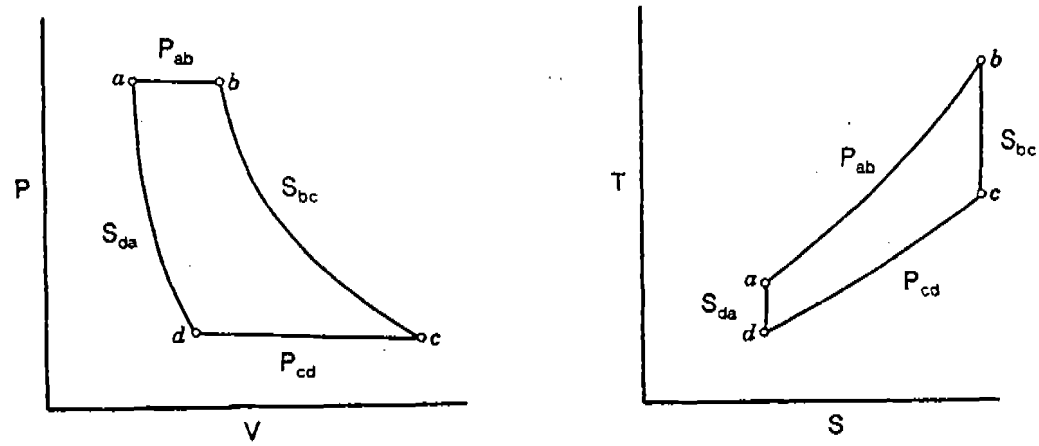


Figure 6a. Conventional Brayton cycle.

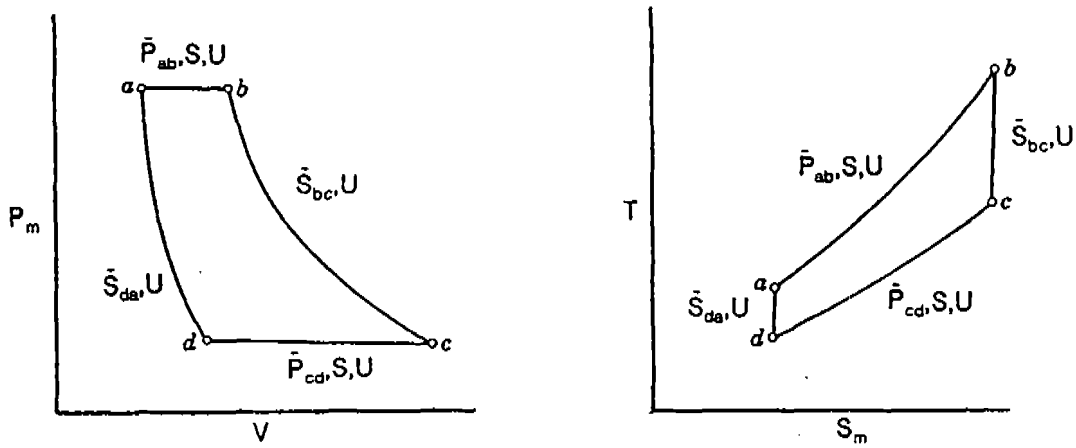


Figure 6b. Internal phase Brayton cycle for incoherent spacetime.
Note, $\bar{S}_{bc} = (S,\theta_S^{bc})$, $\bar{S}_{da} = (S,\theta_S^{da})$, $\bar{P}_{ab} = (P_{ab},\theta_P^{ab})$, $\bar{P}_{cd} = (P_{cd},\theta_P^{cd})$, $P_m = P\cos\theta_P$, $S_m = S\cos\theta_S$.

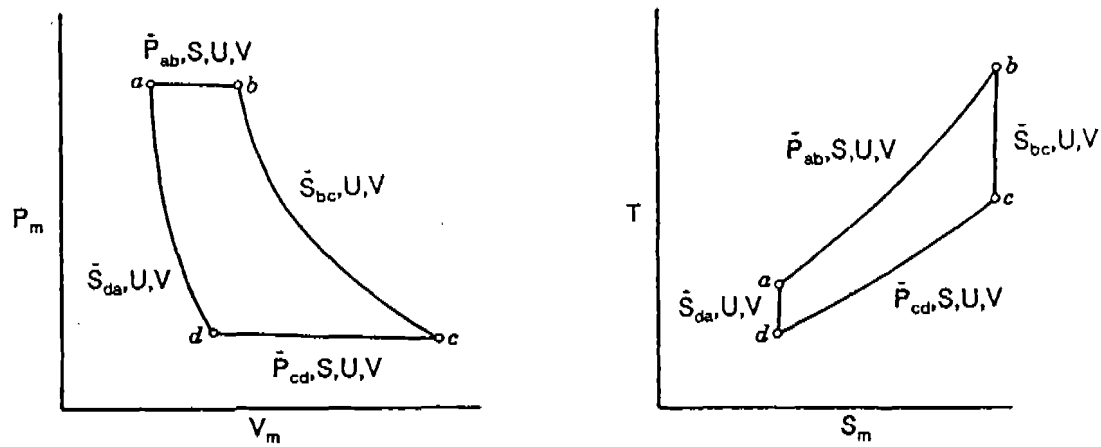


Figure 6c. Internal phase Brayton cycle for coherent spacetime.
Note, $V_m = V\cos\theta_V$.

THERMODYNAMICS AND GRAVITY

Richard A. Weiss
U. S. Army Engineer Waterways Experiment Station
Vicksburg, Mississippi 39180

ABSTRACT. A calculation of the effects of a gravitational field on the state equations of real materials is presented. The effects arise from the broken symmetry of the spacetime coordinates of the region in which the gravity field is located. The form of the laws of thermodynamics for matter located in a gravity field is investigated by applying the broken spacetime symmetry forms of the first and second laws of thermodynamics. In a gravity field the laws of thermodynamics are dependent on the internal phase angles $\theta_r(r)$ of the radial coordinates of a gravitating mass, and this dependence can also be expressed in terms of the gravitational constant $G_r(r)$ whose value depends on radial distance. In this way the effects of a gravity field on the state equations of real gases, solids and quantum liquids is determined. The Debye theory of solids is generalized to include the case of a solid located in a gravity field. This paper suggests thermodynamic measurements that can be used to determine $G_r(r)$ for a planetary gravity field. Because the internal phase angles of the radial coordinates are related to the photon redshift in a gravity field, it is suggested that the photon redshift may be used to determine the variation of $G_r(r)$ with radial distance. The sensitivity of the state equations of matter to an ambient gravity field suggests that the state equations of matter in astronomical compact objects may be considerably different from conventional predictions.

1. INTRODUCTION. Gravitating matter appears in the form of galaxies, stars and planets. Stars are gravitating systems of gases such as hydrogen, helium and smaller amounts of heavier elements. The inner planets are composed of solids with liquid cores, while the outer planets are gaseous with liquid cores. The stability of these systems results from a balance of the outwardly directed pressure and the inwardly directed Newtonian gravitational force.[1-9] This is true also in general relativity theory as described by the Tolman-Oppenheimer-Volkoff equations.[10-12] The possibility of non-Newtonian effects in gravity has also been considered and searches for these smaller corrections to Newtonian gravity are still in progress.[13-29] An explanation of the apparent non-Newtonian behaviour of gravity has been given in terms of broken spacetime symmetries which can be related to the skewed nature of pressure in an internal space.[27] The broken symmetries of spacetime and pressure suggest that the measured gravitational constant of the gravity field of a planet or star should vary with the radial distance from the center of a gravitating body.[27] This paper suggests that a determination of the gravitational constant $G_r(r)$ for the earth can be made from simple thermodynamic measurements that are performed on solids, liquids and gases at various radial distances from the earth's center. In addition, the broken spacetime symmetry in a gravity field suggests that $G_r(r)$ can be determined from photon redshift measurements.

It has been suggested that spacetime has a broken symmetry and that space

and time coordinates have internal phase angles associated with them and can be written as complex numbers in the following manner[27]

$$\bar{x} = xe^{j\theta_x} \qquad \bar{y} = ye^{j\theta_y} \qquad \bar{z} = ze^{j\theta_z} \qquad \bar{t} = te^{j\theta_t} \tag{1}$$

For spherical coordinates the complex number spatial coordinates are written as[27]

$$\bar{r} = re^{j\theta_r} \qquad \bar{\psi} = \psi e^{j\theta_\psi} \qquad \bar{\phi} = \phi e^{j\theta_\phi} \tag{2}$$

Then a volume element can be written as

$$\bar{V} = Ve^{j\theta_V} \tag{3}$$

and the differential change of the volume element is[27]

$$d\bar{V} = e^{j\theta_V}(dV + jVd\theta_V) \tag{4}$$

Equation (4) can also be written as

$$d\bar{V} = e^{j(\theta_V+\beta_{VV})} \sec \beta_{VV}\, dV = e^{j(\theta_V+\beta_{VV})} \csc \beta_{VV}\, Vd\theta_V \tag{5}$$

$$\tan \beta_{VV} = V\partial\theta_V/\partial V \tag{6}$$

The magnitude of a volume element in a gravitational field can be written as[27]

$$V_G = \int|d\bar{V}| = \int \sec \beta_{VV}\, dV = \int \csc \beta_{VV}\, Vd\theta_V \tag{7A}$$

$$\sim V \sec \beta_{VV} \qquad \beta_{VV} \sim 0 \tag{7B}$$

$$\sim V\theta_V \csc \beta_{VV} \qquad \beta_{VV} \sim \pi/2 \tag{7C}$$

where the approximation in equation (7B) holds for incoherent space and the approximation in equation (7C) holds for coherent space. For cartesian coordinates the volume element is[27,29]

$$d\bar{V} = d\bar{x}d\bar{y}d\bar{z} \qquad dV = dxdydz \qquad V = \int dxdydz \tag{8}$$

$$dV_G = |d\bar{V}| = \sec \beta_{xx} \sec \beta_{yy} \sec \beta_{zz}\, dxdydz \tag{9}$$

$$= \sec \beta_{xx} \sec \beta_{yy} \sec \beta_{zz}\, dV$$

$$= \csc \beta_{xx} \csc \beta_{yy} \csc \beta_{zz}\, xyz\, d\theta_x d\theta_y d\theta_z$$

where[27]

$$\tan \beta_{xx} = x\partial\theta_x/\partial x \qquad \tan \beta_{yy} = y\partial\theta_y/\partial y \qquad \tan \beta_{zz} = z\partial\theta_z/\partial z \tag{10}$$

The functions of β_{xx}, β_{yy}, β_{zz} and β_{VV} can depend on the spatial coordinates or be functions only of the local density as in the case of the approximations in equations (7B) and (7C).

For spherical polar coordinates[27]

$$d\bar{V} = \bar{r}^2 \sin\bar{\psi}\, d\bar{\psi}\, d\bar{\phi}\, d\bar{r} \qquad \sin\bar{\psi} = S_\psi e^{j\theta s\psi} \tag{11}$$

From equations (2), (5) and (11) it follows that

$$\begin{aligned} dV_G &= \sec\beta_{\psi\psi} \sec\beta_{\phi\phi} \sec\beta_{rr}\, r^2 S_\psi\, d\psi d\phi dr \\ &= \sec\beta_{VV}\, dV \\ &= \csc\beta_{\psi\psi} \csc\beta_{\phi\phi} \csc\beta_{rr}\, r^2 S_\psi\, \psi\phi r\, d\theta_\psi d\theta_\phi d\theta_r \\ &= \csc\beta_{VV}\, V d\theta_V \end{aligned} \tag{12}$$

$$\theta_V + \beta_{VV} = 3\theta_r + \beta_{rr} + \theta_{s\psi} + \theta_\psi + \beta_{\psi\psi} + \theta_\phi + \beta_{\phi\phi} \tag{13}$$

where for symmetrical space

$$dV = r^2 \sin\psi\, d\psi d\phi dr \qquad V = \int r^2 \sin\psi\, d\psi d\phi dr \tag{14}$$

and where[27]

$$S_\psi = [\sin^2(\psi\cos\theta_\psi) + \sinh^2(\psi\sin\theta_\psi)]^{1/2} \tag{15}$$

$$\tan\theta_{s\psi} = \cot(\psi\cos\theta_\psi)\tanh(\psi\sin\theta_\psi) \tag{16}$$

$$\tan\beta_{\psi\psi} = \psi\partial\theta_\psi/\partial\psi \qquad \tan\beta_{\phi\phi} = \phi\partial\theta_\phi/\partial\phi \qquad \tan\beta_{rr} = r\partial\theta_r/\partial r \tag{17}$$

For spherical symmetry

$$d\bar{V} = 4\pi\bar{r}^2 d\bar{r} \qquad \bar{V} = 4/3\pi\bar{r}^3 \qquad \theta_V = 3\theta_r \tag{18}$$

For a gravitating system with matter located in broken symmetry spacetime the pressure has an internal phase angle and is written in the following complex number form[27]

$$\bar{P} = Pe^{j\theta P} \tag{19}$$

The internal phase angle of the pressure is related to the internal phase angle of the radial coordinate by the following small angle approximation[27]

$$-2\theta_r \sim \theta_P + P(\partial\theta_P/\partial r)/(\partial P/\partial r) \tag{20}$$

For a particle located in a gravitational field the internal phase angle of the time and space coordinates are related by[27,29]

$$\theta_r - 2\theta_t = -2\theta_r \qquad 3\theta_r = 2\theta_t \tag{21}$$

so that the internal phase angle of the time interval associated with an event or process occurring in a gravitational field is given by[29]

$$\theta_t \sim -3/4[\theta_p + P(\partial\theta_p/\partial r)/(\partial P/\partial r)] \tag{22}$$

From equations (6), (18) and (20) it follows that under the approximation of spherical symmetry

$$\tan \beta_{VV} \sim \tan \beta_{rr} = r\partial\theta_r/\partial r \tag{23}$$

$$\sim -1/2\, r\partial/\partial r[\theta_p + P(\partial\theta_p/\partial r)/(\partial P/\partial r)]$$

$$= -1/2\, r\partial/\partial r[\theta_p + P(r\partial\theta_p/\partial r)/(r\partial P/\partial r)]$$

Note that in general $\theta_r = \theta_r(n,T)$, $\theta_p = \theta_p(n,T)$ and $P = P(n,T)$ where for a planet or star the particle number density and temperature depend on the radial distance from the center of the object in the manner $n = n(r,\phi,\psi)$ and $T = T(r,\phi,\psi)$ respectively. Therefore by the chain rule of differentiation

$$r\partial\theta_r/\partial r = \sigma n\partial\theta_r/\partial n + \mu T\partial\theta_r/\partial T \tag{24}$$

$$r\partial\theta_p/\partial r = \sigma n\partial\theta_p/\partial n + \mu T\partial\theta_p/\partial T \tag{25}$$

$$r\partial P/\partial r = \sigma n\partial P/\partial n + \mu T\partial P/\partial T \tag{26}$$

where

$$\sigma(r) = (r/n)(\partial n/\partial r) \qquad \mu(r) = (r/T)(\partial T/\partial r) \tag{27}$$

where for planets and stars $\sigma < 0$ and $\mu < 0$ generally, and for the solid earth and for the earth's atmosphere (low to medium temperature real gas)[27]

$$\theta_r < 0 \qquad \theta_p > 0 \qquad \partial\theta_r/\partial r > 0 \qquad \partial\theta_p/\partial r < 0 \tag{27A}$$

Combining equations (20), (22) and (23) with equations (24) through (26) gives

$$\theta_r \sim -1/2[\theta_p + P(\sigma n\partial\theta_p/\partial n + \mu T\partial\theta_p/\partial T)/(\sigma n\partial P/\partial n + \mu T\partial P/\partial T)] \tag{28}$$

$$\theta_t \sim -3/4[\theta_p + P(\sigma n\partial\theta_p/\partial n + \mu T\partial\theta_p/\partial T)/(\sigma n\partial P/\partial n + \mu T\partial P/\partial T)] \tag{29}$$

$$\tan \beta_{VV} \sim \tan \beta_{rr} = \sigma n\partial\theta_r/\partial n + \mu T\partial\theta_r/\partial T \tag{30}$$

Substituting equation (28) into equation (30) gives

$$\tan \beta_{rr} \sim \beta_{rr} = -1/2[\sigma(A + B + C - D) + \mu(E + F + G - H)] \tag{31}$$

where

$$A = n\partial\theta_p/\partial n \tag{32}$$

$$B = (n\partial P/\partial n)(\sigma n\partial\theta_p/\partial n + \mu T\partial\theta_p/\partial T)/J \tag{33}$$

$$C = P(\sigma n\partial\theta_p/\partial n + \sigma n^2\partial^2\theta_p/\partial n^2 + \mu nT\partial^2\theta_p/\partial n\partial T)/J \tag{34}$$

$$D = P(\sigma n\partial\theta_p/\partial n + \mu T\partial\theta_p/\partial T)(\sigma n\partial P/\partial n + \sigma n^2\partial^2 P/\partial n^2 + \mu nT\partial^2 P/\partial n\partial T)/J^2 \tag{35}$$

$$E = T\partial\theta_p/\partial T \tag{36}$$

$$F = (T\partial P/\partial T)(\sigma n\partial\theta_p/\partial n + \mu T\partial\theta_p/\partial T)/J \tag{37}$$

$$G = P(\mu T\partial\theta_p/\partial T + \mu T^2\partial^2\theta_p/\partial T^2 + \sigma nT\partial^2\theta_p/\partial n\partial T)/J \tag{38}$$

$$H = P(\sigma n\partial\theta_p/\partial n + \mu T\partial\theta_p/\partial T)(\mu T\partial P/\partial T + \mu T^2\partial^2 P/\partial T^2 + \sigma nT\partial^2 P/\partial n\partial T)/J^2 \tag{39}$$

and where

$$J = \sigma n\partial P/\partial n + \mu T\partial P/\partial T \tag{40}$$

At the earth's surface $\theta_r \sim -5.7°$ and $\theta_t \sim -8.6°$.[27,29] By taking both terms on the right hand side of equation (20) to be equal it follows that $\theta_p \sim -\theta_r \sim 5.7°$.

This paper determines the effects of a gravity field on the state equations of gases, liquids, solids and quantum liquids. The effects occur because gravity induces internal phase angles in the space and time coordinates, volume and pressure which are used to characterize matter. Briefly the paper is organized as follows: Section 2 considers the general effects of a gravity field on the laws of thermodynamics, Section 3 studies the relationship of gravity and the broken symmetry of space, Section 4 investigates the specific effects of gravity on the state equation of real gases, Section 5 determines the effects of a gravity field on the state equations of solids and quantum liquids, and finally Section 6 ascertains the structure of the Debye theory of solids located in a gravity field.

2. THERMODYNAMICS OF MATTER IN A GRAVITY FIELD. This section treats the basic thermodynamic formalism that describes matter in a broken symmetry spacetime. The thermodynamic relations that are derived in this section will be used in Sections 4 through 6 to describe the effects of gravity on the state equations of real gases, solids and quantum liquids. These calculations also describe manmade broken symmetry states of the vacuum that may be induced in the laboratory using electromagnetic fields. The gauge invariant and conformal invariant relativistic trace equation for matter in a broken symmetry spacetime that is induced by a gravity field is[27,29]

$$\bar{U}_G + T(d\bar{U}_G/dT)_{\bar{P}_G V_G} - 3V_G d/dV_G(\bar{P}_G V_G)_{\bar{U}_G} = \bar{U}_G^a + T(d\bar{U}_G^a/dT)_{\bar{P}_G^a V_G} \tag{41}$$

where $\bar{U}_G$ and $\bar{U}_G^a$ = renormalized and unrenormalized internal energy respectively of a body located in a gravity field, $\bar{P}_G$ and $\bar{P}_G^a$ = renormalized and unrenormalized pressure respectively of matter located in a gravity field, and where V_G = volume of matter in broken symmetry space due to a gravity field and given by equation (7). The unrenormalized internal energy is assumed to be affected by a gravity field only by the addition of a constant term which may be taken to be zero so that

$$\bar{U}_G^a = \bar{U}^a \qquad \bar{P}_G^a = \bar{P}^a \tag{42}$$

where $\bar{U}^a$ and $\bar{P}^a$ = unrenormalized energy density and pressure respectively for matter in the absence of a gravity field. Thus the effect of a gravity field on the source term on the right hand side of equation (41) comes essentially

through the broken symmetry volume element V_G. The first and second laws of thermodynamics for the broken symmetry space associated with a gravity field are written as[27]

$$\begin{aligned} T d\bar{S}_G &= d\bar{U}_G + \bar{\bar{P}}_G d\bar{V} + \bar{\bar{M}}_G d\bar{\alpha} \\ &= d\bar{U}_G + \bar{P}_G dV_G + \bar{M}_G d\alpha_G \\ &= d\bar{U}_G + \bar{P}_G \sec \beta_{VV}\, dV + \bar{M}_G \sec \beta_{\alpha\alpha}\, d\alpha \end{aligned} \tag{43}$$

where β_{VV} is given in equation (5) while $\beta_{\alpha\alpha}$ is given by

$$\tan \beta_{\alpha\alpha} = \alpha \partial\theta_\alpha / \partial\alpha \qquad d\alpha_G = |d\bar{\alpha}| = \sec \beta_{\alpha\alpha}\, d\alpha \tag{44}$$

where α = generalized coordinate, α_G = generalized coordinate of matter in a gravity field, $\bar{\bar{M}}_G$ and $\bar{M}_G$ = two representations of the complex number generalized force and $\bar{\bar{P}}_G$ and $\bar{P}_G$ = two representations of the complex number pressure of matter in a gravity field. In analogy to equation (3) the generalized coordinate is written as

$$\bar{\alpha} = \alpha e^{j\theta\alpha} \qquad d\bar{\alpha} = e^{j\theta_\alpha}(d\alpha + j\alpha d\theta_\alpha) \tag{45}$$

From equation (43) and neglecting the generalized force it follows that

$$T \partial \bar{S}_G / \partial V_G = \partial \bar{U}_G / \partial V_G + \bar{P}_G \tag{46}$$

$$\partial \bar{S}_G / \partial V_G = \partial \bar{P}_G / \partial T \tag{47}$$

$$\partial \bar{U}_G / \partial V_G = T \partial \bar{P}_G / \partial T - \bar{P}_G \tag{48}$$

Using equation (7) allows equations (46) through (48) to be rewritten as

$$\cos \beta_{VV}\, T \partial \bar{S}_G / \partial V = \cos \beta_{VV}\, \partial \bar{U}_G / \partial V + \bar{P}_G \tag{49}$$

$$\cos \beta_{VV}\, \partial \bar{S}_G / \partial V = \partial \bar{P}_G / \partial T \tag{50}$$

$$\cos \beta_{VV}\, \partial \bar{U}_G / \partial V = T \partial \bar{P}_G / \partial T - \bar{P}_G \tag{51}$$

The corresponding symmetrical space equations, which are valid in the absence of a gravitational field and in the absence of any structurally induced broken spacetime symmetry, are written as

$$T \partial \bar{S} / \partial V = \partial \bar{U} / \partial V + \bar{P} \tag{52}$$

$$\partial \bar{S} / \partial V = \partial \bar{P} / \partial T \tag{53}$$

$$\partial \bar{U} / \partial V = T \partial \bar{P} / \partial T - \bar{P} \tag{54}$$

The broken symmetry of the thermodynamic functions is related to the speed at which thermodynamic processes occur; fast processes have broken symmetry thermodynamic functions while the internal phase angles of the thermodynamic func-

tions can be taken to have zero value for slow processes.

The real and imaginary parts of equations (46) through (54) can be taken in order to determine the relationship between phase angles and magnitudes. For instance, equation (50) is equivalent to

$$\theta_S^G + \beta_{SS}^G = \theta_P^G + \beta_{PP}^G \tag{55}$$

$$\cos^2\beta_{VV}\,[(\partial S_G/\partial V)^2 + (S_G\partial\theta_S^G/\partial V)^2] = (\partial P_G/\partial T)^2 + (P_G\partial\theta_P^G/\partial T)^2 \tag{56}$$

where

$$\tan\beta_{SS}^G = S_G\partial\theta_S^G/\partial S_G \qquad \tan\beta_{PP}^G = P_G\partial\theta_P^G/\partial P_G \tag{57}$$

The thermodynamic functions U_G, S_G, P_G, θ_U^G, θ_S^G and θ_P^G depend on the variables T, V and θ_V. Equations (46) through (48) are homologous to equations (52) through (54) so that the state equations for matter located in broken symmetry spacetime can be obtained from the state equations for matter in symmetrical spacetime by making the substitutions $V \to V_G$ and $dV \to dV_G = dV \sec\beta_{VV}$. Thus all state equations for which the pressure depends on particle number density, even the ideal gas, will be affected by gravity, and the effects can be determined through the substitution $V \to V_G$ in the symmetric spacetime version of the material state equations where V_G is defined in equation (7).

A simple approximate way for calculating the pressure for a broken spacetime symmetry thermodynamic system, such as matter located in a gravity field, can be developed so that the cumbersome substitution $V \to V_G \sim V \sec\beta_{VV}$ in every term of a symmetric spacetime state equation for matter can be avoided. To do this, first note that if β_{VV} is independent of temperature then equations (49) through (51) can be written as

$$T\partial\bar{S}_G/\partial V = \partial\bar{U}_G/\partial V + \bar{P}_G' \tag{58}$$

$$\partial\bar{S}_G/\partial V = \partial\bar{P}_G'/\partial T \tag{59}$$

$$\partial\bar{U}_G/\partial V = T\partial\bar{P}_G'/\partial T - \bar{P}_G' \tag{60}$$

where in a gravity field the effective pressure $\bar{P}_G'$ is calculated by the standard form (symmetric spacetime form) of thermodynamic equations provided that

$$\bar{P}_G' = \bar{P}_G \sec\beta_{VV} \tag{61}$$

If in addition $\bar{U}_G$ and $\bar{U}$ are not greatly different in value then a comparison of equations (52) through (54) and equations (58) through (60) suggests that $\bar{P}_G' \sim \bar{P}$ and therefore equation (61) becomes

$$\bar{P}_G \sim \bar{P}\cos\beta_{VV} = \bar{P}_G^c \quad \text{or} \quad \bar{P}_G \sim \bar{P}_G^c \tag{62}$$

where the superscript c refers to the pressure calculated within the cosine approximation. If equation (62) is a reasonable approximation it becomes very useful because it allows a simple way of estimating the effects of a gravity

field (or other force that induces a broken spacetime symmetry) on the state equation of matter without knowing the form of the state equation $P = P(V,T)$ and without making the meticulous change $V \to V_G \sim V \sec \beta_{VV}$ in each term of the symmetric spacetime form of the state equation. If the approximation in equation (62) is not valid then the substitution $V \to V_G \sim V \sec \beta_{VV}$ must be made in each term of the state equation of matter for symmetric spacetime. Finally it should be pointed out that equations (49) through (51) can be written equivalently as

$$\sin \beta_{VV}\ T/V\ \partial \bar{S}_G/\partial\theta_V = \sin \beta_{VV}\ 1/V\ \partial \bar{U}_G/\partial\theta_V + \bar{P}_G \tag{63}$$

$$\sin \beta_{VV}\ 1/V\ \partial \bar{S}_G/\partial\theta_V = \partial \bar{P}_G/\partial T \tag{64}$$

$$\sin \beta_{VV}\ 1/V\ \partial \bar{U}_G/\partial\theta_V = T\partial \bar{P}_G/\partial T - \bar{P}_G \tag{65}$$

which are useful for large spacetime asymmetries.

3. GRAVITY AND THE BROKEN SYMMETRY OF SPACE. This section establishes the relationship between the fundamental laws of thermodynamics and the gravitational constant, and suggests the possibility that the value of the gravitational constant can be determined from the measurement of thermodynamic properties of matter located in the gravity field of the earth. The connection between gravity and thermodynamics is established by first determining the relationship between the internal phase angle of a volume element of matter and the internal phase angles of the spatial coordinates of the volume element, and then relating the coordinate internal phase angles to the value of the gravitational constant.

A. Determination of β_{VV} from Thermodynamic Measurements.

From equations (5), (8) and (9) it follows for cartesian coordinates that

$$\sec \beta_{VV} = \sec \beta_{xx} \sec \beta_{yy} \sec \beta_{zz} \tag{66}$$

$$\csc \beta_{VV}\ V d\theta_V = \csc \beta_{xx} \csc \beta_{yy} \csc \beta_{zz}\ xyz\ d\theta_x d\theta_y d\theta_z \tag{67}$$

From equations (5), (12) and (14) it follows for spherical polar coordinates that

$$\sec \beta_{VV} = (S_\psi/\sin\psi)\sec \beta_{\psi\psi} \sec \beta_{\phi\phi} \sec \beta_{rr} \tag{68}$$

$$\csc \beta_{VV}\ V d\theta_V = \csc \beta_{\psi\psi} \csc \beta_{\phi\phi} \csc \beta_{rr}\ r^2 S_\psi\ \psi\phi r\ d\theta_\psi d\theta_\phi d\theta_r \tag{69}$$

If $\theta_\phi = 0$ and $\theta_\psi = 0$ then equation (68) gives

$$\sec \beta_{VV} = \sec \beta_{rr} \qquad \beta_{VV} = \beta_{rr} \tag{70}$$

In this case equation (70) is exact but often the approximation $\beta_{VV} \sim \beta_{rr}$ is used in the general case as in equation (23). In the spherically symmetric case given by equations (18) and (70) the magnitude of the complex number volume element is given by [27,29]

$$dV_G = |d\bar{V}| = 4\pi r^2 \sec \beta_{rr}\ dr \tag{71}$$

Therefore only if the zenith and azimuthal angles are scalars does the simple result $\beta_{VV} = \beta_{rr}$ hold exactly. The case $\theta_\psi << \theta_r$ and $\theta_\phi << \theta_r$ is generally true for the gravity field of the earth so that $\beta_{VV} \sim \beta_{rr}$ is a reasonable approximation.[27]

From equations (50), (51) and (66) it follows that for a broken symmetry cartesian coordinate system the form of the Gibbs-Helmholtz-Maxwell equations are

$$(\cos \beta_{xx} \cos \beta_{yy} \cos \beta_{zz})\partial \bar{U}_G/\partial V = T\partial \bar{P}_G/\partial T - \bar{P}_G \tag{72}$$

$$(\cos \beta_{xx} \cos \beta_{yy} \cos \beta_{zz})\partial \bar{S}_G/\partial V = \partial \bar{P}_G/\partial T \tag{73}$$

where $dV = dxdydz$. Equivalently the use of equations (64), (65) and (67) lets these equations be written as

$$(\sin \beta_{xx} \sin \beta_{yy} \sin \beta_{zz})/(xyz)\partial^3 \bar{U}_G/\partial\theta_x\partial\theta_y\partial\theta_z = T\partial \bar{P}_G/\partial T - \bar{P}_G \tag{74}$$

$$(\sin \beta_{xx} \sin \beta_{yy} \sin \beta_{zz})/(xyz)\partial^3 \bar{S}_G/\partial\theta_x\partial\theta_y\partial\theta_z = \partial \bar{P}_G/\partial T \tag{75}$$

For a broken symmetry spherical polar coordinate system, equations (50), (51) and (68) give the Gibbs-Helmholtz-Maxwell equations as

$$[(\sin \psi / S_\psi)\cos \beta_{\psi\psi} \cos \beta_{\phi\phi} \cos \beta_{rr}]\partial \bar{U}_G/\partial V = T\partial \bar{P}_G/\partial T - \bar{P}_G \tag{76}$$

$$[(\sin \psi / S_\psi)\cos \beta_{\psi\psi} \cos \beta_{\phi\phi} \cos \beta_{rr}]\partial \bar{S}_G/\partial V = \partial \bar{P}_G/\partial T \tag{77}$$

where $dV = r^2 \sin \psi \, d\psi d\phi dr$. Alternatively, equations (64), (65) and (69) gives

$$(\sin \beta_{\psi\psi} \sin \beta_{\phi\phi} \sin \beta_{rr})/(r^2 S_\psi \, \psi\phi r)\partial^3 \bar{U}_G/\partial\theta_\psi\partial\theta_\phi\partial\theta_r = T\partial \bar{P}_G/\partial T - \bar{P}_G \tag{78}$$

$$(\sin \beta_{\psi\psi} \sin \beta_{\phi\phi} \sin \beta_{rr})/(r^2 S_\psi \, \psi\phi r)\partial^3 \bar{S}_G/\partial\theta_\psi\partial\theta_\phi\partial\theta_r = \partial \bar{P}_G/\partial T \tag{79}$$

Within the approximation $\theta_\phi \sim 0$ and $\theta_\psi \sim 0$ for the earth's coordinate system equations (70), (76) and (77) give

$$\cos \beta_{rr} \, \partial \bar{U}_G/\partial V = T\partial \bar{P}_G/\partial T - \bar{P}_G \tag{80}$$

$$\cos \beta_{rr} \, \partial \bar{S}_G/\partial V = \partial \bar{P}_G/\partial T \tag{81}$$

In the remainder of this section the approximation in equation (70) will be assumed valid for the earth's gravity field. Therefore equations (80) and (81) and the discussion in Section 2 suggests that β_{rr} (and hence β_{VV}) can be determined from the measurement of pressure in a substance, such as a gas, at various radial distances from the earth's center.

B. Determination of G_r from Thermodynamic Measurements.

It has been shown in the literature that in broken symmetry spacetime the radial component of the Newtonian gravitation constant for a non-rotating planet is given by[27]

$$G_r = G \cos(2\theta_r) \cos^2\theta_r \tag{82}$$

$$\sim G(1 - 3\theta_r^2 + 3\theta_r^4 - \cdots) \tag{83}$$

where G_r = radial component of the Newtonian gravitation constant for broken symmetry spacetime, and G = Newtonian gravitation constant for a totally symmetric spacetime. Actually G_r along with G_ϕ and G_ψ are the three components of the gravitation constant for the earth, but in this paper G_ϕ and G_ψ are not required because it is assumed that θ_ϕ and θ_ψ are negligible compared to θ_r.[27] However, all three components of the gravity constant G_r, G_ϕ and G_ψ would be needed if the exact set of thermodynamic equations (76) and (77) are used to analyse thermodynamic measurements rather than the simplified set of equations given in equation (80) and (81). The first step in the determination of G_r from thermodynamic measurements is to use the approximate equation (83) to determine θ_r. Using only the second order terms in equation (83) gives

$$\theta_r \sim - [1/3(1 - G_r/G)]^{1/2} \tag{84}$$

and

$$1/G\ \partial G_r/\partial r \sim - 6\theta_r/r\ (r\partial\theta_r/\partial r) \tag{85}$$

$$\sim - 6\theta_r\beta_{rr}/r \tag{86}$$

where equation (86) is obtained by using the approximate form of equation (17) that is valid for small values of β_{rr}. Combining equations (84) and (86) gives

$$\beta_{rr} \sim 1/6(r/G\ \partial G_r/\partial r)/[1/3(1 - G_r/G)]^{1/2} \tag{87}$$

where $\beta_{rr} > 0$ because $\partial G_r/\partial r > 0$ for the earth. Then cos β_{rr} can be calculated by a power series approximation

$$\cos\ \beta_{rr} = 1 - \beta_{rr}^2/2 + \beta_{rr}^4/24 - \cdots \tag{88}$$

where cos β_{rr} appears in the thermodynamic equations (80) and (81) for matter in a gravity field. In this way the gravitation constants G and G_r and the radial coordinate distance r from the earth's center will enter the basic thermodynamic equations (80) and (81). Therefore thermodynamic measurements of pressure and heat capacity of matter at various radial distances from the earth's center may possibly yield values of β_{rr} and G_r and determine their variation with radial distance. Thermodynamic measurements at various elevations would yield values of $\beta_{rr}(r)$ from equations (80) and (81). Then the value of G_r would be obtained as a solution to the nonlinear differential equation (87) which can be rewritten as follows

$$(r\partial\psi/\partial r)^2 - 12\beta_{rr}^2\psi = 0 \qquad \psi = (G - G_r)/G \tag{89}$$

where $\psi \geq 0$.

C. Determination of β_{rr} and G_r from Redshift of Photon.

It has been shown in the literature that the value of the internal phase angle of the radial coordinate θ_r at any radial distance from the earth's center can be calculated from the difference between the measured value and conventionally predicted value of the gravitational redshift of a photon as follows[27]

$$\sin^2\theta_r = (z_m - z_c)/z_c \qquad \sin\ \theta_r = - [(z_m - z_c)/z_c]^{1/2} \tag{90}$$

where z_m = measured gravitational redshift, z_c = conventionally predicted gravitational redshift, and where $\theta_r < 0$.[27] If the value of θ_r is small the following approximate forms of equation (90) can be used

$$\theta_r^2 \sim (z_m - z_c)/z_c \qquad \theta_r \sim -[(z_m - z_c)/z_c]^{1/2} \tag{91}$$

For the general case combining equations (17) and (90) gives

$$\tan \beta_{rr} = r\partial\theta_r/\partial r = -r\partial/\partial r\{\sin^{-1}[(z_m - z_c)/z_c]^{1/2}\} \tag{92}$$

or using the small angle and slow variation approximations

$$\beta_{rr} \sim -r\partial/\partial r\{[(z_m - z_c)/z_c]^{1/2}\} \tag{93}$$

Note that in equations (92) and (93) $\beta_{rr} > 0$, $\theta_r < 0$ and $d\theta_r/dr > 0$.[27] Therefore β_{rr} is obtainable from the measured gravitational redshift of the photon at various distances from the center of the earth so that the derivative in equation (93) can be obtained from measured data. Also, from equations (82) and (90) it follows that

$$G_r = G(1 - 2\sin^2\theta_r)(1 - \sin^2\theta_r) \tag{94}$$

$$= G[1 - 2(z_m - z_c)/z_c][1 - (z_m - z_c)/z_c] \tag{95}$$

$$= G(3z_c - 2z_m)(2z_c - z_m)/z_c^2 \tag{96}$$

$$\sim G[1 - 3(z_m - z_c)/z_c] = G(4z_c - 3z_m)/z_c \tag{97}$$

where the small angle approximation in equation (97) can be obtained directly from equations (83) and (91). Combining equations (85) and (91) gives

$$1/G\ \partial G_r/\partial r \sim -6\theta_r\ \partial\theta_r/\partial r \sim -3\partial/\partial r[(z_m - z_c)/z_c] \tag{98}$$

Equations (93), (97) and (98) agree with equation (87). In this way β_{rr}, G_r and dG_r/dr can be determined from photon gravitational redshift measurements conducted at various radial distances from the earth's center.

D. Determination of β_{rr} from the Complex Number Pressure.

Combining the slow variation approximation form of equation (17) with equation (20) gives

$$\beta_{rr} \sim r\partial\theta_r/\partial r \sim -1/2r\partial/\partial r[\theta_p + P(\partial\theta_p/\partial r)/(\partial P/\partial r)] \tag{99}$$

$$= -1/2r\partial/\partial r[\theta_p + P(r\partial\theta_p/\partial r)/(r\partial P/\partial r)]$$

which can be evaluated as in equation (31) by using the relation

$$r\partial/\partial r = \sigma n\partial/\partial n + \mu T\partial/\partial T \tag{100}$$

Measurements of gravity are done in the atmosphere whether above the earth's surface or below in mine shafts and boreholes. Therefore θ_r and θ_p refer to

the atmosphere when gravity measurements are considered. The values of θ_p can be obtained from measurements of the third virial coefficient of real gases.[27] If the third virial coefficient is measured over a range of elevations it may be possible to determine the function $\beta_{rr}(r)$ by using equations (99) and (100). The measurement of the third virial coefficient of the real gases has been suggested as a method of detecting gravity waves.[27] It should be no surprise then that the measurement of the variation of the third virial coefficient with elevation can be used to determine the broken spacetime symmetry that exists within a real gas in a gravity field.

4. EQUATION OF STATE OF REAL GASES IN A GRAVITY FIELD. This section determines the form of the state equation of real gases located in a spacetime that has a broken symmetry (coordinate internal phase) that is induced by gravity or by altering the vacuum in the laboratory such as may be done by the application of a magnetic field. This calculation is done first by the exact method of incorporating broken symmetry space into each term of the pressure state equation, and secondly by the cosine approximation technique of equation (62). The gravitational constant of the earth may possibly be determined from the measurement of the pressure of real gases in a container placed at several elevations in a gravity field.

A. Calculation of β_{VV} for the Real Gases.

In Section 2 it was shown that the calculation of the pressure of matter located in a gravity field requires the angle β_{VV} that describes the degree of the broken symmetry of space. It can be shown that for the real gases the internal phase angle of the pressure is given under the small angle approximation by[27]

$$\theta_p \sim (n^2 C \sin\theta_C)/(1 + Bn + C\cos\theta_C\, n^2 + \cdots) \tag{101}$$
$$= \varepsilon_2 n^2 + \varepsilon_3 n^3 + \varepsilon_4 n^4 + \cdots$$

where C and θ_C = magnitude and internal phase angle of the third virial coefficient of the real gases, n = particle number density, B = second virial coefficient and where $\varepsilon_j = \varepsilon_j(T)$. The third virial coefficient C can be positive or negative depending on the value of the temperature.[27] Combining equations (28) and (101) gives

$$\theta_r \sim -1/2[\theta_p + P(\partial\theta_p/\partial r)/(\partial P/\partial r)] \tag{102}$$
$$= -1/2(\delta_2 n^2 + \delta_3 n^3 + \delta_r n^4 + \cdots)$$

where $\delta_j = \delta_j(T)$. Then the value of β_{VV} is obtained from equations (30) and (102) for small angles and for a slow variation of θ_r with radial distance

$$\beta_{VV} \sim \beta_{rr} \sim -1/2(a_2 n^2 + a_3 n^3 + a_4 n^4 + \cdots) \tag{103}$$

where

$$a_2(T) = 2\sigma\delta_2 + \mu T\partial\delta_2/\partial T \tag{104}$$

$$a_3(T) = 3\sigma\delta_3 + \mu T\partial\delta_3/\partial T \tag{105}$$

$$a_4(T) = 4\sigma\delta_4 + \mu T\partial\delta_4/\partial T \tag{106}$$

where $\sigma < 0$ and $\mu < 0$ as described in Section 1, so that $a_j < 0$ and $\beta_{VV} \sim \beta_{rr} > 0$ for the earth. Then equation (103) gives

$$\cos \beta_{VV} = 1 - 1/2\beta_{VV}^2 + \cdots \tag{107}$$
$$= 1 - \alpha_4 n^4 - \alpha_5 n^5 - \alpha_6 n^6 - \cdots$$

where

$$\alpha_4 = 1/8a_2^2 \tag{108}$$

$$\alpha_5 = 1/4a_2a_3 \tag{109}$$

$$\alpha_6 = 1/8(2a_2a_4 + a_3^2) \tag{110}$$

B. Exact Real Gas State Equation for Broken Symmetry Spacetime.

The equation of state of real gases for symmetrical spacetime is given by[27,28]

$$P^a = nRT(1 + B^a n + C^a n^2 + D^a n^3 + E^a n^4 + \cdots) \tag{111}$$

$$E^a = nRT(3/2 - nT\partial B^a/\partial T - 1/2n^2 T\partial C^a/\partial T - 1/3n^3 T\partial D^a/\partial T \tag{112}$$
$$- 1/4n^4 T\partial E^a/\partial T - \cdots)$$

$$\bar{P} = nRT(1 + Bn + \bar{C}n^2 + \bar{D}n^3 + \bar{E}n^4 + \cdots) \tag{113}$$

$$\bar{E} = nRT(3/2 - nT\partial B/\partial T - 1/2n^2 T\partial\bar{C}/\partial T - 1/3n^3 T\partial\bar{D}/\partial T \tag{114}$$
$$- 1/4n^4 T\partial\bar{E}/\partial T - \cdots)$$

where

$$n = N/V = \text{particle number density} \tag{115}$$

and P^a, E^a = unrenormalized pressure and energy density respectively, and $\bar{P}$, $\bar{E}$ = renormalized pressure and energy density respectively. The connection between B^a, C^a and the corresponding renormalized values B and $\bar{C}$ are given by a solution of the trace equation of relativistic thermodynamics.[27,28] The third and higher renormalized virial coefficients $\bar{C}$, $\bar{D}$, $\bar{E}$, $\cdots$, are generally complex number solutions of the relativistic trace equation for symmetrical spacetime.[27]

For a real gas in a gravity field that induces broken symmetry spacetime the state equations (111) through (115) become

$$P_G^a = n_G RT(1 + B^a n_G + C^a n_G^2 + D^a n_G^3 + E^a n_G^4 + \cdots) \tag{116}$$

$$E_G^a = n_G RT(3/2 - n_G T\partial B^a/\partial T - 1/2n_G^2 T\partial C^a/\partial T - 1/3n_G^3 T\partial D^a/\partial T - 1/4n_G^4 T\partial E^a/\partial T - \cdots) \tag{117}$$

$$\bar{P}_G = n_G RT(1 + Bn_G + \bar{C}n_G^2 + \bar{D}n_G^3 + \bar{E}n_G^4 + \cdots) \tag{118}$$

$$\bar{E}_G = n_G RT(3/2 - n_G T\partial B/\partial T - 1/2n_G^2 T\partial \bar{C}/\partial T - 1/3n_G^3 T\partial \bar{D}/\partial T - 1/4n_G^4 T\partial \bar{E}/\partial T - \cdots) \tag{119}$$

which give the pressure and energy density in a gravity field, where the particle number density in broken symmetry spacetime is given by

$$n_G = N/V_G = n \cos \beta_{VV} \tag{120}$$

where V_G is given by equation (7). Therefore the calculation of P_G^a, E_G^a and $\bar{P}_G$, $\bar{E}_G$ requires the evaluation of β_{VV} for the real gases. It is assumed that β_{VV} is a function of the local temperature and density. Combining equations (116) -(119) and (120) gives

$$P_G^a = nRT \cos \beta_{VV} (1 + B^a n \cos \beta_{VV} + C^a n^2 \cos^2 \beta_{VV} + D^a n^3 \cos^3 \beta_{VV} + E^a n^4 \cos^4 \beta_{VV} + \cdots) \tag{121}$$

$$E_G^a = nRT \cos \beta_{VV} (3/2 - n \cos \beta_{VV} T\partial B^a/\partial T - 1/2n^2 \cos^2 \beta_{VV} T\partial C^a/\partial T - 1/3n^3 \cos^3 \beta_{VV} T\partial D^a/\partial T - 1/4n^4 \cos^4 \beta_{VV} T\partial E^a/\partial T - \cdots) \tag{122}$$

$$\bar{P}_G = nRT \cos \beta_{VV} (1 + Bn \cos \beta_{VV} + \bar{C}n^2 \cos^2 \beta_{VV} + \bar{D}n^3 \cos^3 \beta_{VV} + \bar{E}n^4 \cos^4 \beta_{VV} + \cdots) \tag{123}$$

$$\bar{E}_G = nRT \cos \beta_{VV} (3/2 - n \cos \beta_{VV} T\partial B/\partial T - 1/2n^2 \cos^2 \beta_{VV} T\partial \bar{C}/\partial T - 1/3n^3 \cos^3 \beta_{VV} T\partial \bar{D}/\partial T - 1/4n^4 \cos^4 \beta_{VV} T\partial \bar{E}/\partial T - \cdots) \tag{124}$$

Placing equation (107) into equations (121) through (124) gives

$$P_G^a = nRT[1 + B^a n + C^a n^2 + D^a n^3 + (E^a - \alpha_4)n^4 + (F^a - 2\alpha_4 B^a - \alpha_5)n^5 + \cdots] \quad (125)$$

$$E_G^a = nRT[3/2 - nT\partial B^a/\partial T - 1/2n^2 T\partial C^a/\partial T - 1/3n^3 T\partial D^a/\partial T - 1/4n^4(T\partial E^a/\partial T + 6\alpha_4) - 1/5n^5(T\partial F^a/\partial T - 10\alpha_4 T\partial B^a/\partial T + 15/2\alpha_5) - \cdots] \quad (126)$$

$$\bar{P}_G = nRT[1 + Bn + \bar{C}n^2 + \bar{D}n^3 + (\bar{E} - \alpha_4)n^4 + (\bar{F} - 2\alpha_4 B - \alpha_5)n^5 + \cdots] \quad (127)$$

$$\bar{E}_G = nRT[3/2 - nT\partial B/\partial T - 1/2n^2 T\partial \bar{C}/\partial T - 1/3n^3 T\partial \bar{D}/\partial T - 1/4n^4(T\partial \bar{E}/\partial T + 6\alpha_4) - 1/5n^5(T\partial \bar{F}/\partial T - 10\alpha_4 T\partial B/\partial T + 15/2\alpha_5) - \cdots] \quad (128)$$

The effective fifth and sixth virial coefficients appearing in the pressure equation (127) are

$$\bar{E}_G = \bar{E} - \alpha_4 \quad (129)$$

$$\bar{F}_G = \bar{F} - 2\alpha_4 B - \alpha_5 \quad (130)$$

The effect of broken spacetime symmetry occurs in the fifth and higher virial coefficients $\bar{E}, \bar{F}, \bar{G}, \cdots$, of the real gases, and therefore the effects of broken spacetime symmetry should be observed in the real gases only at high densities. The fifth virial coefficient $\bar{E}$ is lowered in value due to a gravity field or some other means of inducing a broken symmetry in spacetime. For comparison it should be pointed out that the effects of the gauge and conformal invariant trace equation occur in the third and higher virial coefficients $\bar{C}, \bar{D}, \bar{E}, \cdots$, and should be more readily detectable.[27,28]

C. The Cos β_{VV} Approximation.

According to equation (62) the pressure of a real gas in a gravity field, or more generally in broken symmetry spacetime, can be obtained from the symmetrical spacetime pressure given in equations (111) and (113) by simply multiplying by cos β_{VV} and assuming the energy densities are unchanged as follows

$$P_G^{ac} = nRT \cos \beta_{VV} (1 + B^a n + C^a n^2 + D^a n^3 + E^a n^4 + F^a n^5 + \cdots) \quad (131)$$

$$E_G^a \sim nRT(3/2 - nT\partial B^a/\partial T - 1/2n^2 T\partial C^a/\partial T - 1/3n^3 T\partial D^a/\partial T - 1/4n^4 T\partial E^a/\partial T - 1/5n^5 T\partial F^a/\partial T - \cdots) \quad (132)$$

$$\bar{P}_G^c = nRT \cos \beta_{VV} \, (1 + Bn + \bar{C}n^2 + \bar{D}n^3 + \bar{E}n^4 + \bar{F}n^5 + \cdots) \tag{133}$$

$$\bar{E}_G \sim nRT(3/2 - nT\partial B/\partial T - 1/2n^2 T\partial \bar{C}/\partial T - 1/3n^3 T\partial \bar{D}/\partial T \tag{134}$$
$$- 1/4n^4 T\partial \bar{E}/\partial T - 1/5n^5 T\partial \bar{F}/\partial T - \cdots)$$

where the superscript c is used to designate the cosine approximation value for the thermodynamic pressure. Note that $\bar{E}_G^a \sim E^a$ and $\bar{E}_G \sim \bar{E}$ in this procedure. Expanding the cosine terms in the form of equation (107) gives

$$P_G^{ac} = nRT[1 + B^a n + C^a n^2 + D^a n^3 + (E^a - \alpha_4)n^4 \tag{135}$$
$$+ (F^a - \alpha_4 B^a - \alpha_5)n^5 + \cdots]$$

$$E_G^a \sim nRT(3/2 - nT\partial B^a/\partial T - 1/2n^2 T\partial C^a/\partial T - 1/3n^3 T\partial D^a/\partial T \tag{136}$$
$$- 1/4n^4 T\partial E^a/\partial T - 1/5n^5 T\partial F^a/\partial T - \cdots)$$

$$\bar{P}_G^c = nRT[1 + Bn + \bar{C}n^2 + \bar{D}n^3 + (\bar{E} - \alpha_4)n^4 \tag{137}$$
$$+ (\bar{F} - \alpha_4 B - \alpha_5)n^5 + \cdots]$$

$$\bar{E}_G \sim nRT(3/2 - nT\partial B/\partial T - 1/2n^2 T\partial \bar{C}/\partial T - 1/3n^3 T\partial \bar{D}/\partial T \tag{138}$$
$$- 1/4n^4 T\partial \bar{E}/\partial T - 1/5n^5 T\partial \bar{F}/\partial T - \cdots)$$

Therefore for the cosine approximation rule for calculating the pressure of a real gas in a gravity field the state equations for the pressure are affected only in the fifth and higher virial coefficients in such a way that within this approximation the fifth and sixth virial coefficients are obtained from equation (137) to be

$$\bar{E}_G^c = \bar{E} - \alpha_4 \tag{139}$$

$$\bar{F}_G^c = \bar{F} - \alpha_4 B - \alpha_5 \tag{140}$$

The cosine approsimation pressure equations (131) and (133) agree out to the fifth order virial coefficient with the pressure values obtained from the exact equation (121) and (123) as can be seen by comparing equations (129) and (139). However, differences between the exact and approximate calculations appear in the sixth virial coefficient as can be seen by comparing equations (130) and (140). Therefore the real gases $P_G^{ac} \sim P_G^a$ and $\bar{P}_G^c \sim \bar{P}_G$ at least out to the fifth virial coefficient, and so the cosine rule for calculating the pressure of a real gas in a gravity field is a reasonable approximation.

5. STATE EQUATIONS OF SOLIDS AND QUANTUM LIQUIDS IN A GRAVITY FIELD.

This section considers the effects of gravity on the thermodynamic functions of solids and quantum liquids, and suggests that the measurement of the Grüneisen function can be used to determine the elevation dependent gravitational constant $G_r(r)$. The solids and quantum liquids are assumed to have a simple Mie-Grüneisen type of state equation wherein the pressure is given for symmetric spacetime (absence of gravity field) by[27,28]

$$\bar{P} = \bar{P}_o + \bar{\gamma}_o \bar{E}_T = \bar{P}_o + \bar{\gamma}_o \bar{E}_\nu T^\nu = \bar{P}_o + \bar{P}_\nu T^\nu \tag{141}$$

$$P^a = P^a_o + \gamma^a_o E^a_T = P^a_o + \gamma^a_o E^a_\nu T^\nu = P^a_o + P^a_\nu T^\nu \tag{142}$$

and the energy densities and internal energies are

$$\bar{E} = \bar{E}_o + \bar{E}_T = \bar{E}_o + \bar{E}_\nu T^\nu \qquad \bar{U} = \bar{U}_o + \bar{U}_T = \bar{U}_o + \bar{U}_\nu T^\nu \tag{143}$$

$$E^a = E^a_o + E^a_T = E^a_o + E^a_\nu T^\nu \qquad U^a = U^a_o + U^a_T = U^a_o + U^a_\nu T^\nu \tag{144}$$

where the thermal energy densities are written as

$$\bar{E}_T = E_T e^{j\theta E T} = \bar{E}_\nu T^\nu = E_\nu T^\nu e^{j\theta E \nu} \qquad \theta_{ET} = \theta_{E\nu} \tag{145A}$$

$$E^a_T = E^a_\nu T^\nu$$

where ν = number that depends on type of system being considered. The complex number energy densities and pressures are written as

$$\bar{E} = E e^{j\theta E} \qquad \bar{E}_o = E_o e^{j\theta^o_E} \qquad \bar{E}_T = E_T e^{j\theta E T} \qquad \bar{E}_\nu = E_\nu e^{j\theta E \nu} \tag{146}$$

$$\bar{P} = P e^{j\theta P} \qquad \bar{P}_o = P_o e^{j\theta^o_P} \qquad \bar{P}_\nu = P_\nu e^{j\theta P \nu} \tag{147}$$

while the complex number zero-temperature value of the Grüneisen function is written as

$$\bar{\gamma}_o = \gamma_o e^{j\theta^o_\gamma} \tag{148}$$

where $\bar{P}$, $\bar{P}_o$, $\bar{P}_\nu$, $\bar{\gamma}_o$, $\bar{E}_T$ and $\bar{E}_\nu$ = renormalized values of the symmetric spacetime pressure, zero temperature pressure, thermal pressure coefficient, zero temperature Grüneisen function, thermal energy density, and thermal energy density coefficient respectively, and where P^a, P^a_o, P^a_ν, γ^a_o, E^a_T and E^a_ν = corresponding unrenormalized values of these quantities. The relation between the renormalized and unrenormalized thermodynamic functions is given by a relativistic trace equation.[28]

The thermodynamic quantities that appear in equations (141) through (148) are functions of n and T, where n = N/V = average particle number density for symmetric spacetime. In a gravity field, or some laboratory created field such as a magnetic field, the spacetime symmetry is broken and the particle number density n_G is given by equation (120), and for this case the renormalized and unrenormalized state equations for solids and quantum liquids are written as

$$\bar{P}_G = \bar{P}_{Go} + \bar{\gamma}_{Go}\bar{E}_{GT} = \bar{P}_{Go} + \bar{\gamma}_{Go}\bar{E}_{G\nu}T^{\nu} = \bar{P}_{Go} + \bar{P}_{G\nu}T^{\nu} \tag{149}$$

$$P^a_G = P^a_{Go} + \gamma^a_{Go}E^a_{GT} = P^a_{Go} + \gamma^a_{Go}E^a_{G\nu}T^{\nu} = P^a_{Go} + P^a_{G\nu}T^{\nu} \tag{150}$$

$$\bar{U}_G = \bar{U}_{Go} + \bar{U}_{GT} = \bar{U}_{Go} + \bar{U}_{G\nu}T^{\nu} \tag{150A}$$

$$U^a_G = U^a_{Go} + U^a_{GT} = U^a_{Go} + U^a_{G\nu}T^{\nu} \tag{150B}$$

$$\bar{E}_G = \bar{E}_{Go} + \bar{E}_{GT} = \bar{E}_{Go} + \bar{E}_{G\nu}T^{\nu} \tag{150C}$$

$$E^a_G = E^a_{Go} + E^a_{GT} = E^a_{Go} + E^a_{G\nu}T^{\nu} \tag{150D}$$

where the thermodynamic quantities of equations (149) through (150D) are functions of n_G and T and can be obtained from the corresponding thermodynamic quantities in the symmetric spacetime equations (141) and (142) by making the substitution $n \rightarrow n_G = n \cos \beta_{VV}$. Therefore the angle β_{VV} must be calculated for the solid and quantum liquid type of state equations given in equations (149) through (150D). The relativistic trace equation that connects the renormalized and unrenormalized thermodynamic quantities of equations (149) through (150D) is just the standard form for solids and quantum liquids (reference 28) with the substitution $n \rightarrow n_G$. It should be pointed out that the cosine approximation, equation (62), for calculating the pressure of a system in the presence of a gravity field can be applied to equations (141) and (142) with the result that

$$\bar{P}^c_G = \bar{P}^c_{Go} + \bar{\gamma}^c_{Go}\bar{E}^c_{GT} = \cos \beta_{VV} \, (\bar{P}_o + \bar{\gamma}_o\bar{E}_T) \tag{151}$$

$$P^{ac}_G = P^{ac}_{Go} + \gamma^{ac}_{Go}E^{ac}_{GT} = \cos \beta_{VV} \, (P^a_o + \gamma^a_o E^a_T) \tag{152}$$

where the superscript c designates quantities calculated within the cosine approximation. If the cosine approximation has validity for solids and quantum liquids, then $\bar{P}^c_G \sim \bar{P}_G$ and $\bar{P}^{ac}_G \sim P^a_G$ where $\bar{P}_G$ and P^a_G are the exact pressures for solids and quantum liquids in a gravity field which are obtained from the symmetric spacetime pressures $\bar{P}$ and P^a by making the substitution $n \rightarrow n_G = n \cos \beta_{VV}$ in each term of the symmetric spacetime form of the state equations (141) and (142).

The value β_{VV} can be calculated from equations (28) and (30) provided that the internal phase angle of the pressure θ_P is known for solids and quantum liquids. The phase angle θ_P can be calculated from the real and imaginary parts of equation (141) which can be written as

$$P \cos \theta_P = P_o \cos \theta^o_P + \gamma_o E_T \cos(\theta^o_\gamma + \theta_{E\nu}) \tag{153}$$

$$P \sin \theta_P = P_o \sin \theta^o_P + \gamma_o E_T \sin(\theta^o_\gamma + \theta_{E\nu}) \tag{154}$$

The magnitude and internal phase angle of the pressure can be obtained from equations (153) and (154). The magnitude of the pressure is given by

$$P^2 = P_o^2 + \gamma_o^2 E_T^2 + 2\gamma_o P_o E_T \cos(\theta_\gamma^o + \theta_{E\nu} - \theta_P^o) \tag{155}$$

or approximately as

$$P \sim P_o + \gamma_o E_T = P_o + P_\nu T^\nu \tag{156}$$

where

$$P_\nu = \gamma_o E_\nu \qquad \gamma_o = P_\nu / E_\nu \tag{157}$$

The internal phase angle of the pressure for solids and quantum liquids is obtained from equations (153) and (154) to be

$$\tan \theta_P = A/B \tag{158}$$

where

$$A = P_o \sin \theta_P^o + \gamma_o E_T \sin(\theta_\gamma^o + \theta_{E\nu}) \tag{158A}$$

$$B = P_o \cos \theta_P^o + \gamma_o E_T \cos(\theta_\gamma^o + \theta_{E\nu}) \tag{158B}$$

For small internal phase angles, the internal phase angle of pressure is given by

$$\theta_P = [\theta_P^o + (\gamma_o E_T / P_o)(\theta_\gamma^o + \theta_{E\nu})]/(1 + \gamma_o E_T / P_o) \tag{159}$$

If furthermore $\gamma_o E_T << P_o$ then

$$\theta_P = \theta_P^o + \theta_P^T = \theta_P^o + fE_T = \theta_P^o + fE_\nu T^\nu \tag{160}$$

where

$$f(n) = (\gamma_o / P_o)(\theta_\gamma^o + \theta_{E\nu} - \theta_P^o) \tag{161}$$

The value of the angle β_{VV} can be obtained from equations (28), (30) and (160) to be of the general form

$$\beta_{VV} = \beta_{VV}^o + b_\nu T^\nu \tag{162}$$

where b_ν is a function of density $b_\nu = b_\nu(n)$. From equation (162) it follows that

$$\cos \beta_{VV} = \cos(\beta_{VV}^o + b_\nu T^\nu) \tag{163}$$

$$\sim \cos \beta_{VV}^o - b_\nu T^\nu \sin \beta_{VV}^o$$

where for the approximation in equation (163) it is assumed that $b_\nu T^\nu << \beta_{VV}^o$.

Combining equations (51), (149), (150A) and (163) gives the following approximation equations for the pressure of solids and quantum liquids in a gravity field

$$\bar{P}_{Go} \sim - d\bar{U}_{Go}/dV \cos \beta^{o}_{VV} \tag{164}$$

$$\sim \bar{P}_{o} \cos \beta^{o}_{VV} \tag{165}$$

$$\bar{P}_{G\nu} \sim (\nu - 1)^{-1}(d\bar{U}_{G\nu}/dV \cos \beta^{o}_{VV} - b_{\nu} d\bar{U}_{Go}/dV \sin \beta^{o}_{VV}) \tag{166}$$

$$= (\nu - 1)^{-1}(d\bar{U}_{G\nu}/dV \cos \beta^{o}_{VV} + b_{\nu} \bar{P}_{Go} \tan \beta^{o}_{VV}) \tag{167}$$

$$\sim (\nu - 1)^{-1}(d\bar{U}_{G\nu}/dV + b_{\nu} \bar{P}_{Go} \beta^{o}_{VV}) \tag{168}$$

$$\sim \bar{P}_{\nu} \cos \beta^{o}_{VV} + (\nu - 1)^{-1} b_{\nu} \bar{P}_{o} \sin \beta^{o}_{VV} \tag{169}$$

$$\sim \bar{P}_{\nu} + (\nu - 1)^{-1} b_{\nu} \bar{P}_{o} \beta^{o}_{VV} \tag{170}$$

$$\sim (\nu - 1)^{-1} d\bar{U}_{G\nu}/dV \cos \beta^{o}_{VV} \qquad b_{\nu} \sim 0 \tag{171}$$

$$\sim \bar{P}_{\nu} \cos \beta^{o}_{VV} \qquad b_{\nu} \sim 0 \tag{172}$$

where the approximation in equation (165) assumes that $\bar{U}_{Go} \sim \bar{U}_{o}$, the approximation in equation (168) is valid for small values of the angle β^{o}_{VV}, the approximation in equation (169) holds for $\bar{U}_{G\nu} \sim \bar{U}_{\nu}$, the approximation in equation (170) is appropriate for small values of β^{o}_{VV} and for $\bar{U}_{G\nu} \sim \bar{U}_{\nu}$, the approximation in equation (171) is valid for small b_{ν}, and finally the approximation in equation (172) is valid for small b_{ν} and for $\bar{U}_{G\nu} \sim \bar{U}_{\nu}$. In a gravity field $\bar{P}_{Go} < \bar{P}_{o}$ but $\bar{P}_{G\nu} > \bar{P}_{\nu}$ in general, but if $b_{\nu} = 0$ then $\bar{P}_{G\nu} < \bar{P}_{\nu}$ as shown by equation (172).

From the definition of the zero-temperature Grüneisen function and using equations (166) through (172) it follows that

$$\bar{\gamma}_{Go} = \bar{P}_{G\nu}/\bar{E}_{G\nu} \tag{173}$$

$$\sim (\nu - 1)^{-1}(V/\bar{U}_{G\nu}\, d\bar{U}_{G\nu}/dV \cos \beta^{o}_{VV} - b_{\nu} V/\bar{U}_{G\nu}\, d\bar{U}_{Go}/dV \sin \beta^{o}_{VV}) \tag{174}$$

$$\sim (\nu - 1)^{-1}(V/\bar{U}_{G\nu}\, d\bar{U}_{G\nu}/dV \cos \beta^{o}_{VV} + b_{\nu} \bar{P}_{Go}/\bar{E}_{G\nu} \tan \beta^{o}_{VV}) \tag{175}$$

$$\sim (\nu - 1)^{-1}(V/\bar{U}_{G\nu}\, d\bar{U}_{G\nu}/dV + b_{\nu} \bar{P}_{Go} \beta^{o}_{VV}/\bar{E}_{G\nu}) \tag{176}$$

$$\sim \bar{\gamma}_{o} \cos \beta^{o}_{VV} + (\nu - 1)^{-1} b_{\nu} \bar{P}_{o}/\bar{E}_{\nu} \sin \beta^{o}_{VV} \tag{177}$$

$$\sim \bar{\gamma}_{o} + (\nu - 1)^{-1} b_{\nu} \bar{P}_{o} \beta^{o}_{VV}/\bar{E}_{\nu} \tag{178}$$

$$\sim (\nu - 1)^{-1} V/\bar{U}_{G\nu}\, d\bar{U}_{G\nu}/dV \cos \beta^{o}_{VV} \qquad b_{\nu} \sim 0 \tag{179}$$

$$\sim \bar{\gamma}_{o} \cos \beta^{o}_{VV} \qquad b_{\nu} \sim 0 \tag{180}$$

where the approximation in equation (176) is valid for small values of β^{o}_{VV}, equation (177) holds for $\bar{U}_{G\nu} \sim \bar{U}_{\nu}$, equation (178) assumes that β^{o}_{VV} is small and $\bar{U}_{G\nu} \sim \bar{U}_{\nu}$, the approximation in equation (179) is valid for small b_{ν} (temperature independence of β_{VV}), while equation (180) is valid for small b_{ν} and for $\bar{U}_{G\nu} \sim \bar{U}_{\nu}$. In a gravitational field $\bar{\gamma}_{Go}$ is generally larger than the zero field value $\bar{\gamma}_{o}$ as shown by equation (178), however when $b_{\nu} = 0$ it follows from equation (180) that $\bar{\gamma}_{Go} < \bar{\gamma}_{o}$. Using the approximation in equation (179) with $b_{\nu} \sim 0$ gives

$$\bar{U}_{G\nu} \sim A \exp[(\nu - 1)\int \bar{\gamma}_{Go} \sec \beta^{o}_{VV} \, dV/V] \tag{181}$$

$$= A \exp[-(\nu - 1)\int \bar{\gamma}_{Go} \sec \beta^{o}_{VV} \, dn/n]$$

For a symmetrical spacetime in the absence of a gravity field[27,28]

$$\bar{U}_{\nu} \sim A \exp[-(\nu - 1)\int \bar{\gamma}_{o} \, dn/n] \tag{182}$$

The components of the pressure given in equations (164) and (166) and the zero temperature value of the Grüneisen parameter given in equation (174) are derived from the exact thermodynamic equation (54) by making the substitutions $n \to n_G = n \cos \beta_{VV}$ and $dV \to dV_G = \sec \beta_{VV} \, dV$ which results in equation (51). Therefore aside from the approximation $b_{\nu}T^{\nu} << \beta^{o}_{VV}$ that is used in equation (163), equations (164), (166) and (174) are exact equations for the pressure components and Grüneisen parameter for a solid or quantum liquid in a gravity field.

Now the cosine approximation is used to calculate the components of the pressure and the zero temperature Grüneisen function for solids and quantum liquids. Combining equations (145A), (149), (151) and (163) gives the following expressions for the pressure components within the cosine approximation

$$\bar{P}^{c}_{Go} = \bar{P}_{o} \cos \beta^{o}_{VV} \tag{183A}$$

$$\bar{P}^{c}_{G\nu} = \bar{P}_{\nu} \cos \beta^{o}_{VV} - b_{\nu}\bar{P}_{o} \sin \beta^{o}_{VV} \tag{183B}$$

$$\sim \bar{P}_{\nu} - b_{\nu}\bar{P}_{o}\beta^{o}_{VV} \tag{183C}$$

$$\sim \bar{P}_{\nu} \cos \beta^{o}_{VV} \tag{183D}$$

The following are three values for the zero temperature Grüneisen function corresponding respectively to the case of symmetric spacetime (no gravity), the case of a gravity field, and the cosine approximation for the thermodynamic functions in a gravity field

$$\bar{\gamma}_{o} = \bar{P}_{\nu}/\bar{E}_{\nu} \qquad \bar{\gamma}_{Go} = \bar{P}_{G\nu}/\bar{E}_{G\nu} \qquad \bar{\gamma}^{c}_{Go} = \bar{P}^{c}_{G\nu}/\bar{E}^{c}_{G\nu} \tag{183E}$$

Combining equations (183B) and (183E) gives

$$\bar{\gamma}^{c}_{Go}\bar{E}^{c}_{G\nu} = \bar{\gamma}_{o}\bar{E}_{\nu} \cos \beta^{o}_{VV} - b_{\nu}\bar{P}_{o} \sin \beta^{o}_{VV} \tag{184}$$

If it is assumed that $\bar{E}^{c}_{G\nu} \sim \bar{E}_{G\nu} \sim \bar{E}_{\nu}$ then equation (184) becomes

$$\bar{\gamma}^{c}_{Go} \sim \bar{\gamma}_o \cos \beta^{o}_{VV} - b_\nu \bar{P}_o / \bar{E}_\nu \sin \beta^{o}_{VV} \tag{185}$$

$$\sim \bar{\gamma}_o - b_\nu \bar{P}_o \beta^{o}_{VV} / \bar{E}_\nu \tag{186}$$

$$\sim \bar{\gamma}_o \cos \beta^{o}_{VV} \tag{187}$$

where the approximations in equations (183C) and (186) are valid for small values of β^{o}_{VV}, while the approximations in equations (183D) and (187) are valid for $b_\nu = 0$ which is the condition for the temperature independence of β_{VV}. A comparison of equations (165) and (183A) shows that the correct T = 0 value of the pressure is reproduced by the cosine approximation, but a comparison of equations (169) and (183B) shows that the temperature dependent pressure term is not given correctly by the cosine approximation except when $b_\nu = 0$ in which case equation (172) agrees with (183D) and equation (180) agrees with (187). Therefore only when β_{VV} is temperature independent does the cosine approximation give accurate values of the pressure and zero temperature Grüneisen parameter for solids and quantum liquids.

6. DEBYE THEORY OF SOLIDS IN A GRAVITATIONAL FIELD. This section examines the effects of gravity on the Debye theory of the thermal state equation of solids, and suggests that these effects can be used to determine the gravitational constant $G_r(r)$. The Debye theory of the thermal state of a solid is based on a calculation of the normal modes of the longitudinal and transverse vibrations of a solid combined with the quantum theory expression for the average energy per normal mode.[30-34] This procedure is described in detail in the literature and will be briefly reviewed in this paragraph by writing the standard Debye equations in a complex number form which are needed to describe the effects of gravity on the internal energy and heat capacity of a solid. The number of phonon normal modes in a complex number frequency interval $d\bar{\nu}$ is given by[30-34]

$$d\bar{N} = \bar{A}\bar{\nu}^2 d\bar{\nu} \qquad \bar{A} = 9N/\bar{\nu}^3_{MG} \tag{188}$$

where $\bar{\nu}$ = complex number frequencies of the normal modes which are represented by[35]

$$\bar{\nu} = \nu e^{j\theta_\nu} \qquad \theta_\nu = -\theta_{tR} \tag{190}$$

where θ_{tR} = internal phase angle of the periods of vibration, $\bar{\nu}_{MG}$ = complex number Debye frequency of a solid in a gravitational field which is given by the following generalization of the standard result[30-34]

$$\bar{\nu}_{MG} = \nu_{MG} e^{j\theta^{G}_{\nu M}} = \bar{c}_G [(3N)/(4\pi\bar{V})]^{1/3} = \bar{\nu}^{G}_{M} \tag{191}$$

where

$$\theta^{G}_{\nu M} = -\theta_{tR} \tag{191A}$$

and where N = number of atoms in a solid, $\bar{V}$ = complex number volume of a solid in a gravity field which is represented by equation (3), $\bar{c}_G$ = complex number average wave speed in a solid in the presence of a gravitational field and is given by[30-34]

$$3/\bar{c}_G^2 = 1/\bar{c}_{\ell G}^2 + 2/\bar{c}_{tG}^2 \tag{192}$$

where $\bar{c}_{\ell G}$ = complex number longitudinal wave speed for a solid in a gravitational field, and $\bar{c}_{tG}$ = complex number transverse (shear) wave speed for a solid located in a gravitational field. The complex number average energy per normal mode is given by the following generalization of the standard result[27,30-35]

$$\bar{\varepsilon} = h\bar{\nu}/2 + h\bar{\nu}/[\exp(h\bar{\nu}/kT) - 1] \tag{193}$$

which includes the zero point energy corresponding to $T = 0$.

The internal energy of a Debye solid that is located in a gravity field is then given by equations (188) and (193) as

$$\bar{U}_G = \bar{A}h/2 \int_0^{\bar{\nu}_{MG}} \bar{\nu}^3 d\bar{\nu} + \bar{A}h \int_0^{\bar{\nu}_{MG}} \bar{\nu}^3 [\exp(h\bar{\nu}/kT) - 1]^{-1} \, d\bar{\nu} \tag{194}$$

$$= \bar{U}_o^G + \bar{U}_T^G$$

where $\bar{U}_o^G$ and $\bar{U}_T^G$ are simplified to

$$\bar{U}_o^G = U_o^G \exp(j\theta_{Uo}^G) = 9/8Nh\bar{\nu}_M^G \tag{195}$$

$$\bar{U}_T^G = U_T^G \exp(j\theta_{UT}^G) = 9NkT(\bar{x}_D^G)^{-3} \int_0^{\bar{x}_D^G} \bar{x}^3 [\exp(\bar{x}) - 1]^{-1} \, d\bar{x} \tag{196}$$

where $\bar{x}_D^G = \bar{T}_D^G/T$ and

$$\bar{x} = h\bar{\nu}/(kT) \qquad x = h\nu/(kT) \qquad \theta_x = \theta_\nu = -\theta_{tR} \tag{197}$$

where the complex number Debye temperature $\bar{T}_D^G$ for a solid in a gravitational field is given by

$$k\bar{T}_D^G = h\bar{\nu}_M^G \qquad \bar{T}_D^G = T_D^G \exp(j\theta_{TD}^G) \tag{198}$$

or

$$kT_D^G = h\nu_M^G \qquad \theta_{TD}^G = \theta_{\nu M}^G = -\theta_{tR} \tag{199}$$

Equation (195) gives

$$U_o^G = 9/8Nh\nu_M^G \qquad \theta_{Uo}^G = \theta_{\nu M}^G = -\theta_{tR} \tag{200}$$

and the measured zero point energy is

$$U_{om}^G = U_o^G \cos\theta_{Uo}^G = 9/8Nh\nu_M^G \cos\theta_{\nu M}^G = 9/8Nh\nu_{Mm}^G \tag{201}$$

where ν_{Mm}^G = measured Debye frequency given by

$$\nu^G_{Mm} = \nu^G_M \cos\theta^G_{\nu M} = (k/h)T^G_D \cos\theta^G_{TD} = (k/h)T^G_{Dm} \tag{202}$$

where T^G_{Dm} = measured Debye temperature for a solid in a gravitational field. Equations (200) and (201) are essentially the standard result that the zero point energy is linearly dependent on the Debye frequency.

The complex number Debye function for a solid in a gravitational field is defined as the following generalization of the standard form[30-34]

$$\bar{D}_G(\bar{T}^G_D/T) = 3(\bar{x}^G_D)^{-3} \int_o^{\bar{x}^G_D} \bar{x}^3[\exp(\bar{x}) - 1]^{-1}d\bar{x} \tag{203}$$

so that equation (196) can be written as

$$\bar{U}^G_T = 3NkT\bar{D}_G(\bar{T}^G_D/T) \tag{204}$$

Equation (203) can be rewritten as

$$\bar{D}_G(\bar{T}^G_D/T) = 3(\bar{x}^G_D)^{-3} \int_o^{\bar{x}^G_D} \sec\beta_{xx}\, x^3 \exp[j(4\theta_x + \beta_{xx})]\bar{F}dx \tag{205}$$

where

$$\bar{F} = \{[\cos(x\sin\theta_x) + j\sin(x\sin\theta_x)]\exp(x\cos\theta_x) - 1\}^{-1} \tag{206}$$

$$\tan\beta_{xx} = x\partial\theta_x/dx \tag{207}$$

For the case when $\theta_x = \theta^c_x$ = constant equation (205) becomes

$$\bar{D}_G(\bar{T}^G_D/T) = 3(\bar{x}^G_D)^{-3} \exp(j4\theta^c_x) \int_o^{\bar{x}^G_D} x^3\bar{F}_c dx \tag{208}$$

where

$$\bar{F}_c = \{[\cos(x\sin\theta^c_x) + j\sin(x\sin\theta^c_x)]\exp(x\cos\theta^c_x) - 1\}^{-1} \tag{209}$$

At low temperatures the complex number Debye function that appears in equation (203) can be simplified by writing[30-34]

$$(e^{\bar{x}} - 1)^{-1} = e^{-\bar{x}}(1 - e^{-\bar{x}})^{-1} = \sum_{n=1}^{\infty} e^{-n\bar{x}} \tag{210}$$

so that

$$\bar{D}_G(\bar{T}^G_D/T) = 3(\bar{x}^G_D)^{-3} \sum_{n=1}^{\infty} \int_o^{\infty} \bar{x}^3 e^{-n\bar{x}}d\bar{x} \tag{211}$$

$$= 3(\bar{x}^G_D)^{-3} \sum_{n=1}^{\infty} \int_o^{\infty} \sec\beta_{xx}\, x^3 \exp[j(4\theta_x + \beta_{xx})]\, \bar{H}_n dx$$

where

$$\bar{H}_n = e^{-nx \cos \theta_x}[\cos(nx \sin \theta_x) - j \sin(nx \sin \theta_x)] \tag{212}$$

For the case of constant $\theta_x = \theta_x^c$, or equivalently $\theta_\nu = \theta_\nu^c$, the integral in equation (211) can be written as

$$\bar{D}_G(\bar{T}_D^G/T) = 3(T/\bar{T}_D^G)^3 \sum_{n=1}^{\infty} [\exp(j4\theta_x^c) \int_o^{\infty} x^3 e^{-n\bar{x}} dx] \tag{213}$$

It has been shown in Reference 35 that the quantity in the square brackets of equation (213) is a real number given by $6/n^4$ so that this equation can be written as

$$\bar{D}_G(\bar{T}_D^G/T) = 3(T/\bar{T}_D^G)^3 \sum_{n=1}^{\infty} 6/n^2 = (3/15)\pi^4(T/\bar{T}_D^G)^3 \tag{214}$$

Combining equations (204) and (214) gives the following low temperature form for the internal energy of a solid in a gravitational field

$$\bar{U}_T^G = (9/15)\pi^4 RT(T/\bar{T}_D^G)^3 \tag{215}$$

or in terms of the magnitude and the internal phase angle

$$U_T^G = (9/15)\pi^4 RT(T/T_D^G)^3 \tag{216}$$

$$\theta_{UT}^G = -3\theta_{TD}^G = -3\theta_{\nu M}^G = 3\theta_{tR} \tag{217}$$

The complex number heat capacity for a low temperature solid in a gravitational field is obtained from equation (215) to be

$$\bar{C}_T^G = 12/5\ \pi^4 R(T/\bar{T}_D^G)^3 \tag{218}$$

The measured value of the lattice vibration (phonon) energy of a low temperature Debye solid in a gravitational field is then given by

$$\begin{aligned} U_m^G &= U_o^G \cos \theta_{Uo}^G + U_T^G \cos \theta_{UT}^G \\ &= 9/8Nh\nu_M^G \cos \theta_{tR} + 9/15\pi^4 RT(T/T_D^G)^3 \cos(3\theta_{tR}) \\ &= 9/8Nh\nu_{Mm}^G + 9/15\pi^4 RT(T/T_{Dm}^G)^3 \cos^3\theta_{tR} \cos(3\theta_{tR}) \\ &= 9/8Nh\nu_{Mm}^G + 9/15\pi^4 RT(T/T_{Dm}^G)^3 \cos^3\theta_r \cos(3\theta_r) \end{aligned} \tag{219}$$

where ν_{Mm}^G and T_{Dm}^G are the measured values of the Debye frequency and Debye temperature respectively for a solid in a gravitational field and are defined in equation (202). The measured lattice heat capacity of a solid in gravity field is obtained from equation (218) to be

$$C^{G}_{Tm} = 12/5\pi^4 R(T/T^{G}_{D})^3 \cos(3\theta_{tR}) \qquad (220)$$

$$= 12/5\pi^4 R(T/T^{G}_{Dm})^3 \cos^3\theta_{tR} \cos(3\theta_{tR})$$

$$= 12/5\pi^4 R(T/T^{G}_{Dm})^3 \cos^3\theta_{r} \cos(3\theta_{r})$$

Equations (219) and (220) can also be written in terms of θ_r by remembering that the redshift of a phonon in a gravity field is related to the change in gravitational potential energy which gives immediately $\theta_{tR} = \theta_r$. This is also the phase angle condition for photons in a gravitational field. Equations (219) and (220) show that the internal energy and heat capacity for a solid in a gravitational field are reduced compared to their free space values. At the earth's surface the effect is small because $\theta_r = -5.7°$ so that $\cos^3\theta_r \cos(3\theta_r) = 0.94$ and this represents only a 6% reduction in the measured values of the internal energy and heat capacity at the earth's surface as compared to the corresponding measured values in a gravity free area at a large distance from the earth. In a large gravitational field as found in neutron stars and white dwarf stars the value of θ_r may be large and of the order $\theta_r \sim \pi/6$ so that the measured heat capacity and internal energy of vibrations of a solid can be zero in value. These conditions may also hold in the normal state of a high-T_c superconductor because the superconducting state is associated with $\theta_r = \pi/3$ and $\theta_t = \pi/6$ for the electrons in a Cooper pair. Because the factor $\cos^3\theta_r \cos(3\theta_r)$ appears in the expression for the measured heat capacity in equation (220) it follows that the measurement of the heat capacity of a solid at various radial distance from the earth's center can determine $\theta_r(r)$, and this will give values of the radial coordinate dependence of the gravitational constant $G_r(r)$ from equation (82).

7. CONCLUSION. The thermodynamic state equations of matter are affected by gravitational fields through an induced broken symmetry of the spacetime in which the matter is located. The effects are small at the earth's surface but may have appreciable effects in compact stellar objects. Applications to the real gases show that the effects of gravity occur in the fifth and higher virial coefficients, so that the effects are small except at very high densities. For solids the effects of gravity on the lattice phonon component of the internal energy and heat capacity is about 6% at the surface of the earth, and arises through the internal phase angle of the Debye temperature. Thermodynamic measurements may possibly be used to determine the broken symmetry of spacetime and the values of the gravitational constant that depends on the radial distance from the center of the earth.

ACKNOWLEDGEMENT

Many thanks go to Elizabeth Klein for typing this paper.

REFERENCES

1. Chandrasekhar, S., An Introduction to the Study of Stellar Structure, Dover, New York, 1939.

2. Eddington, A. S., The Internal Constitution of the Stars, Dover, New York, 1926.

3. Schwarzschild, M., Structure and Evolution of the Stars, Dover, New York, 1958.

4. Unsöld, A., The New Cosmos, Springer-Verlag, New York, 1969.

5. Aller, L. H., Astrophysics - Nuclear Transformations, Stellar Interiors, and Nebulae, Ronald Press, New York, 1954.

6. Kuiper, G. P., editor, The Sun, Univ. of Chicago Press, Chicago, 1953.

7. Kuiper, G. P., editor, The Earth as a Planet, Univ. of Chicago Press, Chicago, 1954.

8. Jeffreys, H., The Earth, Cambridge University Press, New York, 1962.

9. Stacey, F. D., Physics of the Earth, John Wiley, New York, 1977.

10. Weinberg, S., Gravitation and Cosmology, John Wiley, New York, 1972.

11. Saakyan, G. S., Equilibrium Configurations of Degenerate Gaseous Masses, John Wiley, New York, 1974.

12. Misner, C. W., Thorne, K. S. and Wheeler, J. A., Gravitation, W. H. Freeman, San Francisco, 1973.

13. Stacey, F. D. and Tuck, G. J., "Geophysical Evidence for Non-Newtonian Gravity," Nature, Vol. 292, p. 230-232, 1981.

14. Stubbs, C. W., Adelberger, E. G., Heckel, B. R., Rogers, W. F., Swanson, H. E., Watanabe, R., Gundlach, J. H. and Raab, F. J., "Limits on Composition-Dependent Interactions Using a Laboratory Source: Is There a "Fifth Force" Coupled to Isospin," Phys. Rev. Lett., Vol. 62, p. 609, 6 Feb. 1989.

15. Kuroda, K. and Mio, N., "Test of a Composition-Dependent Force by a Free-Fall Interferometer," Phys. Rev. Lett., Vol 62, p. 1941, 24 Apr. 1989.

16. Bizzeti, P. G., Bizzeti-Sona, A. M., Fazzini, T., Perego, A. and Taccetti, N., "Search for a Composition-Dependent Fifth Force," Phys. Rev. Lett., Vol. 62, p. 2901, 19. June 1989.

17. Bartlett, D. F. and Tew, W. L., "Possible Effect of the Local Terrain on the Australian Fifth-Force Measurement," Phys. Rev. D, Vol. 40, p. 673, 15 July 1989.

18. Thomas, J., "Testing the Inverse-Square Law of Gravity: Error and Design with the Upward Continuation Integral," Phys. Rev. D, Vol. 40, p. 1735, 15 Sept. 1989.

19. Thomas, J., Kasameyer, P., Fackler, O., Felske, D., Harris, R., Kammeraad, J., Millett, M. and Mugge, M., "Testing the Inverse-Square Law of Gravity on a 465-m Tower," Phys. Rev. Lett., Vol. 63, p. 1902, 30 Oct. 1989.

20. Müller, G., Zürn, W., Lindner, K. and Rösch, N., "Determination of the Gravitational Constant by an Experiment at a Pumped-Storage Reservoir," Phys, Rev. Lett., Vol. 63, p. 2621, 11 Dec. 1989.

21. Cowsik, R., Krishnan, N., Tandon, S. N. and Unnikrishnan, S., "Strength of Intermediate-Range Forces Coupling to Isospin," Phys. Rev. Lett., Vol. 64, p. 336, 22 Jan. 1990.

22. Jekeli, C., Eckhardt, D. H. and Romaides, A. J., "Tower Gravity Experiment: No Evidence for Non-Newtonian Gravity," Phys. Rev. Lett., Vol. 64, p. 1204, 12 Mar. 1990.

23. Nelson, P. G., Graham, D. M. and Newman, R. D., "Search for an Intermediate-Range Composition-Dependent Force Coupling to N-Z," Phys. Rev. D, Vol. 42, p. 963. 15 Aug. 1990.

24. Thomas, J. and Vogel, P., "Testing the Inverse-Square Law of Gravity in Boreholes at the Nevada Test Site," Phys. Rev. Lett., Vol. 65, p. 1173, 3 Sept. 1990.

25. Zumberge, M. A., Ander, M. E., Lautzenhiser, T. V., Parker, R. L., Aiken, C. L. V., Gorman, M. R., Nieto, M. M., Cooper, A. P. R., Ferguson, J. F., Fisher, E., Greer, J., Hammer, P., Hansen, B. L., McMechan, G. A., Sasagawa, G. S., Sidles, C., Stevenson, J. M., and Wirtz, J., "The Greenland Gravitational Constant Experiment," J. G. R., Vol. 95, p. 15,483, Sept. 10, 1990.

26. Speake, C. C., Niebauer, T. M., McHugh, M. P., Keyser, P. T., Faller, J. E., Cruz, J. Y., Harrison, J. C., Mäkinen, J. and Beruff, R. B., "Test of the Inverse-Square Law of Gravitation Using the 300-m Tower at Erie, Colorado," Phys. Rev. Lett., Vol. 65, p. 1967, 15 Oct. 1990.

27. Weiss, R. A., Gauge Theory of Thermodynamics, K&W Publications, Vicksburg, MS, 1989.

28. Weiss, R. A., Relativistic Thermodynamics, Exposition Press, New York, 1976.

29. Weiss, R. A., "Electromagnetism and Gravity," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990, p. 265.

30. Huang, K., Statistical Mechanics, John Wiley, New York, 1963.

31. Mayer, J. E.and Mayer, M. G., Statistical Mechanics, John Wiley, New York, 1977.

32. Hill, T. L., An Introduction to Statistical Mechanics, Addison-Wesley, Reading, MA, 1960.

33. Eyring, H., Henderson, D., Stover, B. J. and Eyring, E. M., Statistical Mechanics and Dynamics, John Wiley, New York, 1964.

34. Tolman, R. C., The Principles of Statistical Mechanics, Oxford, New York, 1938.

35. Weiss, R. A., "Thermal Radiation of High-T_c Superconductors," Eighth Army Conference on Applied Mathematics and Computing, Cornell University, Ithaca, NY, ARO 91-1, June 19-22, 1990. p. 399.

Robust Stabilization, Robust Performance, and Disturbance Attenuation for Uncertain Linear Systems

Yeih J. Wang and Leang S. Shieh †

Department of Electrical Engineering, Cullen College of Engineering
University of Houston
University park, Houston, TX 77204–4793, USA

John W. Sunkel

Avionics Systems Division, NASA–Johnson Space Center
Houston, TX 77058, USA

Abstract: This paper presents a linear quadratic regulator approach to the robust stabilization, robust performance, and disturbance attenuation of uncertain linear systems. The state-feedback designed systems provide both robust stability with optimal performance and disturbance attenuation with H_∞-norm bounds. The proposed approach can be applied to *matched* and/or *mismatched* uncertain linear systems. For a matched uncertain linear system, it is shown that the disturbance-attenuation robust-stabilizing controllers with or without optimal performance always exist and can be easily determined without searching; whereas, for a mismatched uncertain linear system, the introduced tuning parameters greatly enhance the flexibility of finding the disturbance-attenuation robust-stabilizing controllers.

† This work was supported in part by the U.S. Army Research Office, under contract DAAL–03–87–K0001, and NASA–Johnson Space Center, under grants NAG 9–380 and NAG 9–385.

1. Introduction

The problems of robust stabilization, robust performance, and disturbance attenuation of uncertain linear systems have drawn much attention recently. Nonlinear robust control laws that stabilize uncertain linear systems satisfying *matching conditions* were developed by Leitmann [7]. Feedback control designs based on the algebraic Riccati equation (ARE), which adjust a scalar to achieve stabilization of the systems with uncertainty parameters bounded by constraint sets, were derived by Petersen and Hollot [9], Petersen [10], Schmitendorf [12], and Khargonekar *et al.* [6]. These approaches have generally utilized the concept that a given ARE-based control law guarantees the existence of a quadratic Lyapunov function (and hence, stability) for the closed-loop uncertain linear system. Also, other recent research attention, e.g., Bernstein and Haddad [2], Doyle *et al.* [3], Glover and Doyle [4], and Petersen [11], has been given to the ARE-based control designs which stabilize a nominal system and reduce the effect of disturbances on the output to a prespecified level. More recently, Veillette *et al.* [15] has proposed an ARE-based design which not only robustly stabilizes an uncertain linear system with the structured uncertainty in the system matrix, but also provides disturbance attenuation with a robust H_∞-norm bound.

In this paper, based on linear quadratic regulator theory and Lyapunov stability theory, we develop linear state-feedback control laws for robust stabilization, robust performance, and disturbance attenuation of a given uncertain linear system with the uncertainties existing both in the system matrix and the input matrix. The proposed design procedures can be applied to both matched and mismatched systems. The paper is organized as follows. First, the matching conditions for uncertain linear systems to be stabilized with prespecified disturbance attenuation level are defined in Section 2. It is shown that many dynamic systems, described by second-order monic vector differential equations, often satisfy these matching conditions. Next, linear robust stabilizing controllers which provide disturbance attenuation and optimal performance for matched systems with norm-bounded or structured uncertainty matrices are developed in Section 3. Also, it is shown that linear disturbance-attenuation robust-stabilizing controllers with optimal performance for matched systems always exist and can be easily determined without seaching. Then, in order to achieve the stabilization and disturbance attenuation of mismatched systems

with norm-bounded or structured uncertainty matrices, alternative linear distrurbance-attenuation robust-stabilizing controllers are proposed in Section 4. To demonstrate the proposed methods, two examples are illustrated in Section 5, and the results are summarized in the conclusion in Section 6.

2. Nomenclature, Systems, and Definitions

Throughout this paper, we denote:

$\sigma_{\max}(M)$	maximum singular value of a matrix M;
$\sigma_{\min}(M)$	minimum singular value of a matrix M;
$\|M\|$	matrix norm, $\|M\| \triangleq \sigma_{\max}(M) = \lambda_{\max}^{1/2}(M^T M)$;
I	identity matrix of appropriate dimension;
0	null matrix of appropriate dimension;
$M > (\geq)\ 0$	matrix M is symmetric positive (semi)definite;
$M < (\leq)\ 0$	matrix M is symmetric negative (semi)definite;
$P > (\geq)\ Q$	means $P - Q > (\geq)\ 0$;
$P < (\leq)\ Q$	means $P - Q < (\leq)\ 0$.

Consider the uncertain linear system

$$\dot{x}(t) = [A + \Delta A]x(t) + [B + \Delta B]u(t) + Dw(t), \tag{1a}$$

$$y(t) = Cx(t), \tag{1b}$$

where $x(t) \in \mathcal{R}^n$ is the state, $u(t) \in \mathcal{R}^m$ is the control, $w(t) \in \mathcal{R}^q$ is the disturbance, $y(t) \in \mathcal{R}^p$ is the output, $A \in \mathcal{R}^{n\times n}$, $B \in \mathcal{R}^{n\times m}$, $D \in \mathcal{R}^{n\times q}$, and $C \in \mathcal{R}^{p\times n}$ are the nominal system matrix, input matrix, disturbance matrix, and output matrix, respectively, and ΔA and ΔB are the associated uncertainty matrices of appropriate dimensions with respect to A and B. We assume that the nominal system (A, B) is controllable. Without loss of generality, we also assume that B has full rank. Our objective is to design a linear state-feedback control law $u(t) = Kx(t)$ such that the resulting closed-loop system matrix $A_c \triangleq [A + \Delta A + (B + \Delta B)K]$ is asymptotically stable, and the resulting closed-loop system is optimal with respect to a certain performance index, and the H_∞-norm of the closed-loop transfer functon matrix $H(s) \triangleq C[sI - A_c]^{-1}D$ from the disturbance input $w(t)$ to

the output $y(t)$ is less than or equal to some prespecified disturbance-attenuation value δ, i.e., $H^T(-j\omega)H(j\omega) \leq \delta^2 I$ for all $\omega \in \mathcal{R}$.

To proceed with the derivation for such a control law, we need to consider two classes of uncertain linear systems which are matched and mismatched. The system in (1) is called a matched uncertain linear system if there exist matrices $E \in \mathcal{R}^{m\times n}$, $F \in \mathcal{R}^{m\times m}$, and $G \in \mathcal{R}^{m\times q}$ such that

(i) $\Delta A = BE$,

(ii) $\Delta B = BF$, and $\|F\| < 1$ or $2I + F + F^T > 0$, and

(iii) $D = BG$.

The matching conditions (i) and (ii) constitute sufficient conditions [7] for the system to be stabilizable. We shall show that the uncertain linear system is, in fact, linearly stabilizable with any disturbance attenuation $\delta > 0$ if it satisfies conditions (i–iii).

It is important to note that a dynamical system [13] which can be modeled by a second-order monic vector differential equation is often a matched system. This fact can be verified as follows. Consider the second-order monic vector differential equation

$$\ddot{q}(t) + (A_1 + \Delta A_1)\dot{q}(t) + (A_2 + \Delta A_2)q(t) = (B_1 + \Delta B_1)u(t) + D_1 w(t), \tag{2a}$$

$$y(t) = C_1\dot{q}(t) + C_2 q(t), \tag{2b}$$

where $q(t) \in \mathcal{R}^m$, $u(t) \in \mathcal{R}^m$, $w(t) \in \mathcal{R}^m$, and $y(t) \in \mathcal{R}^m$ are partial state, input, disturbance, and output, respectively. The state-variable realization of the second-order vector differential equation in (2) in a block companion form is given by

$$\dot{x}(t) = [A + \Delta A]x(t) + [B + \Delta B]u(t) + Dw(t), \tag{3a}$$

$$y(t) = Cx(t), \tag{3b}$$

where

$$A = \begin{bmatrix} 0 & I \\ -A_2 & -A_1 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 \\ B_1 \end{bmatrix}, \qquad D = \begin{bmatrix} 0 \\ D_1 \end{bmatrix} = BG, \qquad C = [C_2, C_1],$$

$$\Delta A = \begin{bmatrix} 0 & 0 \\ -\Delta A_2 & -\Delta A_1 \end{bmatrix} = BE, \qquad \Delta B = \begin{bmatrix} 0 \\ \Delta B_1 \end{bmatrix} = BF,$$

with $E = [-B_1^{-1}\Delta A_2, -B_1^{-1}\Delta A_1]$, $F = B_1^{-1}\Delta B_1$, and $G = B_1^{-1}D_1$ assuming $\det(B_1) \neq 0$. Obviously, the system in (3) satisfies the matching conditions (i–iii) provided that $\|F\| < 1$ or $2I + F + F^T > 0$.

Remark 1. In general, if the uncertain linear system in (1) satisfies the matching conditions (i–iii), the matrices E, F, and G can be obtained from the given ΔA, ΔB, and D, respectively, using a technique based on the singular value decomposition (SVD) (see Appendix). ∎

3. Guaranteed Disturbance-Attenuation Robust-Stabilizing Controllers with Optimal Performance for Matched Systems

Consider the following matched uncertain linear system:

$$\dot{x}(t) = [A + BE]x(t) + [B + BF]u(t) + BGw(t), \tag{4a}$$

$$y(t) = Cx(t). \tag{4b}$$

Suppose that the only information about the uncertainty matrices in (4) is that their matrix norms are bounded by

$$\|E\| \leq \alpha \qquad \text{and} \qquad \|F\| \leq \beta < 1. \tag{5}$$

The following theorem guarantees that a disturbance-attenuation robust-stabilizing controller (with optimal performance if $\|F\| \leq \beta < \frac{1}{2}$) exists for the matched uncertain linear system in (4) having the constraints in (5).

Theorem 1. Consider the matched uncertain linear system in (4) with the norm-bounded uncertainty matrices described in (5). Let $\delta > 0$ be any given disturbance-attenuation constant and $Q \in \mathcal{R}^{n\times n}$ any given symmetric positive-definite matrix. Select any positive constants ε_1 and ε_2 satisfying $\varepsilon_1 \in \left(0, \frac{1-\beta}{\alpha}\right)$ and $\varepsilon_2 \in \left(0, \frac{(1-\beta-\varepsilon_1\alpha)\delta}{\sigma^2_{\max}(G)}\right)$ and let $P \in \mathcal{R}^{n\times n}$ be the symmetric positive-definite solution of the following Riccati equation:

$$A^TP + PA - PB\left[(1-\beta-\varepsilon_1\alpha)I - \frac{\varepsilon_2}{\delta}GG^T\right]B^TP + \frac{\alpha}{\varepsilon_1}I + \frac{1}{\varepsilon_2\delta}C^TC + Q = 0. \tag{6}$$

Then, a disturbance-attenuation robust-stabilizing control law with the attenuation constant δ is given by $u(t) = Kx(t)$, where $K = -\gamma B^T P$ with $\gamma \geq \frac{1}{2}$. That is, the closed-loop system matrix $A_c = A + BE + (B + BF)K$ is asymptotically stable and the H_∞-norm of the closed-loop transfer functon matrix $H(s) = C[sI - A_c]^{-1}D$ (here, $D = BG$) is less than or equal to the δ for all admissible uncertainty matrices E and F in (5). Furthermore, if $\|F\| \leq \beta < \frac{1}{2}$, then the state-feedback control law $u(t) = -\gamma B^T P x(t)$ with $\gamma \geq \frac{1-\beta}{1-2\beta}$ is also optimal with respect to a certain quadratic performance index.

Proof. To show the robust stabilization, we define

$$Q_c \triangleq -A_c^T P - P A_c. \tag{7a}$$

Then

$$Q_c = -A^T P - PA - E^T B^T P - PBE + \gamma PB(2I + F^T + F)B^T P. \tag{7b}$$

From (6), it follows that

$$\begin{aligned} Q_c =& PB\left[(2\gamma - 1 + \beta)I + \gamma(F^T + F)\right]B^T P \\ &+ \varepsilon_1 \alpha PBB^T P + \frac{\alpha}{\varepsilon_1} I - E^T B^T P - PBE + \frac{\varepsilon_2}{\delta} PBGG^T B^T P + \frac{1}{\varepsilon_2 \delta} C^T C + Q \\ \geq& (2\gamma - 1)(1 - \beta)PBB^T P + \left[\sqrt{\frac{\varepsilon_1}{\alpha}} PBE - \sqrt{\frac{\alpha}{\varepsilon_1}} I\right]\left[\sqrt{\frac{\varepsilon_1}{\alpha}} PBE - \sqrt{\frac{\alpha}{\varepsilon_1}} I\right]^T \\ &+ \frac{\varepsilon_2}{\delta} PDD^T P + \frac{1}{\varepsilon_2 \delta} C^T C + Q. \end{aligned} \tag{7c}$$

Hence

$$Q_c \geq \frac{\varepsilon_2}{\delta} PDD^T P + \frac{1}{\varepsilon_2 \delta} C^T C + Q > 0 \quad \text{for} \quad \|F\| \leq \beta < 1 \quad \text{and} \quad \gamma \geq \frac{1}{2}, \tag{7d}$$

or

$$Q_c > \frac{\varepsilon_2}{\delta} PDD^T P + \frac{1}{\varepsilon_2 \delta} C^T C \geq 0 \quad \text{for} \quad \|F\| \leq \beta < 1 \quad \text{and} \quad \gamma \geq \frac{1}{2}. \tag{7e}$$

Thus, based on Lyapunov stability theory [1], A_c is asymptotically stable for $\|F\| \leq \beta < 1$ and $\gamma \geq \frac{1}{2}$.

To show the disturbance attenuation, we utilize the equality in (7a) and the inequality in (7e) as follows:

$$(-j\omega I - A_c)^T P + P(j\omega I - A_c) - \frac{\varepsilon_2}{\delta} PDD^T P - \frac{1}{\varepsilon_2 \delta} C^T C > 0 \tag{8a}$$

for all $\omega \in \mathcal{R}$. Now, we define $\phi(j\omega) \triangleq (j\omega I - A_c)^{-1}$, and premultiply $D^T\phi^T(-j\omega)$ and postmultiply $\phi(j\omega)D$ to the inequality in (8a) to obtain

$$D^TP\phi(j\omega)D + D^T\phi^T(-j\omega)PD - \frac{\varepsilon_2}{\delta}D^T\phi^T(-j\omega)PDD^TP\phi(j\omega)D - \frac{1}{\varepsilon_2\delta}D^T\phi^T(-j\omega)C^TC\phi(j\omega)D \geq 0. \quad (8b)$$

Then, we complete a square term as follows:

$$\left[\sqrt{\frac{\delta}{\varepsilon_2}}I - \sqrt{\frac{\varepsilon_2}{\delta}}D^T\phi^T(-j\omega)PD\right]\left[\sqrt{\frac{\delta}{\varepsilon_2}}I - \sqrt{\frac{\varepsilon_2}{\delta}}D^T\phi^T(j\omega)PD\right]^T \geq 0. \quad (8c)$$

Thus, from (8b) and (8c) we obtain

$$\frac{\delta}{\varepsilon_2}I \geq \frac{1}{\varepsilon_2\delta}D^T\phi^T(-j\omega)C^TC\phi(j\omega)D = \frac{1}{\varepsilon_2\delta}H^T(-j\omega)H(j\omega) \quad (8d)$$

Hence, $\|H(j\omega)\| \leq \delta$ for all $\omega \in \mathcal{R}$.

To show the robust performance, we let $\hat{A} = A + BE$, $\hat{B} = B + BF$, and $\hat{R} = \frac{1}{\gamma}I$, where $\hat{R}$ is an input weighting matrix of a quadratic performance index. From (7b) and (7c), we have the following Riccati equation:

$$\hat{Q} \triangleq -\hat{A}^TP - P\hat{A} + P\hat{B}\hat{R}^{-1}\hat{B}^TP$$
$$\geq PB[(\gamma - 2\gamma\beta - 1 + \beta)I + \gamma FF^T]B^TP + \frac{\varepsilon_2}{\delta}PDD^TP + \frac{1}{\varepsilon_2\delta}C^TC + Q. \quad (9)$$

Therefore, if $\|F\| \leq \beta < \frac{1}{2}$, then $\hat{Q} > 0$ for $\gamma \geq \frac{1-\beta}{1-2\beta}$, where $\hat{Q}$ is a state weighting matrix of a quadratic performance index. That is, the state-feedback control law $u(t) = -\gamma B^TPx(t)$ for $\gamma \geq \frac{1-\beta}{1-2\beta}$ is optimal with respect to the quadratic performance index

$$J = \frac{1}{2}\int_0^\infty [x^T(t)\hat{Q}x(t) + u^T(t)\hat{R}u(t)]dt.$$

■

Remark 2. The Riccati equation in (6) is constructed to account for the uncertain linear system in (4) with the uncertainty matrices in (5) and the disturbance attenuation δ. If there is no system uncertainty (i.e., $\alpha = 0$ and $\beta = 0$) and the disturbance attenuation is not required (i.e., $\delta \to \infty$), the *augmented* Riccati equation in (6) reduces to an *ordinary* Riccati equation which arises in the linear quadratic regulator problem [1]. We assume $Q > 0$ to facilitate the proof; however, if (A, C) is observable, this assumption can be relaxed to $Q \geq 0$. ■

Corollary 1. Consider the matched uncertain linear system in (4) with the norm-bounded uncertainty matrices described in (5). Let $\delta > 0$ be any given disturbance-attenuation constant, $Q \in \mathcal{R}^{n\times n}$ any given symmetric positive-definite matrix, and $h \geq 0$ a prescribed degree of stability [1]. Select any positive constants ε_1 and ε_2 satisfying $\varepsilon_1 \in \left(0, \frac{1-\beta}{\alpha}\right)$ and $\varepsilon_2 \in \left(0, \frac{(1-\beta-\varepsilon_1\alpha)\delta}{\sigma^2_{\max}(G)}\right)$ and let $P \in \mathcal{R}^{n\times n}$ be the symmetric positive-definite solution of the following Riccati equation:

$$(A+hI)^T P + P(A+hI) - PB\left[(1-\beta-\varepsilon_1\alpha)I - \frac{\varepsilon_2}{\delta}GG^T\right]B^T P + \frac{\alpha}{\varepsilon_1}I + \frac{1}{\varepsilon_2\delta}C^T C + Q = 0. \quad (10)$$

Then, a disturbance-attenuation robust-stabilizing control law with the attenuation constant δ is given by $u(t) = Kx(t)$, where $K = -\gamma B^T P$ with $\gamma \geq \frac{1}{2}$. Furthermore, the closed-loop system matrix $A_c = A + BE + (B + BF)K$ has a prescribed degree of stability h [1] for all admissible uncertainty matrices E and F in (5). ■

Now we consider the matched uncertain linear system in (4) with structured uncertainty matrices $E \in \mathcal{R}^{m\times n}$ and $F \in \mathcal{R}^{m\times m}$ described by

$$E = \sum_{i=1}^{k} e_i E_i \qquad \text{with} \quad |e_i| \leq \bar{e}_i, \quad (11a)$$

and

$$F = \sum_{i=1}^{l} f_i F_i \qquad \text{with} \quad |f_i| \leq \bar{f}_i, \quad (11b)$$

respectively, where e_i and f_i are uncertain parameters, and E_i and F_i are known constant matrices with each matrix may having rank greater than one. Applying the SVD in (A5) to the matrices E_i and F_i, we can decompose each E_i and F_i as (see Appendix)

$$E_i = T_i U_i^T \qquad \text{and} \qquad F_i = V_i W_i^T, \quad (11c)$$

where T_i, U_i, V_i, and W_i are weighted unitary matrices with appropriate dimensions.

To derive the disturbance-attenuation robust-stabilizing controllers for the matched system in (4) with the structured uncertainty matrices described in (11), we define symmetric positive-semidefinite matrices $T \in \mathcal{R}^{m\times m}$, $U \in \mathcal{R}^{n\times n}$, and $V \in \mathcal{R}^{m\times m}$ as follows:

$$T \triangleq \sum_{i=1}^{k} \bar{e}_i T_i T_i^T, \qquad U \triangleq \sum_{i=1}^{k} \bar{e}_i U_i U_i^T, \quad (12a)$$

$$V \triangleq \frac{1}{2}\sum_{i=1}^{l} \bar{f}_i(V_i V_i^T + W_i W_i^T), \tag{12b}$$

with the matrices T_i, U_i, V_i, and W_i as in (11). It can be shown that $2V + F + F^T \geq 0$. Also, from the matching condition (ii), we require $2I + F + F^T > 0$. As a result, we assume that

$$I - V > 0. \tag{12c}$$

The following theorem guarantees that a disturbance-attenuation robust-stabilizing controller with optimal performance exists for the matched uncertain linear system in (4) with the structured uncertainty matrices in (11).

Theorem 2. Consider the matched uncertain linear system in (4) with the structured uncertainty matrices described by (11). Let $\delta > 0$ be any given disturbance-attenuation constant and $Q \in \mathcal{R}^{n\times n}$ any given symmetric positive-definite matrix. Select any positive constants ε_1 and ε_2 satisfying $\varepsilon_1 \in \left(0, \frac{1-\sigma_{\max}(V)}{\sigma_{\max}(T)}\right)$ and $\varepsilon_2 \in \left(0, \frac{(1-\sigma_{\max}(V)-\varepsilon_1\sigma_{\max}(T))\delta}{\sigma^2_{\max}(G)}\right)$ and let $P \in \mathcal{R}^{n\times n}$ be the symmetric positive-definite solution of the following Riccati equation:

$$A^T P + PA - PB\left[I - V - \varepsilon_1 T - \frac{\varepsilon_2}{\delta} GG^T\right] B^T P + \frac{1}{\varepsilon_1} U + \frac{1}{\varepsilon_2 \delta} C^T C + Q = 0, \tag{13}$$

where the matrices T, U, and V are defined in (12). Then, a disturbance-attenuation robust-stabilizing control law with the attenuation constant δ is given by $u(t) = Kx(t)$, where $K = -\gamma B^T P$ with $\gamma \geq \frac{1}{2}$. Furthermore, if $0 \leq V < \frac{1}{2} I$, then the state-feedback control law $u(t) = -\gamma B^T P x(t)$ with $\gamma \geq \frac{1-\sigma_{\min}(V)}{1-2\sigma_{\max}(V)}$ is also optimal with respect to a certain quadratic performance index.

Proof. Define Q_c as in (7a). From (13), it follows that

$$\begin{aligned} Q_c =& PB\left[(2\gamma - 1)I + V + \gamma(F^T + F)\right] B^T P \\ &+ \varepsilon_1 PBTB^T P + \frac{1}{\varepsilon_1} U - E^T B^T P - PBE + \frac{\varepsilon_2}{\delta} PBGG^T B^T P + \frac{1}{\varepsilon_2 \delta} C^T C + Q. \end{aligned}$$

Since

$$\begin{aligned} 2V + F^T + F &= \sum_{i=1}^{l} \left[\bar{f}_i(V_i V_i^T + W_i W_i^T) + f_i(V_i W_i^T + W_i V_i^T)\right] \\ &\geq \sum_{i=1}^{l} |f_i|[V_i \pm W_i][V_i \pm W_i]^T \geq 0, \end{aligned}$$

and

$$\left[\varepsilon_1 PBTB^T P + \frac{1}{\varepsilon_1}U - E^T B^T P - PBE\right]$$
$$= \sum_{i=1}^{k}\left[\bar{e}_i\left(\varepsilon_1 PBT_iT_i^T B^T P + \frac{1}{\varepsilon_1}U_iU_i^T\right) - e_i(U_iT_i^T B^T P + PBT_iU_i^T)\right]$$
$$\geq \sum_{i=1}^{k}|e_i|\left[\sqrt{\varepsilon_1}PBT_i \pm \frac{1}{\sqrt{\varepsilon_1}}U_i\right]\left[\sqrt{\varepsilon_1}PBT_i \pm \frac{1}{\sqrt{\varepsilon_1}}U_i\right]^T \geq 0.$$

It follows that

$$Q_c \geq PB[(2\gamma-1)I + V - 2\gamma V]B^T P + \frac{\varepsilon_2}{\delta}PBGG^T B^T P + \frac{1}{\varepsilon_2\delta}C^T C + Q$$
$$= (2\gamma-1)PB(I-V)B^T P + \frac{\varepsilon_2}{\delta}PBGG^T B^T P + \frac{1}{\varepsilon_2\delta}C^T C + Q.$$

Hence, $Q_c \geq \frac{\varepsilon_2}{\delta}PDD^T P + \frac{1}{\varepsilon_2\delta}C^T C + Q > 0$ for $I - V > 0$ and $\gamma \geq \frac{1}{2}$. Thus, based on Lyapunov stability theory [1], A_c is asymptotically stable for $I - V > 0$ and $\gamma \geq \frac{1}{2}$.

The proofs for disturbance attenuation and the optimality condition when $0 \leq V < \frac{1}{2}I$ are similar to those in Theorem 1 and hence omitted. ■

4. Disturbance-Attenuation Robust-Stabilizing Controllers for Mismatched Systems

Consider the following mismatched uncertain linear system described by

$$\dot{x}(t) = [A + \Delta A]x(t) + [B + \Delta B]u(t) + Dw(t), \tag{14a}$$
$$y(t) = Cx(t). \tag{14b}$$

Suppose that the only information about the uncertainty matrices $\Delta A \in \mathcal{R}^{n\times n}$ and $\Delta B \in \mathcal{R}^{n\times m}$ in (14) is that the matrix norms are bounded by

$$\|\Delta A\| \leq \alpha \qquad \text{and} \qquad \|\Delta B\| \leq \beta. \tag{15}$$

The following theorem will be utilized to find a disturbance-attenuation robust-stabilizing controller for the mismatched uncertain system in (14) with the constraints in (15).

Theorem 3. Consider the mismatched uncertain system in (14) with the norm-bounded uncertainty matrices described in (15). Let $\delta > 0$ be any given disturbance-attenuation constant and $Q \in \mathcal{R}^{n\times n}$ any given symmetric positive-definite matrix. Suppose that there exist any positive constants $\varepsilon_1 > 0$, $\varepsilon_2 \in \left(0, \frac{2}{\beta}\right)$, and $\varepsilon_3 > 0$, such that the Riccati equation

$$A^TP+PA-P\Big[\Big(1-\frac{\varepsilon_2\beta}{2}\Big)BB^T-\Big(\varepsilon_1\alpha+\frac{\beta}{2\varepsilon_2}\Big)I-\frac{\varepsilon_3}{\delta}DD^T\Big]P+\frac{\alpha}{\varepsilon_1}I+\frac{1}{\varepsilon_3\delta}C^TC+Q=0 \quad (16)$$

has a symmetric positive-definite solution $P \in \mathcal{R}^{n\times n}$. Then, a disturbance-attenuation robust-stabilizing control law with the attenuation constant δ is given by $u(t) = Kx(t)$, where $K = -\gamma B^TP$ with γ satisfying either

$$\frac{1}{\varepsilon_2\beta}-\frac{1}{2}\geq\gamma\geq\frac{1}{2} \qquad \text{or} \qquad \frac{1}{2}\geq\gamma\geq\frac{1}{\varepsilon_2\beta}-\frac{1}{2}>0. \quad (17)$$

That is, the closed-loop system matrix $A_c = A+\Delta A+(B+\Delta B)K$ is asymptotically stable and the H_∞-norm of the closed-loop transfer functon matrix $H(s) = C[sI-A_c]^{-1}D$ is less than or equal to the δ for all admissible uncertainty matrices ΔA and ΔB in (15).

Proof. Suppose that the Riccati equation in (16) has a symmetric positive-definite solution P. Define Q_c as in (7a). From (16), it follows that

$$\begin{aligned}Q_c =& P\Big[(2\gamma-1)BB^T+\frac{\varepsilon_2\beta}{2}BB^T+\frac{\beta}{2\varepsilon_2}I+\gamma B\Delta B^T+\gamma\Delta BB^T\Big]P\\ &+\Big[\varepsilon_1\alpha PP+\frac{\alpha}{\varepsilon_1}I-\Delta A^TP-P\Delta A\Big]+\frac{\varepsilon_3}{\delta}PDD^TP+\frac{1}{\varepsilon_3\delta}C^TC+Q.\end{aligned}$$

Since

$$\begin{aligned}&2\gamma^2\varepsilon_2\beta BB^T+\frac{\beta}{2\varepsilon_2}I+\gamma B\Delta B^T+\gamma\Delta BB^T\\ &\qquad\geq\Big[\gamma\sqrt{2\varepsilon_2\beta}B+\frac{1}{\sqrt{2\varepsilon_2\beta}}\Delta B\Big]\Big[\gamma\sqrt{2\varepsilon_2\beta}B+\frac{1}{\sqrt{2\varepsilon_2\beta}}\Delta B\Big]^T\geq 0\end{aligned}$$

and

$$\begin{aligned}&\varepsilon_1\alpha PP+\frac{\alpha}{\varepsilon_1}I-\Delta A^TP-P\Delta A\\ &\qquad\geq\Big[\sqrt{\frac{\varepsilon_1}{\alpha}}P\Delta A-\sqrt{\frac{\alpha}{\varepsilon_1}}I\Big]\Big[\sqrt{\frac{\varepsilon_1}{\alpha}}P\Delta A-\sqrt{\frac{\alpha}{\varepsilon_1}}I\Big]^T\geq 0,\end{aligned}$$

we obtain the following inequality:

$$Q_c \geq \Big[2\gamma - 1 + \frac{\varepsilon_2\beta}{2} - 2\gamma^2\varepsilon_2\beta\Big]PBB^TP + \frac{\varepsilon_2}{\delta}PDD^TP + \frac{1}{\varepsilon_2\delta}C^TC + Q$$
$$= \Big[(2\gamma-1)\Big(1 - \frac{\varepsilon_2\beta}{2}(2\gamma+1)\Big)\Big]PBB^TP + \frac{\varepsilon_3}{\delta}PDD^TP + \frac{1}{\varepsilon_3\delta}C^TC + Q.$$

If γ satisfies either inequality in (17), which is equivalent to satisfying the inequality

$$(2\gamma-1)\Big(1 - \frac{\varepsilon_2\beta}{2}(2\gamma+1)\Big) \geq 0,$$

then, $Q_c \geq \frac{\varepsilon_3}{\delta}PDD^TP + \frac{1}{\varepsilon_3\delta}C^TC + Q > 0$. Thus, based on Lyapunov stability theory [1], the obtained controller $u(t)$ stabilizes the mismatched system in (14) with the constraints in (15).

The proof for $\|H\|_\infty \leq \delta$ is similar to that in Theorem 1 and hence omitted. ■

Remark 3. The parameter ε_2 in (16) is restricted to be in the range of $\Big(0, \frac{2}{\beta}\Big)$ such that the term $\Big(1 - \frac{\varepsilon_2\beta}{2}\Big)$ in (16) is greater than zero. ■

Now we consider the uncertain linear system in (14) with structured uncertainty matrices $\Delta A \in \mathcal{R}^{n\times n}$ and $\Delta B \in \mathcal{R}^{n\times m}$ described by

$$\Delta A = \sum_{i=1}^{k} a_i A_i \qquad \text{with} \quad |a_i| \leq \bar{a}_i, \tag{18a}$$

and

$$\Delta B = \sum_{i=1}^{l} b_i B_i \qquad \text{with} \quad |b_i| \leq \bar{b}_i, \tag{18b}$$

respectively, where a_i and b_i are uncertain parameters, and A_i and B_i are known constant matrices with each matrix may having rank greater than one. Applying the SVD in (A5) to A_i and B_i, we can decompose each A_i and B_i as (see Appendix)

$$A_i = T_i U_i^T \qquad \text{and} \qquad B_i = V_i W_i^T, \tag{18c}$$

where T_i, U_i, V_i, and W_i are weighted unitary matrices with appropriate dimensions.

To derive the disturbance-attenuation robust-stabilizing controllers for the system in (14) with the structured uncertainty matrices described by (18), we define symmetric

positive-semidefinite matrices $T \in \mathcal{R}^{n\times n}$, $U \in \mathcal{R}^{n\times n}$, $V \in \mathcal{R}^{n\times n}$, and $W \in \mathcal{R}^{m\times m}$ as follows:

$$T \triangleq \sum_{i=1}^{k} \bar{a}_i T_i T_i^T, \qquad U \triangleq \sum_{i=1}^{k} \bar{a}_i U_i U_i^T, \tag{19a}$$

$$V \triangleq \frac{1}{2}\sum_{i=1}^{l} \bar{b}_i V_i V_i^T, \qquad W \triangleq \frac{1}{2}\sum_{i=1}^{l} \bar{b}_i W_i W_i^T, \tag{19b}$$

with the matrices T_i, U_i, V_i, and W_i as in (18). The following theorem will be utilized to find a disturbance-attenuation robust-stabilizing controller for the mismatched uncertain system in (14) having the constraints in (18).

Theorem 4. Consider the mismatched uncertain linear system in (14) with the structured uncertainty matrices described in (18). Let $\delta > 0$ be any given disturbance-attenuation constant and $Q \in \mathcal{R}^{n\times n}$ any given symmetric positive-definite matrix. Suppose that there exist any positive constants $\varepsilon_1 > 0$, $\varepsilon_2 \in \left(0, \frac{1}{\sigma_{\max}(W)}\right)$, and $\varepsilon_3 > 0$, such that the Riccati equation

$$A^TP + PA - P\Big[BB^T - \varepsilon_1 T - \varepsilon_2 BWB^T - \frac{1}{\varepsilon_2}V - \frac{\varepsilon_3}{\delta}DD^T\Big]P + \frac{1}{\varepsilon_1}U + \frac{1}{\varepsilon_3\delta}C^TC + Q = 0 \tag{20}$$

has a symmetric positive-definite solution $P \in \mathcal{R}^{n\times n}$, where T, U, V, and W are defined in (19). Then, a disturbance-attenuation robust-stabilizing control law with the attenuation constant δ is given by $u(t) = Kx(t)$, where $K = -\gamma B^TP$ with γ satisfying either

$$\frac{1}{2\varepsilon_2\sigma_{\max}(W)} - \frac{1}{2} \geq \gamma \geq \frac{1}{2} \qquad \text{or} \qquad \frac{1}{2} \geq \gamma \geq \frac{1}{2\varepsilon_2\sigma_{\min}(W)} - \frac{1}{2} > 0. \tag{21}$$

Proof. Suppose that the Riccati equation in (20) has a symmetric positive-definite solution P. Define Q_c as in Theorem 1. From (20), it follows that

$$\begin{aligned} Q_c =& P\Big[(2\gamma-1)BB^T + \varepsilon_2 BWB^T + \frac{1}{\varepsilon_2}V + \gamma B\Delta B^T + \gamma\Delta BB^T\Big]P \\ &+ \Big[\varepsilon_1 PTP + \frac{1}{\varepsilon_1}U - \Delta A^TP - P\Delta A\Big] + \frac{\varepsilon_3}{\delta}PDD^TP + \frac{1}{\varepsilon_3\delta}C^TC + Q. \end{aligned}$$

Since

$$\begin{aligned} &4\gamma^2\varepsilon_2 BWB^T + \frac{1}{\varepsilon_2}V + \gamma B\Delta B^T + \gamma\Delta BB^T \\ &\qquad \geq \sum_{i=1}^{l} |b_i|\Big[\gamma\sqrt{2\varepsilon_2}BW_i \pm \frac{1}{\sqrt{2\varepsilon_2}}V_i\Big]\Big[\gamma\sqrt{2\varepsilon_2}BW_i \pm \frac{1}{\sqrt{2\varepsilon_2}}V_i\Big]^T \geq 0 \end{aligned}$$

and

$$\varepsilon_1 PTP + \frac{1}{\varepsilon_1}U - \Delta A^T P - P\Delta A$$
$$\geq \sum_{i=1}^{k} |a_i| \Big[\sqrt{\varepsilon_1} PT_i \pm \frac{1}{\sqrt{\varepsilon_1}} U_i\Big] \Big[\sqrt{\varepsilon_1} PT_i \pm \frac{1}{\sqrt{\varepsilon_1}} U_i\Big]^T \geq 0,$$

we obtain the following inequality:

$$Q_c \geq PB\Big[(2\gamma - 1)I + \varepsilon_2 W - 4\gamma^2 \varepsilon_2 W\Big] B^T P + \frac{\varepsilon_3}{\delta} PDD^T P + \frac{1}{\varepsilon_3 \delta} C^T C + Q$$
$$= PB\Big[(2\gamma - 1)(I - \varepsilon_2(2\gamma + 1)W)\Big] B^T P + \frac{\varepsilon_3}{\delta} PDD^T P + \frac{1}{\varepsilon_3 \delta} C^T C + Q.$$

If γ satisfies either inequality in (21), which is equivalent to satisfying the inequality

$$(2\gamma - 1)(I - \varepsilon_2(2\gamma + 1)W) \geq 0,$$

then, $Q_c \geq \frac{\varepsilon_3}{\delta} PDD^T P + \frac{1}{\varepsilon_3 \delta} C^T C + Q > 0$. Thus, based on Lyapunov stability theory [1], the obtained controller $u(t)$ stabilizes the mismatched system in (14) with the constraints in (18).

The proof for disturbance attenuation is similar to that in Theorem 1 and hence omitted. ■

Remark 4. The introduction of tuning parameters, ε_1, ε_2, and ε_3 in (16) and (20), makes the proposed approach more flexible in obtaining disturbance-attenuation robust-stabilizing controllers. For instance, assuming that (A, C) is observable, the following Riccati equation

$$A^T P + PA - P\Big[BB^T - \frac{1}{\delta^2} DD^T\Big] P + C^T C = 0, \tag{22}$$

which is the standard Riccati equation for H_∞ control problem in [3] (i.e., if there exists a $P > 0$ satisfying (22), then $u(t) = -\frac{1}{2} B^T P x(t)$ is the desired disturbance-attenuation controller), corresponds to a special case of (16) or (20) (when $\Delta A = 0$ and $\Delta B = 0$) with $\varepsilon_3 = \frac{1}{\delta}$ and $Q = 0$. Also, it should be noted that the inequality in (21) gives an explicit bound for which the control gain is allowed to vary without affecting robust stability and disturbance attenuation of the closed-loop system. ■

5. Illustrative Examples

Example 1. Consider a version of the pitch-axis model for the AFTI/F-16 flying at 3000 ft. and Mach 0.6 [5,12,14]. The equations of motion are represented in the state-space form as

$$\dot{x}(t) = [A + \Delta A]x(t) + [B + \Delta B]u(t) + Dw(t),$$
$$y(t) = Cx(t),$$

where the nominal system are described by

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -0.87 & 43.22 \\ 0 & 0.99 & -1.34 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 0 \\ -17.25 & -1.58 \\ -0.17 & -0.25 \end{bmatrix},$$

$$D = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad C = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix},$$

and the structured uncertainty matrices are described by

$$\Delta A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & a_1 & a_2 \\ 0 & a_3 & a_4 \end{bmatrix}, \qquad \Delta B = \begin{bmatrix} 0 & 0 \\ b_1 & b_2 \\ b_3 & b_4 \end{bmatrix}$$

with $|a_1| \leq 0.7$, $|a_2| \leq 35$, $|a_3| \leq 0.7$, $|a_4| \leq 1.05$, $|b_1| \leq 2$, $|b_2| \leq 0.2$, $|b_3| \leq 0.02$, and $|b_4| \leq 0.03$.

Note that this system is matched and the structured uncertainty matrices can be expressed as $\Delta A = BE$ and $\Delta B = BF$, where

$$E = \begin{bmatrix} 0 & -0.0618a_1 + 0.3907a_3 & -0.0618a_2 + 0.3907a_4 \\ 0 & 0.0420a_1 - 4.2657a_3 & 0.0420a_2 - 4.2657a_4 \end{bmatrix}$$

and

$$F = \begin{bmatrix} -0.0618b_1 + 0.3907b_3 & -0.0618b_2 + 0.3907b_4 \\ 0.0420b_1 - 4.2657b_3 & 0.0420b_2 - 4.2657b_4 \end{bmatrix},$$

and the disturbance matrix can be written as $D = BG$ with

$$G = \begin{bmatrix} -0.0618 & 0.3907 \\ 0.0420 & -4.2657 \end{bmatrix}.$$

The eigenvalues of A are -7.65, 0, 5.44 and the nominal system is unstable. To find a disturbance-attenuation robust-stabilizing control law for this matched uncertain system, we determine T, U, and V as in (12) and obtain

$$T = \begin{bmatrix} 1.8874 & -1.9219 \\ -1.9219 & 8.2777 \end{bmatrix}, \qquad U = \text{diag}[0, 3.0508, 7.1143],$$

and

$$V = \begin{bmatrix} 0.17472 & -0.04797 \\ -0.04797 & 0.20393 \end{bmatrix}.$$

Set the disturbance-attenuation constant $\delta = 1$ and choose $Q = I$, $\varepsilon_1 = 0.04 \in (0, 0.086)$, and $\varepsilon_2 = 0.01 \in (0, 0.022)$. The Riccati equation in (13) has a symmetric positive-definite solution

$$P = \begin{bmatrix} 122.72 & 0.8920 & 3.1551 \\ 0.8920 & 0.5816 & -0.0804 \\ 3.1551 & -0.0804 & 54.211 \end{bmatrix}.$$

Then, from Theorem 2, a disturbance-attenuation robust-stabilizing control law with $\delta = 1$ can be constructed as $u(t) = Kx(t)$, where

$$K = -\gamma B^T P = \gamma \begin{bmatrix} 15.924 & 10.019 & 7.8291 \\ 2.1982 & 0.8988 & 13.426 \end{bmatrix}$$

with $\gamma \geq \frac{1}{2}$. Furthermore, the state-feedback control law $u(t) = -\gamma B^T P x(t)$ with $\gamma \geq \dfrac{1 - \sigma_{\min}(V)}{1 - 2\sigma_{\max}(V)} = 1.652$ is optimal with respect to a certain quadratic performance index.

To guarantee that the closed-loop system has a prescribed degree of stability $h = 1$, we set δ, Q, ε_1, ε_2 as before and replace A by $A + I$ to solve the Riccati equation in (13) for P. Then, a disturbance-attenuation robust-stabilizing control law with $\delta = 1$, which guarantees that the state vector decays no slower than e^{-t}, can be constructed as $u(t) = Kx(t)$, where

$$K = -\gamma B^T P = \gamma \begin{bmatrix} 33.018 & 10.236 & 4.7750 \\ -6.4293 & 0.8007 & 20.907 \end{bmatrix}$$

with $\gamma \geq \frac{1}{2}$.

When the requirement of disturbance attenuation is relaxed, i.e. $\delta \to \infty$, a robust stabilizing control law $u(t) = Kx(t) = -\gamma B^T P x(t)$ for the matched system is determined

by solving the Riccati equation in (13) for P with $Q = I$ and $\varepsilon_1 = 0.04$ as before. The feedback gain is given by

$$K = -\gamma B^T P = \gamma \begin{bmatrix} 5.6870 & 6.6475 & 10.092 \\ -0.1324 & 0.7230 & 3.2596 \end{bmatrix}$$

with $\gamma \geq \frac{1}{2}$. This control law is of the same order of magnitude as the control laws obtained in [5,12], for the same example.

Example 2. The dynamics of a helicopter in a vertical plane for an airspeed range of 60–170 knots are given in [8,12]. There are four state variables — x_1 = horizontal velocity (knot/sec), x_2 = vertical velocity (knot/sec), x_3 = pitch rate (deg/sec), and x_4 = pitch angle (deg) — and two control variables — u_1 = collective pitch control and u_2 = longitudinal cyclic pitch control. In the airspeed range of 60 knots to 170 knots, significant changes occur only in element a_{32}, a_{34}, and b_{21}. For this range of operating conditions,

$$A = \begin{bmatrix} -0.0366 & 0.0271 & 0.0188 & -0.4555 \\ 0.0482 & -1.01 & 0.0024 & -4.0208 \\ 0.1002 & 0.2855 & -0.707 & 1.3229 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0.4422 & 0.1761 \\ 3.0447 & -7.5922 \\ -5.52 & 4.99 \\ 0 & 0 \end{bmatrix},$$

$$D = [0,0,0,1]^T, \qquad C = [0,1,0,0],$$

$$\Delta A = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & r_{32} & 0 & r_{34} \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad \Delta B = \begin{bmatrix} 0 & 0 \\ s_{21} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

with $|r_{32}| \leq 0.2192$, $|r_{34}| \leq 1.2031$, and $|s_{21}| \leq 2.0673$. Define T, U, V, and W as in (19) and obtain

$$T = \text{diag}[0,0,1.4223,0], \qquad U = \text{diag}[0,0.2192,0,1.2031],$$

$$V = \text{diag}[1.03365,0], \qquad W = \text{diag}[0,1.03365,0,0].$$

Set the disturbance-attenuation constant $\delta = 0.5$ and choose $Q = I$, $\varepsilon_1 = 1$, $\varepsilon_2 = 0.25$ and $\varepsilon_3 = 0.25$, the Riccati equation in (20) has a symmetric positive-definite solution

$$P = \begin{bmatrix} 9.9891 & -0.6427 & -1.2810 & -11.2650 \\ -0.6427 & 1.0287 & 0.8892 & 2.0922 \\ -1.2810 & 0.8892 & 1.2521 & 3.4268 \\ -11.2650 & 2.0922 & 3.4268 & 19.4367 \end{bmatrix}.$$

Then, from Theorem 4, a disturbance-attenuation robust-stabilizing controller can be constructed as $u(t) = Kx(t) = -\gamma B^T P x(t)$, where

$$K = -\gamma B^T P = \gamma \begin{bmatrix} -9.5318 & 2.0603 & 4.7707 & 17.5269 \\ -0.2459 & 3.4864 & 0.7284 & 0.7682 \end{bmatrix}$$

with $\dfrac{1}{2\varepsilon_2 \max(v_i)} - \dfrac{1}{2} = 1.2093 \geq \gamma \geq \dfrac{1}{2}$.

To compare our results with that in [3], we let $\Delta A = 0$ and $\Delta B = 0$ (i.e. $T = 0$, $U = 0$, $V = 0$, and $W = 0$), and set the disturbance-attenuation constant $\delta = 0.1$. The Riccati equation in (22) which is now identical to (20) with $Q = 0$ and $\varepsilon_3 = \dfrac{1}{\delta} = 10$ does not have a symmetric positive-definite solution; however, with $Q = 0$ and $\varepsilon_3 = 0.25$, the Riccati equation in (20) has a symmetric positive-definite solution, and the desired state-feedback control gain is given by

$$K = \gamma \begin{bmatrix} -0.0033 & -2.1201 & 0.2444 & 0.4382 \\ 0.0063 & 5.8232 & 0.0755 & -0.3804 \end{bmatrix} \quad \text{for} \quad \gamma \geq \frac{1}{2}.$$

Thus, the developed method is more flexible than that of [3].

6. Conclusion

Based on the LQR theory and Lyapunov stability theory, new disturbance-attenuation robust-stabilizing controllers have been developed for matched and/or mismatched uncertain linear systems. It has been shown that dynamic systems, described by second-order vector differential equations, often satisfy the matching conditions and that disturbance-attenuation robust-stabilizing controllers (with optimal performance if $\|\Delta B\| < \dfrac{1}{2}$) always exist for matched uncertain linear systems which contain structured or norm-bounded uncertainty matrices. For mismatched uncertain linear systems , two theorems have been developed for finding disturbance-attenuation robust-stabilizing controllers. These disturbance-attenuation robust-stabilizing control laws can be easily constructed from the symmetric positive-definite solution of the augmented Riccati equation. Also, the proposed approach is more flexible than some existing methods in the sense that additional tuning parameters (such as ε, γ, and h etc.) have been introduced in the derivations to achieve robust stabilization, robust performance, and disturbance attenuation for uncertain linear systems. Two practical examples have been presented to illustrate the results.

Appendix

Lemma A.1 (Singular value decomposition [13].) Let $M \in \mathcal{R}^{n\times m}$ be any real matrix. Then there exist unitary matrices $U_n = [u_1, u_2, \ldots, u_n] \in \mathcal{R}^{n\times n}$ $(u_i^T u_j = \delta_{i,j})$ and $V_m = [v_1, v_2, \ldots, v_m] \in \mathcal{R}^{m\times m}$ $(v_i^T v_j = \delta_{i,j})$ such that

$$M = U_n \Sigma V_m^T, \tag{A1a}$$

where $\Sigma \in \mathcal{R}^{n\times m}$ is defined as

$$\Sigma = \begin{bmatrix} \Sigma_k & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{with} \qquad \Sigma_k = \text{diag}[\sigma_1, \sigma_2, \ldots, \sigma_k], \tag{A1b}$$

where $k \leq \min(n, m)$ is the rank of the matrix M and $\sigma_1, \sigma_2, \ldots, \sigma_k$ are the nonzero singular values of M. Furthermore, the matrix M can be written as

$$M = \sum_{i=1}^{k} \sigma_i u_i v_i^T = U_k \Sigma_k V_k^T, \tag{A1c}$$

where $U_k = [u_1, u_2, \ldots, u_k] \in \mathcal{R}^{n\times k}$ $(U_k^T U_k = I)$ and $V_k = [v_1, v_2, \ldots, v_k] \in \mathcal{R}^{m\times k}$ $(V_k^T V_k = I)$. ■

Consider the following matched uncertain system

$$\begin{aligned} \dot{x}(t) &= [A + \Delta A]x(t) + [B + \Delta B]u(t) + Dw(t) \\ &= [A + BE]x(t) + [B + BF]u(t) + BGw(t). \end{aligned} \tag{A2}$$

By utilizing the singular-value decomposition technique, the nominal input matrix B with full rank m can be decomposed as

$$B = U_m \Sigma_m V_m^T, \tag{A3a}$$

where $U_m \in \mathcal{R}^{n\times m}$, $\Sigma_m \in \mathcal{R}^{m\times m}$, and $V_m \in \mathcal{R}^{m\times m}$, are defined as in Lemma A.1. It is easy to see that

$$U_m^T \Delta A = U_m^T BE = \Sigma_m V_m^T E, \tag{A3b}$$

$$U_m^T \Delta B = U_m^T BF = \Sigma_m V_m^T F, \tag{A3c}$$

and

$$U_m^T D = U_m^T B G = \Sigma_m V_m^T G. \qquad (A3d)$$

Hence, if the uncertain system satisfies the matching conditons (i–iii), then we can determine E, F, and G from ΔA, ΔB, and D by

$$E = T_m \Delta A, \qquad F = T_m \Delta B, \qquad \text{and} \qquad G = T_m D, \qquad (A4a)$$

where

$$T_m = V_m \Sigma_m^{-1} U_m^T. \qquad (A4b)$$

Consider a real $n \times m$ matrix M of rank k. Immediately from Lemma A.1, the matrix M can be decomposed as the product of two rank-k matrices as follows:

$$M = M_u M_v^T, \qquad (A5a)$$

with

$$M_u = U_k \Sigma_k^{1/2} \qquad \text{and} \qquad M_v = V_k \Sigma_k^{1/2}, \qquad (A5b)$$

where $U_k \in \mathcal{R}^{n \times k}$, $\Sigma_k \in \mathcal{R}^{k \times k}$, and $V_k \in \mathcal{R}^{m \times k}$, are defined as in Lemma A.1.

References

[1] B.D.O. Anderson and J.B. Moore, *Linear Optimal Control* (Prentice-Hall, Englewood Cliffs, New Jersey, 1990).

[2] D.S. Bernstein and W. Haddad, LQG control with an H_∞ performance bound: A Riccati equation approach, *IEEE Transactions on Automatic Control* **34** (1989) 293–305.

[3] J.C. Doyle, K. Glover, P.P. Khargonekar, and B. Francis, State-space solutions to standard H_2 and H_∞ control problems, *IEEE Transactions on Automatic Control* **34** (1989) 831–847.

[4] K. Glover and J.C. Doyle, State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity, *Systems and Control Letters* **11** (1988) 167–172.

[5] F. Jabbari and W.E. Schmitendorf, A non-iterative method for design of linear robust controllers, *Proc. Conf. Decision & Control*, Tampa, Florida (December 1989) 1690–1692.

[6] P.P. Khargonekar, I.R. Petersen, and K. Zhou, Robust stabilization of uncertain linear systems: Quadratic stabilizability and H^∞ control theory, *IEEE Transactions on Automatic Control* **35** (1990) 356–361.

[7] G. Leitmann, Guaranteed asymptotic stability for some linear systems with bounded uncertainties, *Journal of Dynamic Systems, Measurement and Control* **101** (1979) 212–216.

[8] K.S. Narendra and S.S. Tripathi, Identification and optimization of aircraft dynamics, *Journal of Aircraft* **10** (1973) 193–199.

[9] I.R. Petersen and C.V. Hollot, A Riccati equation approach to the stabilization of uncertain linear systems, *Automatica* **22** (1986) 397–411.

[10] I.R. Petersen, A stabilization algorithm for a class of uncertain linear systems, *Systems and Control Letters* **8** (1987) 351–357.

[11] I.R. Petersen, Disturbance attenuation and H^∞ optimization: A design method based on the algebraic Riccati equation, *IEEE Transactions on Automatic Control*, **32** (1987) 427–429.

[12] W.E. Schmitendorf, A design methodology for robust stabilizing controllers, *AIAA Journal of Guidance, Control and Dynamics* **10** (1987) 250–254.

[13] R.E. Skelton, *Dynamic Systems Control* (John Wiley & Sons, New York, 1988).

[14] K.M. Sobel and E.Y. Shapiro, A design methodology for pitch pointing flight control systems, Journal of Guidance, Control, and Dynamics **8** (1985) 181–187.

[15] R.J. Veillette, J.V. Medanic, and W.R. Perkins, Robust stabilization and disturbance rejection for systems with structured uncertainty, *Proc. Conf. Decision & Control*, Tampa, Florida (December 1989) 936–941.

MINIMAX LINEAR SPLINES

Royce W. Soanes
U.S. Army Armament Research, Development, and Engineering Center
Close Combat Armaments Center
Benet Laboratories
Watervliet, NY 12189-4050

ABSTRACT. An algorithm is derived for obtaining a mesh that minimizes the maximum local interpolatory error for a linear spline, where the error is measured in any classical Banach norm. This algorithm is based on the standard method of approximate error equidistribution advocated by C. de Boor.

INTRODUCTION. In order to enable an industrial machine with primitive computational ability to use complicated or difficult to compute functional relationships repeatedly, efficiently, and accurately, it is necessary to supply the machine with these functional relationships as sets of data in tabular form. It is assumed that the machine can deal with continuous, piecewise linear functions (linear splines). A graphics tube is a good example. Such a tube can draw only straight lines, but drawing many short, connected line segments can represent an arbitrary curve well. In order to represent these functions most accurately, a nonuniform mesh must be used. Finding such a mesh is, in principle, a very difficult nonlinear optimization problem, but C. de Boor [1-3] advocated a general method by which the mesh can be found quickly, easily, robustly (and approximately) without any recourse to optimization methods! We present herein a robust addition to de Boor's standard method which improves its accuracy without increasing the essential complexity of his algorithm.

INTERPOLATORY ERROR. Let ℓ be the linear interpolant of function f on a subinterval of length h. The error is given by

$$f(t) = \ell(t) + e(t) \quad (\mu - \frac{h}{2} \leq t \leq \mu + \frac{h}{2})$$

Expand e in a Taylor series around the midpoint (μ) of the subinterval

$$e(t) = \sum_{i=0}^{\infty} \frac{e^{(i)}(\mu)}{i!} (t-\mu)^i$$

Applying the two boundary conditions

$$e(\mu - \frac{h}{2}) = 0 = e(\mu + \frac{h}{2})$$

ultimately yields

$$e(\frac{ht}{2} + \mu) = \sum_{i=1}^{\infty} \frac{f^{(2i)}(\mu)}{(2i)!} (\frac{h}{2})^{2i}(t^{2i}-1)$$

$$+ \sum_{i=1}^{\infty} \frac{f^{(2i+1)}(\mu)}{(2i+1)!} (\frac{h}{2})^{2i+1}t(t^{2i}-1)$$

Taking the first two terms of each sum

$$e(\frac{ht}{2} + \mu) = \frac{f''(\mu)}{2^3} h^2(t^2-1)$$

$$+ \frac{f^{(3)}(\mu)}{2^4 \cdot 3} h^3t(t^2-1)$$

$$+ \frac{f^{(4)}(\mu)}{2^7 \cdot 3} h^4(t^4-1)$$

$$+ \frac{f^{(5)}(\mu)}{2^8 \cdot 3 \cdot 5} h^5t(t^4-1) + O(h^6)$$

Letting

$$\rho_i = f^{(2+i)}(\mu)/f''(\mu)$$

one has

$$e(\frac{ht}{2} + \mu) = \frac{f''(\mu)}{2^3} h^2(t^2-1)\{1 + \frac{\rho_1}{2 \cdot 3} ht + \frac{\rho_2}{2^4 \cdot 3} h^2(t^2+1)$$

$$+ \frac{\rho_3}{2^5 \cdot 3 \cdot 5} h^3t(t^2+1) + O(h^4)\} = \frac{f''(\mu)}{2^3} h^2(t^2-1)(1+S)$$

where

$$S = a_1ht + a_2h^2(t^2+1) + a_3h^3t(t^2+1) + O(h^4)$$

and

$$a_1 = \frac{\rho_1}{2 \cdot 3} \quad , \quad a_2 = \frac{\rho_2}{2^4 \cdot 3} \quad , \quad a_3 = \frac{\rho_3}{2^5 \cdot 3 \cdot 5}$$

INTERPOLATORY ERROR NORM. The local L^n norm of the error on a subinterval of length h is defined by

$$\|e\|_{n,h} = (\int_{\mu-h/2}^{\mu+h/2} |e(t)|^n dt)^{1/n}$$

where $1 \leq n < \infty$ and n is an integer. For $n = \infty$, we have the maximum error. Now,

$$\|e\|^n_{n,h} = \frac{h}{2}\int_{-1}^{1} |e(\frac{ht}{2} + \mu)|^n dt$$

but

$$e(\frac{ht}{2} + \mu) = \frac{f''(\mu)}{2^3} h^2(t^2-1)(1+S)$$

So if we let h be sufficiently small so that $|S| < 1$ on $(-1,1)$, we have

$$\|e\|^n_{n,h} = \frac{|f''(\mu)|^n h^{2n+1}}{2^{3n+1}} \int_{-1}^{1} (1-t^2)^n(1+S)^n dt$$

Since only the even terms of $(1+S)^n$ contribute to the integral, we have

$$\|e\|^n_{n,h} = \frac{|f''(\mu)|^n h^{2n+1}}{2^{3n}} \int_{0}^{1} (1-t^2)^n Ev(1+S)^n dt$$

where $Ev(1+S)^n$ denotes the even terms of $(1+S)^n$.

Hence,

$$Ev(1+S)^n = 1 + \binom{n}{1}a_2h^2(1+t^2) + \binom{n}{2}a_1^2h^2t^2 + O(h^4)$$

Letting

$$I_{n,i} = \int_0^1 (1-t^2)^n t^{2i} dt$$

we therefore have

$$\int_0^1 (1-t^2)^n Ev(1+S)^n dt$$

$$= \int_0^1 (1-t^2)^n(1+h^2(na_2(1+t^2) + \frac{n(n-1)}{2} a_1^2t^2) + O(h^4))dt$$

$$= I_{n,o} + nh^2(a_2(I_{n,o}+I_{n,1}) + \frac{n-1}{2} a_1^2 I_{n,1}) + O(h^4)$$

$$= I_{n,o}(1+nh^2(a_2(1 + \frac{I_{n,1}}{I_{n,o}}) + \frac{n-1}{2} a_1^2 \frac{I_{n,1}}{I_{n,o}}) + O(h^4))$$

Using integration-by-parts on $I_{n,i}$ and solving the resulting recursion ultimately yields

$$I_{n,i} = \frac{2^{2n}n!(2i)!(i+n)!}{i!(2i+2n+1)!}$$

from which we conclude that

$$I_{n,o} = \frac{2^{2n}(n!)^2}{(2n+1)!}$$

$$I_{n,1} = \frac{2^{2n+1}n!(n+1)!}{(2n+3)!}$$

and

$$\frac{I_{n,1}}{I_{n,o}} = \frac{1}{2n+3}$$

Hence,

$$\int_0^1 (1-t^2)^n Ev(1+S)^n dt$$

$$= \frac{2^{2n}(n!)^2}{(2n+1)!}\left(1+nh^2\left(\frac{2(n+1)}{2n+3}a_2 + \frac{n-1}{2(2n+3)}a_1^2\right) + O(h^4)\right)$$

and

$$\|e\|_{n,h}^n = \frac{|f''(\mu)|^n h^{2n+1}(n!)^2}{2^n(2n+1)!}\left(1+nh^2\left(\frac{2(n+2)}{2n+3}a_2 + \frac{n-1}{2(2n+3)}a_1^2\right) + O(h^4)\right)$$

or

$$\|e\|_{n,h} = k\,|f''(\mu)|\,h^{2+1/n}\left(1+h^2\left(\frac{2(n+2)}{2n+3}a_2 + \frac{n-1}{2(2n+3)}a_1^2\right) + O(h^4)\right)$$

where

$$k = \frac{1}{2}\left(\frac{(n!)^2}{(2n+1)!}\right)^{1/n}$$

Using Stirling's approximation to the factorial, it is easy to show that

$$\lim_{n\to\infty} k = \frac{1}{8}$$

Recalling that

$$a_1 = \frac{\rho_1}{2\cdot 3} \quad \text{and} \quad a_2 = \frac{\rho_2}{2^4\cdot 3}$$

we finally have

$$\|e\|_{n,h} = k\,|f''(\mu)|\,h^{2+1/n}\left(1 + \frac{h^2}{24}\left(\frac{n+2}{2n+3}\rho_2 + \frac{n-1}{3(2n+3)}\rho_1^2\right) + O(h^4)\right)$$

as $h\to 0$, where

$$k = \frac{1}{2}\left(\frac{(n!)^2}{(2n+1)!}\right)^{1/n}$$

and

$$\rho_i = f^{(2+i)}(\mu)/f''(\mu)$$

NORM OF ARBITRARY FUNCTION. The local L^p norm of arbitrary function ϕ over a subinterval of length h is defined as

$$\|\phi\|_{p,h} = \left(\int_{\mu-h/2}^{\mu+h/2} |\phi(t)|^p dt\right)^{1/p}$$

where $p > 0$, finite and real. In this context, we allow $p < 1$ even though Minkowski's triangle inequality holds only for $p \geq 1$.

Expand ϕ in a Taylor series around the midpoint of the subinterval

$$\phi(t) = \phi(\mu) \sum_{i=0}^{\infty} \frac{\rho_i}{i!} (t-\mu)^i$$

where

$$\rho_i = \phi^{(i)}(\mu)/\phi(\mu)$$

Now,

$$\|\phi\|_{p,h}^p = \frac{h}{2} \int_{-1}^{1} |\phi(\frac{ht}{2} + \mu)|^p dt$$

but

$$\phi(\frac{ht}{2} + \mu) = \phi(\mu)(1+S)$$

where

$$S = \sum_{i=1}^{\infty} a_i t^i$$

and

$$a_i = \frac{\rho_i}{i!} \left(\frac{h}{2}\right)^i$$

Hence, letting h be sufficiently small so that $|S| < 1$ on $(-1,1)$, we have

$$\|\phi\|^p_{p,h} = \frac{h}{2} \mid \phi(\mu) \mid^p \int_{-1}^{1} (1+S)^p dt = \frac{h}{2} \mid \phi(\mu) \mid^p \int_{-1}^{1} Ev(1+S)^p dt$$

but

$$S = a_1 t + a_2 t^2 + a_3 t^3 + O(h^4)$$

hence

$$Ev(1+S)^p = 1 + \binom{p}{1} a_2 t^2 + \binom{p}{2} a_1^2 t^2 + O(h^4)$$

We therefore have

$$\|\phi\|^p_{p,h} = h \mid \phi(\mu) \mid^p \int_{0}^{1} 1 + pt^2(a_2 + \frac{p-1}{2} a_1^2) + O(h^4) dt$$

$$= h \mid \phi(\mu) \mid^p (1 + \frac{ph^2}{24} (\rho_2 + (p-1)\rho_1^2) + O(h^4))$$

or

$$\|\phi\|_{p,h} = h^{1/p} \mid \phi(\mu) \mid (1 + \frac{h^2}{24} (\rho_2 + (p-1)\rho_1^2) + O(h^4))$$

as $h \to 0$.

STANDARD APPROXIMATION TO $\|e\|_{n,h}$. Recalling that

$$h^{-1/p}\|f''\|_{p,h} = \mid f''(\mu) \mid (1 + \frac{h^2}{24}(\rho_2 + (p-1)\rho_1^2) + O(h^4))$$

and

$$\|e\|_{n,h} = kh^{2+1/n} \mid f''(\mu) \mid (1 + \frac{h^2}{24}(\frac{n+2}{2n+3} \rho_2 + \frac{n-1}{3(2n+3)} \rho_1^2) + O(h^4))$$

we multiply the first equation by $kh^{2+1/n}$ and subtract from the second, getting

$$\|e\|_{n,h} = kh^{2+1/n-1/p}\|f''\|_{p,h}$$

$$+ kh^{2+1/n} \mid f''(\mu) \mid (\frac{h^2}{24}(- \frac{n+1}{2n+3} \rho_2 + (\frac{7n+8}{6n+9} - p)\rho_1^2) + O(h^4))$$

If we now let $p = \frac{n}{2n+1}$, we have

$$\|e\|_{n,h} = k\|f''\|_{n/(2n+1),h}$$

$$+ kh^{4+1/n} \mid f''(\mu) \mid (\frac{1}{24}(- \frac{n+1}{2n+3} \rho_2 + \frac{8n^2+14n+8}{12n^2+24n+9} \rho_1^2) + O(h^2))$$

$$= k\|f''\|_{p,h} + kh^{4+1/n} \mid f''(\mu) \mid (\frac{1}{24}(- a\rho_2 + b\rho_1^2) + O(h^2))$$

$$= k\|f''\|_{n/(2n+1),h} + O(h^{4+1/n})$$

For n = 1,2, and ∞, respectively, we have

$$\|e\|_{1,h} = \frac{1}{12} \|f''\|_{1/3,h} + O(h^3)$$

$$\|e\|_{2,h} = \frac{1}{2\sqrt{30}} \|f''\|_{2/5,h} + O(h^{9/2})$$

$$\|e\|_{\infty,h} = \frac{1}{8} \|f''\|_{1/2,h} + O(h^4)$$

STANDARD ERROR EQUIDISTRIBUTION FOR ANY BANACH NORM. In this section, we justify the standard method of error equidistribution with respect to any Banach norm. The global L^n norm of the error over interval (a,b) is

$$\|e\|_n = (\int_a^b |e(t)|^n dt)^{1/n}$$

Hence, for a mesh $a = x_1 < x_2 < \ldots < x_N = b$

$$\|e\|_n^n = \sum_{i=1}^{N-1} \int_{x_i}^{x_{i+1}} |e_i(t)|^n dt = \sum_{i=1}^{N-1} \|e\|_{n,h_i}^n$$

Let single bars around the error denote the standard approximation to the error norm and analogously define

$$|e|_n^n = \sum_{j=1}^{N-1} |e|_{n,h_j}^n$$

but

$$|e|_{n,h_j} = k\|f''\|_{p,h_j} = k(\int_{x_j}^{x_{j+1}} |f''(t)|^p dt)^{1/p}$$

where

$$p = \frac{n}{2n+1}$$

Hence, letting

$$I_{p,h_j} = \int_{x_j}^{x_{j+1}} |f''(t)|^p dt$$

we have

$$|e|_n^n = k^n \sum_{j=1}^{N-1} I_{p,h_j}^{2n+1}$$

We will refer to the integrals $I_{p,h}$ as the standard or de Boor integrals.

It follows trivially, using Leibnitz's rule, that

$$\frac{\partial}{\partial x_i} |e|_n^n = 0 \qquad 1 < i < N$$

implies that

$$I_{p,h_{i-1}} = I_{p,h_i} \qquad 1 < i < N$$

Hence, the condition $I_{p,h}$ = constant determines the mesh which minimizes the standard global approximation to $\|e\|_n$.

For a linear spline approximation to f", it is a fairly simple (see COMPUTATION) matter to find the mesh for which the de Boor integrals are constant.

CONVERGENCE OF STANDARD METHOD. Recall that

$$\|e\|_{n,h} = k\|f''\|_{p,h} + \frac{k}{24} h^{4+1/n} |f''(\mu)| (-a\rho_2 + b\rho_1^2) + O(h^{6+1/n})$$

Letting

$$F = -af''(\mu)f^{(4)}(\mu) + bf^{(3)}(\mu)^2$$

we have the following one term approximation to the difference between $\|e\|_{n,h}$ and $|e|_{n,h}$:

$$\|e\|_{n,h} - |e|_{n,h} \simeq \frac{kFh^{4+1/n}}{24\lceil f''(\mu)\rceil}$$

but

$$\|e\|_{n,h} \simeq kh^{2+1/n} |f''(\mu)|$$

Therefore, we also have

$$\frac{\|e\|_{n,h} - |e|_{n,h}}{\|e\|_{n,h}} \simeq \frac{Fh^2}{24\lceil f''(\mu)\rceil^2}$$

but also

$$\|f''\|_{p,h} \simeq h^{1/p} |f''(\mu)|$$

hence,

$$h \simeq \left(\frac{\|f''\|_{p,h}}{\lceil f''(\mu)\rceil}\right)^p = \left(\frac{I_{p,h}^{1/p}}{\lceil f''(\mu)\rceil}\right)^p = \frac{I_{p,h}}{|f''(\mu)|^p}$$

In addition, for the correct mesh

$$I_{p,h} = \frac{I_p}{N-1}$$

hence,

$$h \simeq \frac{I_p}{(N-1)\,|f''(\mu)|^p}$$

and therefore,

$$\frac{\|e\|_{n,h} - |e|_{n,h}}{\|e\|_{n,h}} \simeq \frac{FI_p^2}{24\,|f''(\mu)|^{2+2p}(N-1)^2}$$

This tells us that the relative difference between $\|e\|_{n,h}$ and $|e|_{n,h}$ is $O(\frac{1}{N^2})$ as $N\to\infty$, which means that the standard method works better and better ($\|e\|_{n,h}$ will be more nearly constant) as N gets larger and larger. This is all true, however, with the proviso that

$$\frac{F}{|f''(\mu)|^{2+2p}}$$

is bounded throughout the region of interest. It stands to reason, therefore, that the standard method will perform worst where f" is not bounded away from zero.

IMPROVED APPROXIMATION TO $\|e\|_{n,h}$. Recall that

$$\|e\|_{n,h} = k|f''(\mu)|h^{2+1/n}(1 + \frac{h^2}{24}(\frac{n+2}{2n+3}\rho_2 + \frac{n-1}{3(2n+3)}\rho_1^2) + O(h^4))$$

and

$$\|f''\|_{q,h} = h^{1/q}|f''(\mu)|(1 + \frac{h^2}{24}(\rho_2 + (q-1)\rho_1^2) + O(h^4))$$

Multiplying h by r in the second equation, we have

$$r^{-1/q}\|f''\|_{q,rh} = h^{1/q}|f''(\mu)|(1 + \frac{h^2}{24}(r^2\rho_2 + r^2(q-1)\rho_1^2) + O(h^4))$$

Multiplying this equation by kh^Q gives us

$$kr^{-1/q}h^Q\|f''\|_{q,rh} = kh^{Q+1/q}|f''(\mu)|(1 + \frac{h^2}{24}(r^2\rho_2 + r^2(q-1)\rho_1^2) + O(h^4))$$

Now, in order to make this equation look as much like the very first one as possible, we set

$$r^2 = \frac{n+2}{2n+3} \quad , \quad r^2(q-1) = \frac{n-1}{3(2n+3)}$$

and

$$Q + \frac{1}{q} = 2 + \frac{1}{n}$$

Solving for r, q, and Q, we have

$$r = \left(\frac{n+2}{2n+3}\right)^{1/2}$$

$$q = \frac{4n+5}{3n+6}$$

and

$$Q = \frac{5n^2+8n+5}{4n^2+5n}$$

A simple subtraction then gives us an improved approximation to $\|e\|_{n,h}$

$$\|e\|_{n,h} = kr^{-1/q}h^{Q}\|f''\|_{q,rh} + O(h^{6+1/n})$$

where before, we had

$$\|e\|_{n,h} = k\|f''\|_{p,h} + O(h^{4+1/n}) = |e|_{n,h} + O(h^{4+1/n})$$

It must be mentioned however, that although this improved approximation is asymptotically more efficient, no such approximation can be uniformly superior in all cases. Bearing this in mind, we dispense with approximations on all subintervals not having f" bounded away from zero and instead use the exact error

$$e_i(x) = \int_{x_i}^{x}\int_{x_i}^{t} f''(u)dudt - \frac{x-x_i}{x_{i+1}-x_i}\int_{x_i}^{x_{i+1}}\int_{x_i}^{t} f''(u)dudt$$

COMPUTATION. In actual computation, we assume the existence of a piecewise linear approximation to $|f''|$. The mesh over which this function is defined is referred to as the "original" mesh. In order to deal with the standard and improved asymptotic integral approximations to the local error norm, we will need to deal with integrals of the form

$$L = \int_{c}^{c+\ell} \lambda(t)^{m/n}dt$$

where λ is a nonnegative linear function with slope s

$$\lambda(t) = \lambda(c) + s(t-c)$$

with

$$\lambda(t) \geq 0 \quad \text{for} \quad c \leq t \leq c + \ell$$

and where m and n are arbitrary positive integers.

In the following, let

$$\alpha = \lambda(c)^{1/n}$$

and

$$S_k = \sum_{i=0}^{k} \alpha^i \beta^{k-i} = \frac{\beta^{k+1} - \alpha^{k+1}}{\beta - \alpha}$$

First, we need to compute L as a function of ℓ

$$L = \frac{\ell n S_{m+n-1}}{(m+n) S_{n-1}} = A(\ell)$$

where

$$\beta = (\lambda(c) + s\ell)^{1/n}$$

Second, we need to compute ℓ as a function of L

$$\ell = \frac{L(m+n) S_{n-1}}{n S_{m+n-1}} = B(L)$$

where

$$\beta = (\lambda(c)^{m/n+1} + (\frac{m}{n} + 1)sL)^{1/(m+n)}$$

A and B are therefore inverse functions, i.e.,

$$A(B(x)) = x = B(A(x))$$

or

$$A^{-1} = B \quad \text{and} \quad B^{-1} = A$$

Now let values of u denote the original mesh and let g be the piecewise linear interpolant to the $(u_i, |f_i''|)$ data.

Define the integral

$$G(x) = \int_{u_1}^{x} g(t)^{m/n} dt$$

Now if $u_i \leq x \leq u_{i+1}$,

$$G(x) = \int_{u_1}^{u_i} g(t)^{m/n} dt + \int_{u_i}^{u_i + x - u_i} g_i(t)^{m/n} dt$$

$$= G(u_i) + L$$

where $\lambda = g_i$, $c = u_i$, and $\ell = x - u_i$.

Hence,

$$G(x) = G(u_i) + A(x-u_i)$$

explicitly defines G for all x in the domain of interest.

In order to get the standard mesh, we will also have to compute the inverse of G (only for m/n = p).

$$B(G(x) - G(u_i)) = B(A(x-u_i)) = x-u_i$$

Hence,

$$x = u_i + B(G(x) - G(u_i))$$

but if $G(x) = \gamma$, then $x = G^{-1}(\gamma)$. Therefore,

$$G^{-1}(\gamma) = u_i + B(\gamma - G(u_i))$$

for

$$G(u_i) \leq \gamma \leq G(u_{i+1})$$

and provided

$$G(u_i) \neq G(u_{i+1})$$

Define

$$I_i = G(x_{i+1}) - G(x_i) = \int_{x_i}^{x_{i+1}} g(t)^p dt$$

where x is the standard or improved mesh, obtained by prescribing values for the I's. The standard method prescribes

$$I_i = \text{const} = \frac{G(x_N)}{N-1} \quad (1 \leq i \leq N)$$

For the improved mesh, the I's will vary, but the mesh is still obtained in the standard way. Since

$$G(x_{i+1}) = G(x_i) + I_i$$

we have immediately that

$$x_{i+1} = G^{-1}(G(x_i) + I_i) \quad i = 1,2,\ldots,N-2$$

ALGORITHM. Let * denote a standard or improved mesh and ** denote the succeeding improved mesh. We have seen that the main contributor to the ratios $\|e\|_{n,h^{**}}/\|e\|_{n,h^{*}}$ and $|e|_{n,h^{**}}/|e|_{n,h^{*}}$ is

$$\left(\frac{h^{**}}{h^{*}}\right)^{2+1/n}\left|\frac{f''(\mu^{**})}{f''(\mu^{*})}\right|$$

We therefore have the approximate asymptotic relation

$$\frac{|e|_{n,h^{**}}}{|e|_{n,h^{*}}} \simeq \frac{\|e\|_{n,h^{**}}}{\|e\|_{n,h^{*}}}$$

But we would like $\|e\|_{n,h^{**}}$ to be constant, hence we have the proportionality

$$|e|_{n,h^{**}} \propto \frac{|e|_{n,h^{*}}}{\|e\|_{n,h^{*}}}$$

or

$$I_{p,h^{**}} \propto \left(\frac{|e|_{n,h^{*}}}{\|e\|_{n,h^{*}}}\right)^{p}$$

We calculate the I's accordingly and multiply them by the appropriate constant to get

$$\sum_{i=1}^{N-1} I_{p,h_i^{**}} = \frac{G(x_N)}{N-1}$$

The quantities $\|e\|_{n,h_i^{*}}$ are computed either from the improved asymptotic approximation or exactly (relative to the original data) depending on whether or not f" is bounded away from zero on the subinterval in question. It is important to note that this approximate relation between the * and ** meshes can lead to exact convergence (rapidly) to the minimax mesh. If the * mesh is the minimax mesh ($\|e\|_{n,h^{*}}$ = constant), then the de Boor integrals ($I_{p,h}$) on the ** mesh will be no different from those on the * mesh.

The practical convergence properties of this algorithm are as follows. If f" is well bounded away from zero, the standard de Boor method gives impeccable results without any iteration. If f" is not bounded away from zero, convergence to a virtually perfect minimax mesh can easily occur in only two iterations. A few iterations may be needed in the presence of multiple inflection points.

In any case, even the very first iteration improves the mesh markedly.

REFERENCES

1. C. de Boor, "Good Approximation by Splines With Variable Knots," in: Spline Functions and Approximation Theory (A. Meir and A. Sharma, eds.), Birkhäuser Verlag, Basel, 1973, pp. 57-72.

2. C. de Boor, "Good Approximation by Splines With Variable Knots, II," in: Numerical Solution of Differential Equations (G.A. Watson, ed.), Lecture Notes in Math, No. 363, Springer Verlag, 1974, pp. 12-20.

3. C. de Boor, A Practical Guide to Splines, Springer-Verlag, New York, 1978.

9th Annual Army Conference on Applied Mathematics Attendees

Name	Address	Phone	Email
George F. Adams	Director U.S. Army Ballistic Research Laboratory ATTN: George F. Adams Interior Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6197	qta@brl.mil
Adair R. Aguiar		623-9096	
Gerald R. Andersen	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4253	jerry@brl.mil
Bruce Anderson	Cornell University Mathematical Sciences Institute Department of Mathematics, White Hall Ithaca, NY 14853	607-272-6132	anderson@mssun7.msi.cornell.edu
A. Arvind	Massachusetts Institute of Technology Laboratory for Computer Science 545 Technology Square Cambridge, MA 02139		
Steven F. Ashby	Lawrence Livermore National Laboratory Computing & Mathematics Research Group Mail Station L-316 P.O. Box 808 Livermore, CA 94551	415-423-2462	ashby@lll-crg.llnl.gov
Donald Austin	Executive Director Army High Performance Computing Research Center 1100 Washington Avenue South Minneapolis, MN 55415	612-626-1550	austin@ahpcrc.umn.edu

Name	Address	Phone	Email
Harry Auvermann	Commander U.S. Army Atmospheric Sciences Lab. ATTN: SLCAS-AR-I (Dr. Harry Auvermann) White Sands Missile Range, NM 88002-5501	505-678-4224	
Marco Avellaneda	Courant Institute 251 Mercer Street New York, NY 10012	212-998-3141	avellane@math1.nyu.edu
Kenneth A. Bannister	Director U.S. Army Ballistic Research Laboratory ATTN: Dr. Kenneth A. Bannister Interior Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6121	kab@brl.mil
Romsesh C. Batra	Department of Mechanical and Aerospace Engineering and Engineering Mechanics University of Missouri-Rolla Rolla, MO 65401-0249	314-341-4589	c2980@umrvmb.bitnet
Michael J. Belczynski	Commander U.S. Army Tank-Automotive Command ATTN: AMSTA-RYA (Michael J. Belczynski) Warren, MI 48397-5000	313-574-7816	
Simeon M. Berman	New York University Warren Seaver Hall 251 Mercer Street New York, NY 10012	212-998-3001	
Dimitris Bertsimas	Alfred P. Sloan School of Management Massachusetts Institute of Technology 50 Memorial Drive, E53-359 Cambridge, MA 02139		
Adam W. Bojanczyk	School of Electrical Engineering Cornell University Phillips Hall Ithaca, NY 14853	607-255-4296	adamb@ee.cornell.edu
Daniel Boley	Department of Computer Science University of Minnesota Minneapolis, MN 55455	612-625-3887	boley@cs.umn.edu
Deborah Brandon	Carnegie Mellon University Mathematics Department Pittsburgh, PA 15213	412-268-2545	
Michael Brewer	Department of Mechanical Engineering Colorado State University Fort Collins, CO 80523	303-491-7479	
Roger Brockett	Division of Applied Sciences Harvard University Cambridge, MA 02138	617-495-3922	brockett@gramian.harvard.edu
Paul Broome	Director U.S. Army Ballistic Research Lab ATTN: SLCBR-SECAD (Dr. Paul Broome) Aberdeen Proving Ground, MD 21005-5066	301-278-6884	broome@brl.mil
Mel Brown	U. S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4336	brown@aro-emh1.army.mil
Jagdish Chandra	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4254	sso@aro-emh1.army.mil
Lang-Mann Chang	Director U.S. Army Ballistic Research Laboratory ATTN: SLCBR-IB-P (Dr. Lang-Mann Chang) Aberdeen Proving Ground, MD 21005-5066		

Name	Address	Phone	Email
Peter C.T. Chen	Research Mechanical Engineer Benet Laboratories ATTN: SMCAR-CCB-RA Watervliet, NY 12189-4050	518-266-5907	
Shih C. Chu	Commander U.S. Army Armament R&D Center ATTN: SMCAR-CCL-EM (Dr. Shih C. Chu) Light Armament Division, CCAC Picatinny Arsenal, NJ 07806-5000	201-724-7316	pica shihchu
Kenneth D. Clark	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4256	clark@adm.csc.ncsu.edu
Susan Coates	Director U.S. Army Ballistic Research Laboratory ATTN: Susan Coates Vulnerability/Lethality Division Aberdeen Proving Ground, MD 21005-5066	301-278-6710	scoates@brl.mil
Norman Coleman, Jr.	ARDEC ATTN: SMCAR-FSF-RC Picatinny Arsenal, NJ 07806-5000	201-724-6275	ncoleman@pica.army.mil
Terry Cronin	Director U.S. Army CECOM Center of Signals Warfare ATTN: AMSEL-RD-SW-TRI (Mr. Terry Cronin) Vint Hill Farms Station Warrenton, VA 22186-5100	703-349-6939	
Benjamin E. Cummings	Director U.S. Army Human Engineering Lab ATTN: Dr. Benjamin E. Cummings Aberdeen Proving Ground, MD 21005-5066	301-278-5811	bigben@brl.mil
Yuefan Deng	State University of New York at Stony Brook Applied Mathematics Department StonyBrook, NY 11794-3600	516-632-8614	
Keith Dennis	Cornell University Mathematical Sciences Institute Department of Mathematics, White Hall Ithaca, NY 14853	607-255-4027	dennis@mssun7.msi.cornell.edu
Max Donath	University of Minnesota Department of Mechanical Engineering 111 Church St. SE Minneapolis, MN 55455	612-625-2304	donath@vx.acs.umn.edu
Devdatt P. Dubhashi	Cornell University Mathematical Sciences Institute Department of Computer Science, White Hall Ithaca, NY 14853	607-255-9206	dubhashi@cs-cornell.edu
Godi Fischer	University of Rhode Island Kelly Hall Annex Room A-109 Kingston, RI 02881	401-792-5879	fischer@quahog.uri.edu
Donald French	Department of Mathematical Sciences University of Cincinnati Old Chemistry Building (ML 25) Cincinnati, OH 45221-0025	513-556-4039	french@ucunix.san.uc.edu
James Glimm	State University of New York at Stony Brook Department of Applied Mathematics and Statistics Stony Brook, NY 11794-3600	516-632-8355	glimm@ams.sunysb.edu
Aaron Das Gupta	Director U.S. Army Ballistic Research Laboratory ATTN: SLCBR-TB-B (Dr. Aaron Das Gupta) Aberdeen Proving Ground, MD 21005-5066	301-278-6026	dasgupta

Name	Address	Phone	Email
Morton E. Gurtin	Department of Mathematics Carnegie Mellon University Pittsburgh, PA 15213-3890	412-268-2545	
Harumi Hattori	West Virginia University Department of Mathematics Morgantown, WV 26506	304-293-2014	u1493@wvnvm.bitnet
Robert Heyman	U.S. Army Materials Technology Laboratory ATTN: SLCMT-MRD Arsenal Street Watertown, MA 02172-0001	617-923-5274	rheyman@watertown-emh1.army.mil
Jieh Hsiang	Department of Computer Science State University of New York Stony Brook, NY 11794-4400	516-632-8449	hsiang@sbcs.sunysb.edu
Bill Hrusa	Carnegie Mellon University Mathematics Department Pittsburgh, PA 15213	412-268-8487	
John S. Hurley	Hampton University Department of Mathematics Hampton, VA 23668-00101	804-727-5352	hurley@hurley.math.hamptonu.edu
E. F. Infante	Dean, Institute of Technology University of Minnesota 107 Walter Library Minneapolis, MN 55455	612-624-2006	infante@mailbox.mail.umn.edu
Richard James	University of Minnesota Department of Aerospace Engineering Minneapolis, MN 55455	612-625-0706	rdjames@umnacvx
Arthur Johnson	Commander U.S. Army Materials & Mechanics Research Center ATTN: AMXMR-SMM (Dr. Arthur Johnson) Watertown, MA 02172	617-923-5272	
Moon S. Jun	Physical Science Laboratory New Mexico State University Box 30002 Las Cruces, NM 88003-0002	505-522-9137	mjun@dante.nmsu.edu
Professor G. Kallianpur	Department of Statistics University of North Carolina-Chapel Hill Chapel Hill, NC 27514	919-962-2187	
R. L. Kashyap	Purdue University Department of Electrical Engineering W. Lafayette, IN 47907	317-494-3437	kashyap@ecn.purdue.edu
Kent D. Kimsey	Director U.S. Army Ballistic Research Laboratory ATTN: Kent D. Kimsey Terminal Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6083	kimsey@brl.mil
Robert V. Kohn	Courant Institute 251 Mercer Street New York, NY 10012	212-998-3217	kohn@math1.nyu.edu
Professor G. S. Ladde	Department of Mathematics University of Texas at Arlington Box 19408 Arlington, TX 76019	817-273-3261	
Steven P. Lalley	Department of Statistics Purdue University Mathematical Sciences Building West Lafayette, IN 47906	317-494-6036	
Professor E. B. Lee	University of Minnesota EE/CS Building, 200 Union Street Minneapolis, MN 55455	612-625-0125	eblee@ee.umn.edu

Name	Address	Phone	Email
M. Howard Lee	Department of Physics and Astronomy University of Georgia Athens, GA 30602	404-542-3539	
Franklin T. Luk	School of Electrical Engineering Cornell University Phillips Hall Ithaca, NY 14853	607-255-5075	luk@ee.cornell.edu
Mitchell Luskin	School of Mathematics University of Minnesota 127 Vincent Hall Minneapolis, MN 55455		
Paul Muzio	Support Director Army High Performance Computing Research Center 1100 Washington Avenue South Minneapolis, MN 55415	612-626-1550	muzio@ahpcrc.umn.edu
Madhura Nirkhe	University of Maryland Department of Computer Science A.V. Williams Building College Park, MD 20742	301-405-2716	madhura@cs.umd.edu
Jorge Nocedal	Department of Electrical Engineering and Computer Science Northwestern University Evanston, IL 60208	601-634-3987	
Michael J. Nusca	U.S. Army Ballistic Research Laboratory ATTN: SLCBR-LF-A Aberdeen Proving Ground, MD 21005 3909 Halls Ferry Road Vicksburg, MS 39180-6199	301-278-2057	nusca@brl.mil
Robert E. Olson	Commander U.S. Army Waterways Engineer Waterways Experiment Station ATTN: CEWES-EN (Robert E. Olson) Environmental Laboratory 3909 Halls Ferry Road Vicksburg, MS 39180-6199		
Dave Olwell	Department of Mathematics United Stated Military Academy West Point, NY 10096-1786		
James L. Overholt	Commander U.S. Army Tank-Automotive Command ATTN: AMSTA-RYA (James L. Overholt) Analytical and Physical Simulation Br. Warren, MI 48397-5000	313-574-8633	
Thomas J. Pence	Department of Metallurgy, Mechanics, and Materials Science Michigan State University East Lansing, MI 48824-1226	517-353-3889	pence@frith.eng.msu.edu
Shietung Peng	Department of Computer Science University of Maryland, Baltimore County Catonsville, MD 21228	301-455-3540	
Linda R. Petzold	Lawrence Livermore National Laboratory L-316 P.O. Box 808 Livermore, CA 94550	415-423-6571	petzold@lll.crg.llnl.gov
Olivier Pironneau	Institut National De Recherche En Informatique Et En Automatique Domaine de Voluceau - Rocquencourt B.P. 105 - 78153 Le Chesnay CEDEX France	011-331-39635483	pironneau@menusin.mraifr

Name	Address	Phone	Email
Alyson Reeves	Cornell University Mathematical Sciences Institute Department of Mathematics, White Hall Ithaca, NY 14853	607-564-9041	reeves@mssun7.cornell.msi.edu
J. R. Rice	Department of Computer Science Purdue University Computer Science Building West Lafayette, IN 47907	317-494-6003	rice@cs.purdue.edu
Jubaraj Sahu	Director U.S. Army Ballistic Research Laboratory ATTN: Jubaraj Sahu Launch and Flight Division Aberdeen Proving Ground, MD 21005-5066	301-278-3707	sahu@br-mil
S. Sathananthan	Division of Science and Mathematics Jarvis Christian College Hawkins, TX 75765	903-769-2174 ext. 304	
Robert B. Schnabel	Department of Computer Science University of Colorado at Boulder Boulder, CO 80309	303-492-7554	bobby@cs.colorado.edu
George R. Sell	Army High Performance Computing Research Center University of Minnesota Minnesota Tech Center 1100 Washington Avenue South Minneapolis, MN 55415	612-626-1550	sell@ahpcrc.umn.edu
Leang S. Shieh	Cullen College of Engineering Department of Electrical Engineering University of Houston University Park Houston, TX 77204-4793	713-749-4418	
Royce Soanes	Chief, Benet Laboratories ATTN: Mr. Royce Soanes Watervliet Arsenal Watervliet, NY 12189-4050	518-383-8203	
Janet Spoonamore	Construction Engineering Research Lab. Corps of Engineers Facility Systems Division P.O. Box 4005 Champaign, IL 61824-4005	217-373-7268	s[pmma,@csrd@io.edu
Ram P. Srivastav	Department of Applied Mathematics and Statistics State University of New York Stony Brook, NY 11794-3600	516-632-8364	rsrivastav@ccmail.sunysb.edu
J. Michael Steele	Department of Statistics The Wharton School University of Pennsylvania 3010 Steinberg Hall-Dietrich Hall Philadelphia, PA 19104-6302		
Kim Stelson	University of Minnesota Mechanical Engineering 125 Mech. Eng. 111 Church Street SE Minneapolis, MN 55455	612-625-6528	
John C. Strikwerda	Department of Computer Science University of Wisconsin-Madison 1210 West Dayton Street Madison, WI 53706	608-262-0822	strik@cs.wisc.edu

Name	Address	Phone	Email
Moss Sweedler	Mathematical Sciences Institute Cornell University White Hall Ithaca, NY 14853	607-255-4373	jesj@cornellc.cit.cornell.edu
James C. Turner, Jr.	Hampton University Department of Mathematics Hampton, VA 23668	804-727-5352	
Donald Truhlar	University of Minnesota Department of Chemistry 207 Pleasant Street Minneapolis, MN 55455	612-624-7575	mf12101@sc.msc.edu
Gregory Tawa	University of Minnesota Department of Chemistry 139 Smith Hall Minneapolis, MN 55455	612-825-8656	tawa@sc.msc.umn.edu
Thomas Ting	University of Illinois at Chicago CEMM Department (M/C246) Box 4348 Chicago, IL 60680		
Philip Van Valkenberg	MSC 1200 Washington Ave. South Minneapolis, MN 55415		
Mark Vangel	U.S. Army Materials Technology Laboratory ATTN: SLCMT-MRS-MM Arsenal Street Watertown, MA 02172		
John D. Vasilakis	Commander US Army Armament R&D Center ATTN: SMCAR-CCB-RA AMCCOM Benet Weapons Laboratory Watervliet, NY 12189-4050	518-266-5019	vasilaki@pica.army.mil
John W. Walter, Jr	Director U.S. Army Ballistic Research Laboratory ATTN: Dr. John W. Walter, Jr. Aberdeen Proving Ground, MD 21005-5066	301-278-6051	jwalter@brl.mil
J.R. Walton	Department of Mathematics Texas A&M University College Station, TX 77840		
Yeih J. Wang	Cullen College of Engineering Department of Electrical Engineering University of Houston University Park Houston, TX 77204-4793	713-749-4418	
Roger Wehage	Commander U.S. Army Tank-Automotive Command ATTN: AMSTA-RYA (Dr. Roger Wehage) Warren, MI 48397-5000	313-574-5378	
Richard Weiss	USACE-WES 3909 Halls Ferry Road Vicksburg, MS 39180	601-634-2194	
J. R. Whiteman	BICOM, Institute of Computational Mathematics Brunel University Uxbridge, Middlesex, UB8 3 PH United Kingdom	44-895-274000	john.whiteman@brunel.ac.uk
Stephen A. Wilkerson	Director U.S. Army Ballistic Research Laboratory ATTN: Dr. Stephen A. Wilkerson Interior Ballistics Division Aberdeen Proving Ground, MD 21005-5066	301-278-6131	swilker@ibd.brl.mil

Name	Address	Phone	Email
Wendy A. Winner	Director U.S. Army Ballistic Research Laboratory ATTN: Wendy A. Winner Vulnerability/Lethality Division Aberdeen Proving Ground, MD 21005-5066	301-278-6655	wendy@brl.mil
Julian J. Wu	Mathematical and Computer Sciences Division U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211	919-549-4332	jjwu@brl. mil
Stephen S. T. Yau	University of Illinois at Chicago Department of Mathematics P.O. Box 4348, M/C 249 Chicago, IL 60680		
Ashraf Zeid	Computer Science Corporation P.O. Box 5156 Warren, MI 48090-5156	313-574-7816	azeid@tacom.emh2.army.mil

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release: Distribution unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) ARO Report 92-1	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Army Research Office 6b. OFFICE SYMBOL (If applicable) SLCRO-MA	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) P.O. Box 12211 Research Triangle Park, NC 27709-2211	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. / PROJECT NO. / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)

Transactions of the Ninth Army Conference on Applied Mathematics and Computing

12. PERSONAL AUTHOR(S)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Technical Report	FROM Jan 91 TO Feb 92	1992 March	742

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Fluid and solid mechanics, mathematical physics and numerical methods, symbolic computation, control theory, and stochastic techniques.

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

(U) This is a technical report resulting from the Ninth Army Conference on Applied Mathematics and Computing. It contains most of the papers in the agenda of this meeting. These treat many Army applied mathematical problems.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL Dr. Francis G. Dressel	22b. TELEPHONE (Include Area Code) (919)549-4319 22c. OFFICE SYMBOL SLCRO-MA

DD Form 1473, JUN 86 *Previous editions are obsolete.*

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED