AFOSR-85-0346

Annual Technical Report Under AFOSR Grant
AFOSR-85-0346
Air Force Systems Command
Air Force Office of Scientific Research

LUMPED MODEL GENERATION AND EVALUATION: SENSITIVITY AND LIE ALGEBRAIC TECHNIQUES WITH APPLICATIONS TO COMBUSTION

By

F.L. Dryer, H. Rabitz* and R. Yetter

School of Engineering and Applied Science
Department of Mechanical & Aerospace Engineering

*Department of Chemistry
PRINCETON UNIVERSITY
Princeton, NJ 08544

October 1987

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

DTIC FILE COPY (2)

REPORT DOCUMENTATION PAGE

Form Approved OMB No 0704-0188

1a. REPORT SEC U:	1b RESTRICTIVE MARKINGS
2a. SECURITY CL	3 DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release; distribution is unlimited.
2b. DECLASSIFIC	
4 PERFORMING ORGANIZATION REPORT NUMBER(S): None	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR- 87-1806
6a. NAME OF PERFORMING ORGANIZATION: Princeton University; 6b. OFFICE SYMBOL (If applicable):	7a. NAME OF MONITORING ORGANIZATION: AFOSR/NA
6c. ADDRESS (City, State, and ZIP Code): Mechanical and Aerospace Engineering Princeton University, Princeton, NJ 08644	7b. ADDRESS (City, State, and ZIP Code): Building 410, Bolling AFB DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR/NA; 8b. OFFICE SYMBOL (If applicable): NA	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR-85-0346
8c. ADDRESS (City, State, and ZIP Code): Building 410, Bolling AFB DC 20332-6448	10 SOURCE OF FUNDING NUMBERS: PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2308; TASK NO A2; WORK UNIT ACCESSION NO

11. TITLE (Include Security Classification)
(U) Lumped Model Generation and Evaluation: Sensitivity and Lie Algebraic Techniques with Applications to Combustion

12. PERSONAL AUTHOR(S)

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Annual Technical Report	FROM 9-1-86 TO 10-1-87	87, 10, 1	

DTIC ELECTE JAN 07 1988 S D H

16. SUPPLEMENTARY NOTATION
F.L. Dryer, H. Rabitz and R. Yetter

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
07	01	
21	02	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
Semi-empirical Kinetics, Lumping, Lie Algebra

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This program dealt with the development and application of new approaches for producing and evaluating lumped parameter models of physical processes. Local and global sensitivity analysis procedures were studied for achieving this goal. Specifically, Lie group formalism was developed to address global parameter space mapping issues of temporal kinetics problems and extended to more complex reactive-diffusive problems. Furthermore, Lie group theoretical techniques were also used to gain analytic insight into the solution of nonlinear kinetic systems. Using local gradient methods, the lumpability (or model reduction) of hydrogen/oxygen and carbon monoxide/hydrogen/oxygen kinetic mechanisms were studied in various physical environments. It was found that the presence of strong scaling and self similarity in the sensitivities allowed for kinetic model simplification. Such scaling and similarity was found associated with strong thermal coupling in the systems. Lastly, a general analysis method for the exact lumping of chemical kinetic mechanisms was developed and illustrated by simple examples.

20 DISTRIBUTION/AVAILABILITY OF ABSTRACT	21 ABSTRACT SECURITY CLASSIFICATION
☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT ☐ DTIC USERS	Unclassified
22a NAME OF RESPONSIBLE INDIVIDUAL: Julian M Tishkoff	22b TELEPHONE (Include Area Code): (202) 767-4935; 22c. OFFICE SYMBOL: AFOSR/NA

DD Form 1473, JUN 86 — Previous editions are obsolete. — SECURITY CLASSIFICATION OF THIS PAGE

Unclassified

INTRODUCTION

The design and optimization of realistic engineering combustion devices involves the construction and execution of complex mathematical models. These models will typically involve combustion kinetics as well as transport processes with the physics and chemistry described by many parameters which are imprecisely known. In addition, existing freedom for choosing combustion chamber design will introduce other potentially controllable parameters into the model. Therefore, a central problem in all design problems concerns an understanding of system performance with respect to its parameter values. Except for the simplest models, such an understanding will necessitate extensive computer calculations, and repeated execution for each new set of parameters will lead to prohibitive expense. The goal of this program is to provide new insights as to how to simplify detailed submodels which cause the overall system calculations to be prohibitively difficult, and to exercise the techniques to develop simplified chemical kinetic models which provide sufficient detail for generating accurate modeling results. This goal is being pursued by developing and applying new techniques in the general areas of sensitivity analysis and Lie algebraic theories. Due to widespread applications of these two analysis methods, the outcome from this research program has important implications to many other problems arising in combustion phenomena as well as to other subjects of interest to the Air Force (e.g., system control, parameter identification and distinguishability, statistical parameter uncertainty, model scaling, etc.).

WORK STATEMENT

The work statement for this program is as follows:

1. Develop global sensitivity analysis techniques using Lie algebra for parameter space mapping and control of temporal systems. Special attention will be given to using the techniques for performing finite excursions through parameter space. As the tools develop, they will be applied to the lumping consideration above, as well as to design and control problems relevant to combustion systems.

2. Appropriate advanced development of item 1 are planned to extend the analysis procedures to more complex combustion chemistry and to include both spatial and temporal calculation comparisons of lumped and detailed models.

3. Determine criteria for exact lumping in chemical kinetics and investigate the implications of these lumpability conditions to model systems.

4. Model systems will be studied to establish the use of elementary sensitivity coefficients, Green's function elements and derived sensitivity coefficients for lumping purposes. Appropriate numerical procedures will be employed including eigenvector-eigenvalue analysis and rank reduction of the appropriate sensitivity matrices.

5. The sensitivity techniques of item 4 will be developed with hydrogen/oxygen and carbon monoxide/hydrogen/oxygen combustion systems as test cases for systematic model reduction and lumping. The ensuing lumped models will be compared with those in the literature for their predictive capabilities.

STATUS OF RESEARCH

During the past year research on several interrelated activities was pursued in the general area of combustion kinetics and sensitivity analysis particularly as related to model lumping and reduction. The thrust of these developments has largely been fundamental with the emphasis on creating and ultimately applying new methodology for treating several critical problems arising in combustion phenomena. Particular emphasis was given to modelling and theoretical concerns, although a portion of the research is directly related to the interface between laboratory measurements and modelling. A summary of these activities is given below.

I. GLOBAL SENSITIVITY ANALYSIS: A NEW PERSPECTIVE

This research is motivated by a number of important, but heretofore difficult problems, in combustion modelling including system lumping, statistical parameter uncertainty and system control. All of these issues as well as others necessitate obtaining an understanding of how the system behaves over broad region of its parameter space of rate constants, transport coefficients, etc. Traditional methods for achieving this information rely on repeated calculations at sample points in the system parameter space. Such an approach is prohibitively expensive and the results will typically provide little insight into the detailed workings of the system. With this information as background, we have been pursuing an entirely new approach based on Lie group techniques for mapping broad regions of system parameter space. The initial developments appear very promising as summarized below.

A. Uniform Temporal Reacting Systems[1]

One-parameter groups of transformations were used to investigate the effects of wide-ranging changes in rate constants and input/output fluxes upon homogeneous chemical reactions involving an arbitrary number of species in reactions of zero, first and second order.

Every transformation group is so chosen that it either exactly or approximately converts each solution of a set of rate equations into corresponding solutions of a one-parameter family of altered rate equations. All of these solutions have topologically equivalent equilibrium points and topologically equivalent phase trajectories in the space of concentration variables. Compounding the transformations yields transformations with the same properties.

The chemical significance of the transformations was illustrated by applying them to kinetic systems involving two reacting species. There are then twelve separate one-parameter groups of transformations available. The generators of all allowed one-parameter groups are obtained for systems of N

species using an algorithm which exactly determines their action on the rate constants, and either exactly determines or systematically approximates their action on the concentrations. The generators determine invariant functions that establish relations between the initial rate constants and the altered rate constants and between the initial concentration variables and the altered concentration variables.

Some mapping of the concentrations simply shift their values and may be used to study the effects of changes in input/output fluxes and rate constants upon concentrations. Other mappings create "lumped" concentration variables and may be used to systematically reduce the number of manifest concentration variables in nonlinear as well as linear kinetic equations.

A number of mappings of nonlinear kinetics may be used to obtain approximate linearizations valid in regions larger than those obtained by the usual power series expansions. In some cases the linearization is global and exact.

B. Reaction-Diffusion Systems[2]

The methodology developed in paragraph A above has an immediate transferral to the more complex and interesting class of reaction-diffusion problems. Although partial differential equations will in general admit a much broader set of Lie group mappings, their full determination will also be more difficult to establish. However, in the case that the system diffusion coefficients are not concentration dependent, exactly the same transformations developed for the purely temporal reacting systems may be applied beneficially to the case of reaction-diffusion. For example, transformations which rigorously, or at least regionally, linearize a reaction network have exactly the same effect on an analogous reaction-diffusion system. Application of these transformations has allowed us to analytically explore the interrelationship between diffusive transport and reaction kinetics. In addition, traditional local methods of stability analysis may now be significantly extended to cover regions of the system state and parameter spaces. We plan to use these tools for the treatment and analysis of realistic combustion models.

II. Analytic Insight Into the Solution of Kinetic Systems[3,4]

This research has a close connection with that of item I above in that it is also based on the use of Lie group theoretical techniques. The goal here is more limited with the purpose largely being the provision of analytical insight into the solution of nonlinear combustion kinetic systems. Traditional numerical methods for this purpose can typically provide varying degrees of accuracy but are quite inadequate with regard to their resultant insight. To deal with this problem we have developed a new analytic approximation scheme for the finite Lie transformation yielding the solution to a set of nonlinear kinetic equations. This work also has immediate applications to the global parameter space Lie generators involved with the study in item I above.

In this work a new method to factorize certain evolution operators into

an infinite product of simple evolution operators is presented. The method uses Lie operator algebra and the evolution operators are restricted to exponential form. The argument of these forms is a first order linear partial differential operator. The method has broad applications including to the areas of sensitivity analysis, the solution of ordinary differential equations and the solution of Liouville's equation. A sequence of ξ-approximants is generated to represent the Lie operators. Under certain conditions the convergence rate of the ξ-approximant sequences is remarkably rapid. This work presented the general formulation of the scheme and some simple illustrative examples.

Additional research was carried out to establish convergence theorems associated with its sequence of ξ-approxmimants. The theorems presented give the conditions which are sufficient for convergence of the sequences. Although the main emphasis was on convergence properties of the one-dimensional case, the generalization to multidimensional cases is quite straightforward. Further development and numerical illustrations are underway.

III. Kinetic System Identifiability and Distinguishability[5,6]

By following the kinetics of a reaction through the use of certain classes of measurable quantities instead of the concentrations of all species neither the parameter values nor the reaction scheme are necessarily unique. Identifiability deals with the problem of determining whether an experiment is able to supply the desired information on the parameters of an assumed kinetic model, whereas indistinguishability means that two different reaction schemes generate the same values for the observed quantities in any possible experiment. This work examined these issues for the case of first-order reaction systems and both problems are solved by the same analytical tools. The method involving Laplace transforms is conceptually simple, easy to apply, and is also used to derive simple rules to test distinguishability of reaction schemes. Another approach based on similarity transformations is used to generate all the first-order reaction schemes that are indistinguishable from a given one. These same concepts have been extended to nonlinear systems for the case of global identifiability.

IV. A General Analysis on Exact Lumping in Chemical Kinetics[7]

A general analysis of exact lumping has been developed. This analysis can be employed to any reaction system with n species described by a set of first-order ordinary differential equations $dy/dt = f(y)$, where y is an n-dimensional vector; $f(y)$ is an arbitrary n-dimensional function vector. Here we only consider lumping by means of a rectangular constant matrix M (i.e., $\hat{y}$ lower dimension than y with $\hat{y} = My$). It is found that a reaction system is exactly lumpable if and only if the intersection of the invariant or the null subspaces of the Jacobian matrix $J(y)$ of $f(y)$ for all values of y is nonempty. If the intersection is less than n, nontrivial lumping schemes can be obtained. It is proved that the Jacobian matrix can be represented as a linear combination of certain matrices and the intersection of the invariant or null subspace of the constant matrices is just that of the Jacobian matrix. After the determination of the intersections, all possible lumping matrices

can be obtained. The kinetic equations of the lumped system can be described as $d\hat{y}/dt = Mf(\bar{M}\hat{y})$, where $\bar{M}$ is any generalized inverse of M satisfying $M\bar{M} = I_m$. Several implications of these lumpability conditions were investigated as well as illustrated by some simple examples.

V. Hydrogen-Air Combustion Revisited Under a Variety of Conditions[8]

Surely the hydrogen-air combustion system has received the closest scrutiny both theoretically and experimentally. Yet, there is still a lack of complete understanding about which aspects of the system are important particularly under a variety of laboratory conditions. In order to further understand this issue, we have carried out modelling and sensitivity analysis studies under: a) purely temporal and isothermal conditions, b) purely temporal and adiabatic conditions, c) steady premixed adiabatic conditions. This triad of studies provides an interesting hierarchy allowing us to understand the role of diffusive transport as well as thermal coupling. One motivation for this work arose from previous studies showing extensive scaling and self similarity behavior amongst the hydrogen-air sensitivity coefficients. This behavior has significant implications for model simplification both in this system as well as complex combustion problems. It had been previously speculated that the temperature was providing the dominant coupling to produce self organization amongst all the system species and rate parameters. This study has confirmed this conjecture as well as revealed a number of other underlying subtleties including the role of diffusion. In addition, it was shown that the presence of strong scaling and self similarity in the premixed flames allowed for kinetic model simplification.

VI. Model Reduction and Lumping of Carbon Monoxide Oxidation Kinetics

In our previous studies, normalized sensitivity coefficients, $S_{ij} = \partial \ell n\, C_i / \partial \ell n \alpha_j$, have been studied to determine the relative importance of elementary reactions or certain groups of reactions in comprehensive mechanisms. We have presently extended this methodology by using the principal component analysis method of Vajda and Turanyi [J. Phys. Chem., March 1986] in order to systematically reduce the size of the original comprehensive mechanism. Briefly, this methodology is based on a least squares fit approach by first defining the response function, Q, as

$$Q(\underline{\alpha}) = \sum_{j=1}^{q} \sum_{i=1}^{m} \left[\frac{C_i(t_j,\underline{\alpha}) - C_i(t_j,\underline{\alpha}^*)}{C_i(t_j,\underline{\alpha}^*)} \right]^2$$

where α^* is the nominal values of the parameters, then by introducing the classical Gauss-approximation to yield $Q(\underline{\alpha}) = \tilde{Q}(\hat{\underline{\alpha}}) = (\Delta\hat{\underline{\alpha}})^T \underline{\underline{S}}^T \underline{\underline{S}} (\Delta\underline{\alpha})$ where $\hat{\alpha}_j = \ell n \alpha_j$, and finally by performing an eigenvalue-eigenvector decomposition on the resulting cross-product matrix $\underline{\underline{S}}^T\underline{\underline{S}}$. Eigenvectors corresponding to small eigenvalues indicate unimportant reactions, thereby enabling one to optimally reduce the mechanism.

Along these lines we have continued our previous work by applying this methodology to the $CO/H_2/O_2$ reaction mechanism. A large number of isothermal

temporal problems were numerically run to generate a data base which would be representative of combustion environments. The data base covered a temperature range from 800 to 1800 K, several equivalence ratios from lean to rich conditions, and several pressures. The principal component analysis was applied to this data base to determine the minimum reaction set that would reproduce all the original species concentrations within 2%.

The results showed that the original 52 reaction mechanism could be successfully reduced to one consisting of 27 reactions while retaining all 12 species in the model.

Obviously, this reduction is still not practical for use in large multidimensional codes. The necessary further reductions are proposed to proceed along several directions. First, the constraint of retaining all species will be lifted. Our earlier research has shown that in addition to the major reactants and products, two intermediate species are necessary in the model. Secondly, we have also found that in more complex environments, such as adiabatic premixed flames, the underlying chemical processes are much more coupled (namely through the heat release of the reaction) and hence, such problems are anticipated to be more directly lumpable.

Cumulative Chronological List of Publications

This list of publications covers the period from September 1986 to October 1987. Other prior publications under this grant may be found in the previous Annual Technical Reports.

1. C. Wulfman and H. Rabitz, "Global Sensitivity Analysis of Temporal Reactive Systems", J. Chem. Phys., submitted.

2. C. Wulfman and H. Rabitz, "Global Sensitivity Analysis of Reaction-Diffusion Systems Using Lie Group Techniques", manuscript in progress.

3. M. Demiralp and H. Rabitz, "Factorization of Certain Evolution Operators Using Lie Algebra: Formulation of the Method", J. Math. Phys., submitted.

4. M. Demiralp and H. Rabitz, "Factorization of Certain Evolution Operators Using Lie Algebra: Convergence Theorems", J. Math. Phys., submitted.

5. S. Vajda and H. Rabitz, "Identifiability and Distinguishability of First-Order Reaction Systems", J. Phys. Chem., 1987, in press.

6. S. Vajda and H. Rabitz, "State Isomorphism Approach to Global Identifiability of Nonlinear Systems", IEEE Control, submitted.

7. G. Li and H. Rabitz, "A General Analysis of Exact Lumping in Chemical Kinetics" Chem. Eng Sci., submitted.

8. S. Vajda, H. Rabitz, R. Yetter, and F.L. Dryer, "Influence of Thermal Coupling and Diffusion on the Mechanism of H_2 Oxidation in Steady Premixed Laminar Flames", manuscript in progress.

Presentations

This list of presentations covers the period from September 1986 to October 1987. Other prior presentations under this grant may be found in the previous Annual Technical Reports.

H. Rabitz, University of Paris, France, December 1986.

H. Rabitz, Marmara Institute of Turkey, December 1986.

H. Rabitz, Middle East Technical University, August 1987.

H. Rabitz, Yale University, April 1987.

H. Rabitz, International Conference on Numerical Modelling in Ankara, August 1987.

H.Rabitz, The Sixth International Conference on Applied Mathematics and Modeling in St. Louis, MO, July 1987.

R. Yetter, University of Kentucky, KY, November 1987.

Consultative and Advisory Functions

H. Rabitz, U.S. Army at Picatinny, NJ.

H. Rabitz, Research Associates of Lawrenceville, NJ.

F. Dryer, TRW Systems, Redondo Beach, CA.

Personal

During the past year, the professional personnel associated with this research effort included Professors F.L. Dryer, H. Rabitz, and C. Wulfman and Drs. R.A. Yetter, S. Vajda, M. Demiralp, and S. Shi. For each, a professional biography is included. Also during this past year, Mr. Li, a third year graduate student in the Department of Chemistry at Princeton University under the supervision of H. Rabitz, contributed to this program.

Vitae

DR. FREDERICK L. DRYER

PROFESSOR
UNDERGRADUATE DEPARTMENTAL REPRESENTATIVE
PRINCETON UNIVERSITY

Dr. Frederick L. Dryer received his Bachelor of Engineering Degree in Aeronautical Engineering from Rensselaer Polytechnic Institute, Troy, New York in 1966 and a Phd. degree in Aerospace and Mechanical Sciences from Princeton University, Princeton, New Jersey in 1972. After serving on the Professional Research Staff in the Mechanical and Aerospace Engineering Department of Princeton University for eight years, he joined the academic faculty of the department as a tenured Associate Professor in 1981. Dr. Dryer was on sabbatical leave during the 1982-83 academic year as a private consultant to industry, and he returned to the University in July 1983 as a Full Professor in the Mechanical and Aerospace Engineering Department. Dr. Dryer has served as Undergraduate Departmental Representative since November of 1984.

Dr. Dryer's principal research interests are in the fundamental combustion sciences with particular emphasis in high temperature combustion chemistry, formation/ignition/secondary atomization/liquid phase chemistry of fuel droplets, and fire-safety-related properties of conventional and synthetic fuels. In collaboration with Professor Irvin Glassman (Princeton University), he developed and maintains a Fuels Research Group typically consisting of seven professionals and ten graduate students, and occupying about 3500 square feet of laboratory space. Research efforts are supported by grants and/or contracts from government, private foundations and industry.

Dr. Dryer has published over seventy-five technical articles, lectured, and consulted extensively on the above as well as other combustion and energy-related subjects. He has contributed invited presentations on two separate occasions to the International Symposiums of the Combustion Institute (1976, 1981) and on numerous other occasions to the regional, national, or international meetings of other organizations including the American Chemical Society, AGARD/NATO, American Institute of Aeronautics and Astronautics, American Physical Society, Eastern, Central, and Western Sectional Meetings of the Combustion Institute, National Bureau of Standards, U.S. Army Research Office, and National Aeronautics and Space Administration. In 1979-80, he served on technical panels and committees for the National Academy of Sciences in assessing automotive emission characteristics and hazards from both catalytically-controlled spark ignition and diesel light vehicles.

Dr. Dryer served as Associate Editor of the international journal, Combustion Science and Technology, from 1977-1986 and is presently a member of the editorial board. He is a member of Tau Beta Pi, Sigma Gamma Tau, and Sigma Xi Engineering Honoraries, The Combustion Institute, The American Chemical Society, and The American Society of Mechanical Engineers.

Herschel Rabitz

Personal and Professional Vita

PII Redacted

[redacted]

Education:	B.S. in Chemistry, University of California Berkeley, 1966
	Ph.D in Chemical Physics, Harvard University, 1970
	Postdoctoral Associate, University of Wisconsin, 1970-1971
Academic Experience:	Assistant Professor of Chemistry, Princeton University 1971-1976
	Associate Professor of Chemistry, Princeton University 1976-1980
	Professor of Chemistry, Princeton University, 1980-
	Applied Mathematics Program, Princeton University, Affiliated Member and Acting Director, 1984
Societies:	American Physical Society American Association for the Advancement of Science American Chemical Society Sigma Xi Phi Beta Kappa
Honors:	Camille and Henry Dreyfus Teacher-Scholar, 1974-1979 Afred P. Sloan Fellow, 1975-1979
Nonacademic Activities:	Chairman, Princeton Section of the American Chemical Society, 1977-1978
	Associate Chairman, Gordon Conference on Few Body Problems in Chemistry and Physics, 1981
	Chairman, Gordon Conference, Dynamics of Simple Systems in Chemistry and Physics, 1984

Professor Carl Wulfman

Professor Wulfman received his B.S. in chemistry and math from the University of Michigan in 1953 and attained his Ph.D. from the University of London in 1957. His thesis research involved theoretical organic chemistry. In 1961, he came to the University of the Pacific, Stockton as Chairman of the Pysics Department where he is currently a Professor of Physics. During the fall of 1983, he came to Princeton as a Visiting Fellow to work with Professor Rabitz. Research interests include group analysis of differential equations, physics and chemistry; molecular physics and atomic physics.

DR. R.A. YETTER

PROFESSIONAL RESEARCH STAFF
FUELS RESEARCH LABORATORY

Dr. Yetter holds four degrees in Mechanical Engineering; a Bachelor of Science degree from Syracuse University (1974), a Master of Science degree from Cornell University (1980), a Master of Arts degree from Princeton University, and a Doctor of Philosophy degree from Princeton University (1985).

Dr. Yetter has both industrial and laboratory experience as a Research Engineer with the Chemical Sciences Laboratory, Fuels and Lubricants Department, of the Ford Motor Company (3 years) and as a Research Assistant with the Combustion Sciences Laboratory, Department of Energy and Environmental Science, of the Brookhaven National Laboratory (1/2 year). He is currently a Research Collaborator for the Department of Applied Science at the Brookhaven National Laboratory. In July 1985, he joined the Professional Research Staff in the Fuels Research Laboratory.

Dr. Yetter's principal research interest is in the combustion science field with current emphasis in high temperature combustion chemistry, flame structure, and sensitivity analysis theory. He has devoted his work equally among experimental and theoretical studies particularly emphasizing the integration of these two efforts.

Dr. Yetter has been a reviewer for major journals in the field of combustion, and has presented a number of papers at both national and international scientific meetings. He is also an author of several scientific publications.

CURRICULUM VITAE

1. Personal details

Name: Sandor Vajda

PII Redacted

[redacted]

Present post: Research Staff Member
Department of Chemistry and
Department of Aerospace and Mechanical Engineering
Princeton University
Princeton,NJ 08544

[redacted]

2. Career history

Education and qualifications

Sept 67 - July 72	I.M.Gubkin University of Oil Chemistry and Natural Gas Industry, Moscow, USSR M.Eng.S. Degree subject: Control Engineering.
Sept 73 - July 80	Eotvos Lorand University Budapest, Hungary M.S. Degree subject: Applied Mathematics.
May 75	Institute of Chemical Engineering Veszprem, Hungary Doctoral degree in chemical engineering Thesis subject: Parameter estimation in chemistry and chemical engineering.
Oct 83	Hungarian Academy of Science Ph.D. in Chemistry and Applied Mathemetics Thesis subject: Identifiability of kinetic models.

3. Work experience

Sept 72 - Sept 73	Assistant Professor Department of Mathematics Institute for Chemical Engineering Veszprem, Hungary Subject: Undergraduate mathematics
Sept 73 - May 85	Assistant (from 76:Associate) Professor, Laboratory for Chemical Cybernetics Eotvos Lorand University, Hungary Subjects: Computational methods in chemistry in biology Modeling and control of chemical and biomedical processes
June 85 - Aug 85	Research Fellow Laboratoire des Signaux et Systemes, Gif-sur-Yvette, France. Subject: Identification of biomedical systems.
Oct 85 - Sept 86	Research Fellow Control Theory Center and Department of Engineering University of Warwick, Coventry, England. Subjects:Identifiability of nonlinear models; and Identification of pharmacokinetic systems.
Sept 86 - date	Research Staff Member Department of Chemistry and Department of Mechanical and Aerospace Engineering Princeton University Research projects: Identifiability analysis in reaction kinetics and chemical dynamics; Simplification and lumping of complex models; Modeling of combustion processes and simplification of complex combustion models;
Further information :	I speak English, Russian, French, and German.I am member of the Editorial Board of the American Journal of Physiology, section Modeling Methodology Forum.

Dr. Shenghua Shi, Research Associate

Dr. Shenghua Shi was trained as an engineer in the Peoples Republic of China and in 1984 attained his Ph.D. in chemical physics from the University of California, Berkeley, under the direction of Professor William Miller. His thesis research concerned quantum mechanical and semiclassical approaches to molecular dynamics. He joined Professor Rabitz' research group in November, 1984. Other research interests include functional Lie group techniques for the solution and analysis of physically important differential equations, the study of intramolecular energy transfer in complex polyatomic molecules; and application of control theory ideas to the design of optical pumping experiments to ultimately control chemical processes.

Dr. Metin Demiralp, Visiting Research Staff

Dr. Demiralp received his M.S. in chemical engineering in 1971 and his Ph.D. in physical chemistry in 1975 from the University of Istanbul, Turkey. From 1970 to 1978 he was affiliated at various levels with the Marmara Research Institute in Turkey, Applied Mathematics Department. He came to Princeton for a year as a research associate in Professor Rabitz' group and then joined the Research Institute of Basic Sciences in Turkey, Applied Mathematics Department. During the summers of 1985 and 1986, he once again returned to Princeton as Visiting Research Staff. His research interests include applied mathematics; quantum mechanics; differential equations; linear operators; quantum chemistry; evolution operators (classical and quantum mechanical cases); algebraic methods of eigenvalue problems; and resolvent techniques.

FACTORIZATION OF CERTAIN EVOLUTION OPERATORS USING LIE OPERATOR ALGEBRA: CONVERGENCE THEOREMS

Metin Demiralp* and Herschel Rabitz
Princeton University
Department of Chemistry
Princeton, New Jersey 08544, USA

*Permanent Address: Tübitak Research Institute for Basic Sciences
Applied Mathematics Department
P.O. Box 74, 41470
Gebze-Kocaeli, TURKEY

ABSTRACT

This work concerns convergence theorems associated with a sequence of ξ - approximants for exponential evolution operators with Lie operator arguments. A companion paper presents the formulation of the ξ - approximants. The theorems presented in this paper give the conditions which are sufficient for convergence of the sequences. Although the main emphasis will be on convergence properties of the one-dimensional case, the generalization to multidimensional cases is quite straightforward.

I. INTRODUCTION

In a companion paper we developed a new factorization scheme[1] for exponential evolution operators with Lie operator arguments. The factorization was based on ordering of contributions to the evolution operator with respect to deviations from a steady-state solution. Hence, in the Lie operator of the form $\underset{\sim}{f}(\underset{\sim}{x})\cdot\underset{\sim}{\nabla}$ the function $\underset{\sim}{f}(\underset{\sim}{x})$ must vanish around the origin $\underset{\sim}{x}=0$. The factorization scheme results in an infinite product of elementary evolution operators and the approximation to the desired overall evolution operators is achieved by a truncation of the infinite product to order n. This procedure produces a sequence of ξ - approximants to the desired evolution operator. The effect of the Lie transformation, or its approximate representation, on the position vector $\underset{\sim}{x}$ is fundamental in the theory since many of the basic operations may be related to certain properties of Lie transformations. A simple first order recursion relation may be found for the ξ - approximants, however they are rich in singularities. As the limit of the sequence of the ξ - approximants is taken, infinitely many branch point trajectories may exist in the complex $\underset{\sim}{x}$-plane. The flexibility inherent in the ξ - approximants suggests that this approach may rapidly converge to accurately approximate the effect of the evolution operator on $\underset{\sim}{x}$. This conjecture was confirmed in a number of applications[1] although certain cases exhibited slow or non-convergent characteristics.

Such empirical evidence is helpful, but a mathematical proof of the convergence behavior is needed in order to intelligently use the method in realistic applications. It is necessary to establish not only the existence of convergence but also determine the criteria under which convergence is expected. The purpose of this paper is to address these latter issues.

In order to mathematically explore the convergence characteristics, Section II will investigate the singularities of the ξ - approximants. This section will also define some useful fundamental concepts. Section III will present the convergence theorems for the ξ - approximant sequences. These latter developments will be carried out for the one-dimensional case and Section IV will generalize the theorems to the multi-dimensional case. Finally, Section V presents concluding remarks.

II. Singularities of ξ - Approximants and Some Fundamental Definitions in the One Dimensional Case

The evolution operator of concern has the form $Q=\exp\left[f(x)\frac{\partial}{\partial x}\right]$ where $f(x)$ is a specified function defining the Lie Operator. The action of Q on x is approximated by a sequence of ξ - approximants, $Qx \sim \xi_n$ such that

$$\xi_{n+1} = \frac{\xi_n}{\left[1 - n\sigma_{n+1}(t)\ \xi_n^n\right]^{1/n}} \quad ; \qquad \xi_1(x,t) = x\ \exp(f_1 t) \tag{II.1}$$

where

$$Q = \exp\left[tf(x)\ \frac{\partial}{\partial x}\right] = \prod_{j=1}^{\infty}\ \exp\left[\sigma_j(t)\ x^j \frac{\partial}{\partial x}\right] \tag{II.2}$$

and

$$f(x) = \sum_{k=1}^{\infty} f_k\ x^k \tag{II.3}$$

The coefficients $\sigma_j(t)$ in each of the elementary exponential operators in Eq. (II.2) are global functions of time. The evaluation of these coefficients establishes the terms of the recursion relation in Eq. (II.1) and the details of this operation were presented in an earlier paper[1]. The iteration in Eq. (II.1) may be written in explicit form as

$$\xi_2(x,t) = \frac{x\ \exp\ (f_1 t)}{1-\sigma_2\ x} \tag{II.4}$$

$$\xi_3(x,t) = \frac{x\ \exp\ (f_1 t)}{\left[(1-\sigma_2\ x)^2 - \sigma_3\ x^2\right]^{1/2}} \tag{II.5}$$

$$\xi_n(x,t) = \Big[\Big[\ldots\Big[1-\bar{\sigma}_2\,x\Big]^2 - \bar{\sigma}_3\,x^2\Big]^{\frac{3}{2}} - \bar{\sigma}_4\,x^3\Big]^{\frac{4}{3}} \quad \ldots.$$

$$\bar{\sigma}_{n-1}\,x^{n-2}\Big]^{\frac{n-1}{n-2}} - \bar{\sigma}_n\,x^{n-1}\Big]^{-\frac{1}{n}}$$

$$\cdot\, x \exp(f_1 t) \qquad \text{(II.6)}$$

where

$$\bar{\sigma}_{n+1}(t) = n\sigma_{n+1}(t)\exp(nf_1 t) \qquad \text{(II.7)}$$

The structure of ξ_n may be identified as a type of continued fraction.

The origin in the complex x-plane is not a singular point for all the ξ_n's as long as t remains finite. Since $\sigma_j(0) = \bar{\sigma}_j(0) = 0$, all singularities of the approximants are gathered at infinity at the initiation of the evolution. Each singularity moves along a trajectory in the complex x-plane as time evolves and may or may not reach the origin when t tends to infinity. As a specific example we will now examine the second approximant, $\xi_2(x,t)$. This approximant has a rather simple singularity, a pole, whose location is given as follows:

$$x_p = \frac{f_1}{f_2[\exp(f_1 t) - 1]} \qquad \text{(II.8)}$$

where we have made use of the formula

$$\sigma_2(t) = \frac{f_2}{f_1}\,(1-\exp(-f_1 t)) \qquad \text{(II.9)}$$

Since the expansion coefficients f_n are assumed to be real, the pole in Eq. (II.8) is evidently located on the real axis of the complex x-plane. The pole starts to move from either

$+\infty$ ($f_2>0$) or $-\infty$ ($f_2<0$) to a limiting point as time, t, tends to infinity. If $f_2=0$, the pole remains at infinity. In general, two different cases occur as time evolves assuming that $f_2\neq 0$,

$$\lim_{t=\infty} x_p(t) = \begin{cases} 0 & f_1 \geq 0 \\ \left|\frac{f_1}{f_2}\right| & f_1 < 0 \end{cases} \qquad \text{(II.10)}$$

It is apparent from Eq. (II.10) that if the system under consideration is unstable, $f_1 \geq 0$, then the trajectory of the singular point ends at the origin. However, if the system is stable, $f_1 < 0$, then the singular point stops at a finite location away from the origin on the real axis. Therefore, at least for this approximant, there is a "clean" region where a singularity can never appear if $f_1 < 0$ and the origin of the complex x-plane is an interior point of this clean region. If the system under consideration is unstable, the origin may again be included in this clean region, however, in this case it becomes a point located on the border of the clean region.

In order to gain further insight into the ξ_n - approximants, we shall now examine the next approximant, ξ_3. This approximant has four branch points, two of which are located at infinity and the remaining ones are given below (where $\bar{\sigma}_3 \geq 0$, otherwise branch points are complex).

$$x_1 = [\bar{\sigma}_2(t) + [\bar{\sigma}_3(t)]^{\frac{1}{2}}]^{-1}$$
$$x_2 = [\bar{\sigma}_2(t) - [\bar{\sigma}_3(t)]^{\frac{1}{2}}]^{-1} \qquad \text{(II.11)}$$

These singularities are algebraic branch points with two Riemann sheets. Depending on the nature of the system,

$\bar{o}_2^2 - \bar{o}_3$ may be positive, zero or negative. If it differs from zero, then the origin becomes an interior point of the clean region for this approximant.

There is a remarkable property about the ξ - approximants which can be stated as follows. If ξ_j has a singularity which is a branch point (except for the case $j = 2$), then every ξ_k - approximant ($k>j$) will have the same singularity. This means that when j tends to infinity there will be an abundance of branch point trajectories in the complex x-plane. Any given trajectory may or may not be in the clean regions in the complex x-plane. As we shall see, the proof of the convergence of the ξ - approximant sequences completely depends on the existence of these regions and their locations.

It is now useful to make some definitions before proceeding. A given system is ultimately prescribed by the behavior of the function f(x) describing the corresponding Lie operator. If the complex x-plane of such a system has a region where any portion of the branch point trajectories of the ξ - approximants never exist there, then we shall call this region a "clean region" in accord with the use of these words above. If additionally, this region includes the origin of the complex x-plane as an interior point, then this region will be called the "main clean" region of the system. We further define a "global normal" system as follows: iff a system described by f(x) has a main clean region with a non zero measure it is a global normal system where we have used a measure in the sense that the measure of any countable infinite set vanishes. This latter measure is employed

to exclude the possibility of having a clean region which only includes the origin. The interpretation of this definition of a global normal region can be made as follows: if we deal with a finite period of time then the system will apparently have a main clean region. If we denote this region by $R(t)$ then we can write

$$\lim_{t\to\infty} \quad R(t) = R_g \supset [0] \qquad m(R_g) > 0 \tag{II.12}$$

In other words, the main clean region will continue to have an infinite number of uncountable points around the origin when time tends to infinity if the system is global normal. This definition may be relaxed by limiting ourselves not just to a semi-infinite time period, but to a finite one starting from $t=0$. Therefore, we can define the "temporary normal" system as follows: a system described by $f(x)$ is temporary normal <u>iff</u> it has a main clean region with a non zero measure (m) for a given time period $[0,T]$. Here, m is defined again in such a way that the measure of every finite or countable infinite set is zero. Finally all remaining systems will be "abnormal". As can be observed all global normal systems are at the same time temporary normal, and all abnormal systems can be considered as a limiting case ($T\to 0$) of temporary normal systems.

III. Convergence Theorems in the One-Dimensional Case

From an examination of Eqs. (II.1), (II.5-7) we may rewrite the approximants $\xi_j(x,t)$ as follows:

$$\xi_j(x,t) = \frac{x \exp(f_1 t)}{\Delta_j(x,t)} \qquad \text{(III.1)}$$

The function $\Delta_j(x,t)$ in the denominator satisfies the recursion relation

$$\Delta_n(x,t) = \left[\Delta_{n-1}^{n-1}(x,t) - \sigma_n x^{n-1}\right]^{\frac{1}{n-1}} \quad ; \Delta_1 = 1 \qquad \text{(III.2)}$$

One may conclude from this relation that the serial representation of $\Delta_n(x,t)$ in positive integer powers of x with time-dependent coefficients will converge within a finite circle of non zero radius around the origin of the complex x-plane for some time period [0,T]. One can then construct a Majorant series for this function such that

$$|\Delta_n(x,t)| < D_n(x,t) \; ; \; D_n > 1 \quad ; \; |x| < \rho_n(t) \qquad \text{(III.3)}$$

where $\rho_n(t)$ denotes the time dependent convergence radius of the Majorant series. The expression for the bound D_n may be established as follows

$$|\Delta_n^n(x,t)| < D_n^n(x,t) \Rightarrow$$

$$|\Delta_n^n(x,t) - \sigma_{n+1} x^n| < |\Delta_n^n(x,t)| + |\sigma_{n+1}| |x|^n$$

$$< D_n^n(x,t)\left[1+|\sigma_{n+1}| |x^n|\right]$$

$$\Rightarrow |\Delta_{n+1}| < D_n(x,t)\left[1+|\sigma_{n+1}| |x|^n\right] \Rightarrow$$

$$D_{n+1}(x,t) = D_n(x,t)\left[1+|\sigma_{n+1}| |x|^n\right] \qquad \text{(III.4)}$$

This latter result implies that

$$D_\infty(x,t) = D_N(x,t) \prod_{j=1}^{\infty} \left[1+\left|\sigma_{N+j}\right| x^{N+j}\right] \qquad \text{(III.5)}$$

If the infinite product in Eq. (III.5) is convergent, this is the region of the complex x-plane defined by $|x| < \rho(t)$ and

$$\rho(t) \geq \rho_{min} > 0 \qquad \text{(III.6)}$$

for all t-values, then $D_\infty(x,t)$ will converge to a finite value. This result also implies that the function $\Delta_\infty(x,t)$ will converge for all t-values in a region defined by $|x| < \rho_{min}$. The existence of such a convergence implies that the zeros of the function $\Delta_n(x,t)$ in the complex x-plane are bounded from below in absolute value for all times. This in turn means that the system is global normal.

The condition for convergence of the infinite product in Eq. (III.5) is equivalent to establishing the convergence of the following expression

$$d_N(x,t) = \sum_{j=1}^{\infty} | \sigma_{N+j} | |x|^{N+j} \qquad \text{(III.7)}$$

If this sum converges and remains smaller than unity for sufficiently large N values, then the infinite product in Eq. (III.5) also converges. If ρ_{min} in Eq. (III.6) vanishes, then two circumstances may occur:

$$\text{i)} \quad \rho(t) \geq \rho_{min}(T) > 0 \qquad t \in [0,T] \qquad \text{(III.8)}$$

$$\text{ii)} \quad \rho(t) \geq \rho_{min}(T) \qquad \rho_{min}(T) = 0 \text{ (except } T=0) \qquad \text{(III.9)}$$

The first of these cases corresponds to a temporary normal system, while the second implies the abnormal case. We have therefore proved the following theorem.

Theorem I: If the following infinite sum

$$d(x,t) = \sum_{j=1}^{\infty} \left| \bar{o}_j \right| x^j \qquad \text{(III.10)}$$

converges in a circle around the origin of the complex x-plane $|x|<\rho(t)$, then the following statements hold

(i) if $\rho(t) \geq \rho_{min} > 0$ for $t \in [0,\infty]$, the system is global normal

(ii) if $\rho(t) \geq \rho_{min}(\tau) > 0$ for $t \in [0,\tau]$ with $\tau>0$, the system is, at least, temporary normal

Corollary I

If the first condition (i) of Theorem I holds, then the sequence of ξ - approximants converges for all x and t values in the regions $(-\rho_{min}, \rho_{min})$ and $[0,\infty]$ respectively, and they have a permanent main clean region with non-zero measure.

Corollary II

If the second condition (ii) of Theorem I holds, the sequence of ξ - approximants converges at least for all x and t values in the regions $(-\rho_{min}(\tau), \rho_{min}(\tau))$ and $[0,\tau]$, $\tau > 0$ respectively and they have at least a temporary clean region around the origin of the complex x-plane.

We now seek to more explicity express the relation between the convergence condition of $d(x,t)$ and the nature of the system. As derived in the companion paper[1], the σ - coefficients are described as

$$\sigma_{n+1}(t) = \mathcal{F}_n(0) \tag{III.11}$$

with

$$\sigma_{n+1}(x) = \frac{\mathcal{F}_n\left[x\left[1+n\ \sigma_{n+1}\ x^n\right]^{-\frac{1}{n}}\right] - \mathcal{F}_n(0)}{x} \tag{III.12}$$

and

$$\mathcal{F}_1(x) = \sum_{j=0}^{\infty} f_{j+2}\ \exp\left(-(j+1)\sigma_1\ x^j\right) \tag{III.13}$$

where the time dependence of the $\mathcal{F}$'s is not shown explicitly.

Now, if we assume that $f(x)$ converges for $|x| \leq (\rho_f > 0)$, we can write the following inequality

$$|f_{j+2}| \leq \frac{A_f}{\rho_f^j} \tag{III.14}$$

This relation, however, permits us to construct the following Majorant function for $\mathcal{F}_1$

$$M_1(x) = \frac{A_f \exp(-\sigma_1)}{1 - \frac{x\ \exp(-\sigma_1)}{\rho_f}} \quad , \quad |x| < \rho_f\ \exp(\sigma_1) = \rho_f\ \exp(f_1 t) \tag{III.15}$$

Let us now assume that we have found a Majorant function for $\mathcal{F}_n$ as follows

$$\mathcal{F}_n(x) = \sum_{j=0}^{\infty} \mathcal{F}_n^{(j)}\ x^j \quad ; \quad \left|\mathcal{F}_n^{(j)}\right| < \frac{A_n}{\rho_n^j} \tag{III.16}$$

where $\mathcal{F}_n^{(j)}$ stands for time-dependent coefficients and A_n, ρ_n denote certain time-dependent constants. The last assumption, however, makes it possible to write the following expression for M_{n+1}, the Majorant function of $\mathcal{F}_{n=1}$, as can be revealed after a careful examination of the recursion given by Eq. (III.12)

$$M_{n+1}(x) \geq \frac{M_n\left[x\left[1-n|\sigma_{n-1}||x|^n\right]^{-\frac{1}{n}}\right]}{x} \qquad \text{(III.18)}$$

If we use the expression of M_n given by Eq. (III.17), we can write

$$M_{n+1}(x) \geq \frac{A_n}{\rho_n} \frac{G_n(x)}{1-\left[n|\sigma_{n+1}| + \frac{1}{\rho_n^n}\right]^{\frac{1}{n}} x} \qquad \text{(III.19)}$$

where

$$G_n(x) = g_{1,n}(x) \;/\; g_{2,n}(x) \qquad \text{(III.20)}$$

$$g_{1,n}(x) = \sum_{j=0}^{n-1} \left[1-n|\sigma_{n+1}|x^n\right]^{1-(j+1)/n} \rho_n^{-j} \; x^j \qquad \text{(III.21)}$$

$$g_{2,n}(x) = \sum_{j=0}^{n+1} \left[n|\sigma_{n+1}| + \frac{1}{\rho_n^n}\right]^{\frac{j}{n}} x^j \qquad \text{(III.22)}$$

Since $G_n(0) = 1$ and $G_n(x)$ is a monotonic decreasing function of x, we can construct the following Majorant function for the right hand side of Eq. (III.18)

$$M_{n+1}(x) = \frac{A_{n+1}}{\left[1 - \frac{x}{\rho_{n+1}}\right]} \qquad \text{(III.23)}$$

where

$$A_{n+1} = \frac{A_n}{\rho_n} \quad ; \quad A_1 = A_f \exp(-f_1 t) \qquad \text{(III.24)}$$

and

$$\frac{1}{\rho_{n+1}} = \left[n|\sigma_{n+1}| + \frac{1}{\rho_n^n}\right]^{\frac{1}{n}} \quad ; \quad \rho_1 = \rho_f \exp(f_1 t) \qquad \text{(III.25)}$$

If we assume that $[n|\sigma_{n+1}|]^{\frac{1}{n}} \exp(f_1 t)$ is bounded by ν, then we can write

$$\alpha_n \leq \exp(-f_1 t)\ \rho_n \qquad \text{(III.26)}$$

$$\alpha_{n+1} = \frac{\alpha_n}{\left[1 + \nu\ \alpha_n^n\right]^{\frac{1}{n}}} \qquad \alpha_1 = \rho_f \qquad \text{(III.27)}$$

Therefore, we have made the convergence radii of the Majorant functions smaller. As can be easily shown after some intermediate steps, α_n monotonically converges to a nonzero limit, say α, as n tends to infinity. This makes it possible to write

$$B_{n+1} = \frac{B_n \exp(-f_1 t)}{\alpha} \quad ; \quad B_1 = A_f \qquad \text{(III.28)}$$

$$M_n = \frac{B_n}{\left[1 - \exp(-f_1 t)\ \frac{x}{\alpha_n}\right]} \qquad \text{(III.29)}$$

Since $\sigma_n < M_{n+1}$ we can obtain

$$|\sigma_{n+1}(t)| < \frac{A_f}{\alpha^n}\left[\frac{1 - \exp(-nf_1 t)}{nf_1}\right] \qquad n \geq 1 \qquad \text{(III.30)}$$

which obviously satisfies the boundedness condition of $[n|\sigma_{n+1}|]^{\frac{1}{n}} \exp(f_1 t)$ globally for $f_1 < 0$ and temporarily for $f_1 > 0$. This result immediately produces the following theorem.

Theorem II: If the descriptive function of a given system is denoted by $f(x)$, $(f(0)=0)$, then the following statements are true.

(i) if $f(x)$ has a finite convergence radius centered at the origin of the complex x-plane and $f_1<0$, then the system is global normal.

(ii) if the same conditions of case (i) hold except that $f_1>0$, then the system is at least temporary normal.

Our third theorem concerns the ξ_n - approximants. Let us consider the inverse relation between ξ_n and ξ_{n+1}.

$$\xi_n = \frac{\xi_{n+1}}{\left[1+n\sigma_{n+1}\ \xi_{n+1}^n\right]^{\frac{1}{n}}} \tag{III.31}$$

If we write

$$\nu = \min\left\{\left[\left|\frac{1}{n\sigma_{n+1}}\right|\right]^{\frac{1}{n}}\right\}, \tag{III.32}$$

and if the following holds for a specific n

$$|\xi_{n+1}(x,t)| \leq \nu_{n+1} \leq \nu \tag{III.33}$$

then,

$$|\xi_n| < \frac{\nu_{n+1}}{\left[1 - |n\sigma_{n+1}|\nu_{n+1}^n\right]^{\frac{1}{n}}} \tag{III.34}$$

Now one can choose ν_{n+1} in a way such that

$$\nu_n = \frac{\nu_{n+1}}{\left[1 - |n\sigma_{n+1}|\nu_{n+1}^n\right]^{\frac{1}{n}}} \Rightarrow \nu_{n+1} = \frac{\nu_n}{\left[1 + |n\sigma_{n+1}|\nu_n^n\right]^{\frac{1}{n}}} \leq \nu_n \tag{III.35}$$

where ν_n is defined as below

$$|\xi_n(x,t)| \leq \nu_n < \nu \qquad \text{(III.36)}$$

Therefore we conclude

Theorem III: If we denote the mimimum of the expression $\left[\left|\frac{1}{n\sigma_{n+1}}\right|\right]^{\frac{1}{n}}$ $n=1,2,..$ by ν, and for a finite fixed N, the approximant ξ_N remains smaller than ν in absolute value, then all higher order approximants behave in the same way.

The interpretation of this theorem is as follows. If the system is globally normal then the limit of the sequence of approximants $\xi(x,t) = \lim_{N=\infty} \xi_N$ will remain permanently in the main clean region.

In the proofs of these theorems we assumed that f(x) is a real function and x is a real variable. We did this for the sake of simplicity. However, if f(x) and x are assumed to be complex quantities, nothing will change except the replacement of f_1 with R $\exp(f_1)$ and changing the intervals into the circles.

IV. GENERALIZATION TO THE MULTIDIMENSIONAL CASE

In the companion paper[1] establishing the approximants it was noted that there is a degree of flexibility in the order of the elementary factors or propagators associated with a multi-dimensional Lie transformation. A convenient ordering for the proof of convergence can be written as follows

$$Q = \exp(t\underset{\sim}{f}(\underset{\sim}{x})\cdot\nabla)=\exp(t\ x^{T}\cdot f^{T(1)}\cdot\nabla)$$

$$\left\{\prod_{j=0}^{\infty}\exp\left[\mu_j^{(1)}x_1^j\ \frac{\partial}{\partial x_1}\right]\right\}\ \cdot\ \cdot\ \cdot\ \cdot\ \left\{\prod_{j=0}^{\infty}\exp\left[\mu_j^{(N)}\ x_N^j\ \ \frac{\partial}{\partial x_N}\right]\ \right\} \tag{IV.1}$$

where $\mu_j^{(N)}$ depends on x_n's except x_N and t.

We have chosen an ordering of a product of elementary exponential operators such that the differentiation with respect to x_N is effected first. This ordering has a practical implication if we consider the effect of Q on x_1, in which case the last $(N-1)$ curly bracketed operators reduce to unity due to the fact that they have no effect on x_1

$$Qx_1 = \exp[t\ \underset{\sim}{x}^{T}.\underset{\sim}{f}^{T(1)}\cdot\nabla]\ \left\{\prod_{j=0}^{\infty}\exp\left[\mu_j^{(1)}x_1^j\ \frac{\partial}{\partial x_1}\right]\right\}\ x_1$$

Similarly, if we deal with Qx_j, then we can choose the ordering or the curly bracketed operators in a way such that

$$Qx_j = \exp[t\ \underset{\sim}{x}^{T}.\underset{\sim}{f}^{T(1)}\cdot\nabla]\ \left\{\prod_{j=0}^{\infty}\exp\left[\mu_k^{(j)}x_j^k\ \frac{\partial}{\partial x_j}\right]\right\}\ x_j \tag{IV.2}$$

can be written. Such changes of ordering will alter the μ_j's and without any loss of generality we may consider the particular ordering in Eq. (IV.1).

To find, for example, $\mu_0^{(1)}(x_2,\ldots x_N,t)$ we can obtain a partial differential equation which must satisfy

$$\frac{\partial \mu_0^{(1)}}{\partial t} = \bar{f}_1\left[-\mu_0^{(1)},\ x_2,\ldots,x_N\right] + \sum_{j=2}^{N} \bar{f}_j\left[-\mu_0^{(1)},x_2,\ldots,x_N\right]\ \frac{\partial \mu_0^{(1)}}{\partial x_j} \tag{IV.3}$$

where $\bar{f}_j$ denotes the new descriptive vector element of the system after extraction of its linear response. This may be equivalently stated as

$$|\underset{\sim}{f}(x)|_{\underset{\sim}{x}=0} = 0\ ,\quad \{\nabla f_j\}_{|x|=0}=0 \qquad j=1,2,\ldots N \tag{IV.4}$$

The same equations are assumed to hold for $\mu_0^{(1)}$

$$\mu_0^{(1)}\left[0,0,\ldots,t\right]=0\ ,\quad \left\{\nabla\mu_0\right\}_{|x|=0}=0 \tag{IV.5}$$

since first degree terms are excluded by extraction of the linear response. Hence, Eq. (IV.3) may be solved by a multi-dimensional Taylor series with the initial condition

$$\mu_0^{(1)}\ \left[x_2,\ldots,x_N,0\right]=0 \tag{IV.6}$$

The convergence of such series have been thoroughly investigated in the theory of partial differential equations[2]. Therefore, $\mu_0^{(1)}$, and the other μ's which satisfy the same kind of

partial differential equations can be assumed convergent and bounded in a closed domain around the (n-1)-tuple manifold formed by the cartesian product of the $x_2,\ldots,x_N$-complex planes.

In analogy with the previous section one may prove theorems about the convergence properties of the sequence of approximants generated by truncating the product of operators in Eq. (IV.2). These same type of statements follow as before except through a change of the x-plane into an n-tuple manifold constructed by the cartesian product of the n-complex planes (x_1-plane,..., x_N-plane).

V. CONCLUDING REMARKS

In the first of these two papers we presented a factorization scheme for Lie transformation evolution operators and in the present paper we have given sufficient conditions for the convergence of the scheme. Under appropriate circumstances, these approximants form a practical tool to produce a rapidly convergent and high precision approximation to the original evolution operator. These new approximants are also richer than, for example, Padé approximants for numerical analysis. This comment follows due to the abundance of branch points which make it possible to characterize many types of functions having various types of singularities. These two papers are actually only the first step in the theoretical development of these new types of approximants and much additional research needs to be done for their deeper understanding and to bring them to a truly practical level.

ACKNOWLEDGEMENTS

The authors acknowledge support for this work from the Office of Naval Research and the Air Force Office of Scientific Research.

REFERENCES

1. M. Demiralp and H. Rabitz "Factorization of Certain Evolution Operators Using Lie Operator Algebra - Formulation of the Method", to be published.

2. P.R. Garabedian, "Partial Differential Equations", John Wiley & Sons, New York, 1962

FACTORIZATION OF CERTAIN EVOLUTION OPERATORS USING LIE ALGEBRA: FORMULATION OF THE METHOD

Metin Demiralp* and Herschel Rabitz
Princeton University
Department of Chemistry
Princeton, New Jersey 08544, USA

*Permanent Address: Tübitak Research Institute for Basic Sciences
Applied Mathematics Department
P.O. Box 74, 41470
Gebze - Kocaeli, Turkey

Abstract

In this work a new method to factorize certain evolution operators into an infinite product of simple evolution operators is presented. The method uses Lie operator algebra and the evolution operators are restricted to exponential form. The argument of these forms is a first order linear partial differential operator. The method has broad applications including to the areas of sensitivity analysis, the solution of ordinary differential equations and the solution of Liouville's equation. A sequence of ξ-approximants are generated to represent the Lie operators. Under certain conditions the convergence rate of the ξ-approximant sequences is remarkably high. This work only presents the general formulation of the scheme and some simple illustrative examples. Investigation of convergence properties is given in a companion paper.

I. Introduction

In this paper a system with n-degrees of freedom will be characterized by n variables, $x_1,\ldots,x_n$, which form real Euclidean space. If any two points in the space are related by a unique transformation Q whose functional structure does not depend on the location of the points, then one can define an evolution operator for the system. Since any two points of an n-dimensional space may be connected by a continuous curve, it is possible to use a tracing parameter which defines the position of the system on this curve during its evolution from its initial state $\underset{\sim}{x}_i$ to its final state $\underset{\sim}{x}_f$. This circumstance often arises where time is the evolutionary parameter and we will accordingly denote the parameter as t. Therefore, the initial and final states of the system can be characterized by the scalar instants of time t_i and t_f. Hence the evolution operator Q can be represented as $Q(t_f,t_i)$ and

$$\underset{\sim}{x}_f = Q(t_f,t_i) \cdot \underset{\sim}{x}_i \qquad (I.1)$$

where the dot is used to symbolically represent the effect of Q on $\underset{\sim}{x}_i$. In many applications one can find practical expressions for the operator Q if t_i and t_f are sufficiently close to each other. Hence, the global evolution operator $Q(t_f,t_i)$ may be factorized into a simple sequence of evolutionary steps

$$Q(t_f,t_i) = Q(t_f,t_m) \cdot Q(t_m,t_{m-1})\ldots Q(t_1,t_i) \qquad (I.2)$$

and by choosing m sufficiently large this factorization can characterize the global evolution of the system. If the simple short time interval solutions were exactly calculable,

then the presence of a large number of such evolutions, m, is not important. However, in reality, even the simple evolutions over the short time intervals can often be only approximately determined. In such a case, the number of increments m is significant since errors can accumulate to possibly create numerical instabilities and inaccuracies. In addition, the factorization requires operators at times other than the initial and final specified values. Therefore, a more global factorization of the evolution operator such as suggested in this paper would be more attractive.

The present work considers the factorization of the evolution operator into a sequence of simple <u>global</u> evolution operators. The scheme presented will maintain its validity only on a special subclass of evolution operators. First, we restrict the system under consideration to being autonomous such that the evolution operator has the following simple structure

$$Q(t_f,t_i) = Q(t_f-t_i) \tag{I.3}$$

We also restrict ourselves to autonomous evolution operators having an exponential form

$$Q(t_f-t_i) = e^{(t_f-t_i)S} \tag{I.4}$$

where S denotes a time-independent operator. An important class of evolution operators is included in the following definition

$$S = L = \sum_{j=1}^{N} f_j(x_1,\ldots,x_N)\,\frac{\partial}{\partial x_j} \tag{I.5}$$

where the dimension or number of degrees of freedom of the system may be finite or infinite. The finite dimensional case may be directly related to the corresponding initial value problem produced by the set of ordinary differential equations[1], $\dot{x}_j = f_j(x_1, x_2, \ldots, x_N)$. Since almost every partial differential equation with initial conditions can be cast into an infinite set of ordinary differential equations through an appropriately chosen basis set expansion, we may consider the Lie operator in Eq. (I.5) as capable of treating a wide class of problems. Some caution is still required since the coefficients in Eq. (I.5) are scalars while some formal reductions of partial differential equations to ordinary differential equations can produce matrix coefficients. In summary, we restrict ourselves to operators having the structure of Eq. (I.5) and of finite order N.

Lie operators also arise in other areas besides that mentioned above. For example, the investigation of analytic simplectic maps[2] and the description of the behavior of trajectories near a reference trajectory for a general Hamiltonian system[3] are also other applications. This latter work is distinct from the present paper where we seek a global approximation to the evolution operator that is valid within a region of space without regard to a reference trajectory. In addition, our approximation of factorizing the exponential evaluation operator into a product sequence of global operators is different from that developed before. Recent additional works[4-6] have considered the use of Lie transformations to perform parameter space mapping of the solution of ordinary differential equations.

space mapping of the solution of ordinary differential equations. Other applications may also be found.

The remainder of this paper is organized in the following fashion. Section II gives the general formulation of the global factorization for one-dimensional systems followed by a generalization to multi-dimensional systems in Section III. Some illustrative examples are treated in Section IV and concluding remarks are given in Section V.

II. Factorization Procedure in the One-Dimensional Case

Lie exponential evolution operators defined by Eqs. (I.4) and (I.5) frequently arise in many applications. One application that was mentioned above arises in the treatment of ordinary differential equations. In particular, if we can evaluate the effect of the Lie transformation

$$Q = e^{tL} \quad ; \; L = \underset{\sim}{f}(\underset{\sim}{x}) \cdot \underset{\sim}{\nabla} \tag{II.1}$$

on the position vector $\underset{\sim}{x}$ around a point $\underset{\sim}{a}$ in the phase space of a system defined by

$$\dot{\underset{\sim}{x}} = \underset{\sim}{f}(\underset{\sim}{x}) \; , \tag{II.2}$$

then the solution to these equations may be written as

$$\underset{\sim}{x}(\underset{\sim}{a},t) = \left[e^{tL}\, \underset{\sim}{x}\right]_{\underset{\sim}{x}=\underset{\sim}{a}} \tag{II.3}$$

where $\underset{\sim}{x}$,$\underset{\sim}{a}$ and $\underset{\sim}{\nabla}$ are defined in the following manner

$$\underset{\sim}{x}^{T} = [x_1, x_2, \ldots, x_N] \tag{II.4}$$

$$\underset{\sim}{a}^T = [a_1, a_2, \ldots, a_N] \tag{II.5}$$

$$\underset{\sim}{\nabla}^T = \left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_N}\right] \tag{II.6}$$

This relation between the solution of ordinary differential equations and Lie transformations may conversely be used to determine the action of the operator Q on the position vector by solving the following ordinary differential equation

$$\dot{\underset{\sim}{\xi}}(\underset{\sim}{x},t) = \underset{\sim}{f}(\underset{\sim}{\xi}) \quad ; \quad \underset{\sim}{\xi}(\underset{\sim}{x},0) = \underset{\sim}{x} \tag{II.7}$$

This approach to determining Q is generally not preferable since Eq. (II.7) is often only soluble by elaborate numerical techniques which will hide the important structure of the desired transformation. Although the approach pursued here is also approximate, it will still leave the structure of the evolution operator rather apparent.

In order to appreciate the approach taken below, we recall some important properties of Lie transformations

$$e^{tL}[f(\underset{\sim}{x})g(\underset{\sim}{x})] = [e^{tL}f(\underset{\sim}{x})]\,[e^{tL}g(\underset{\sim}{x})] \tag{II.8}$$

$$e^{tL}f(\underset{\sim}{x}) = f(e^{tL}\underset{\sim}{x}) \tag{II.9}$$

The first of these equations states that a Lie transformation on a product of two functions f(x), g(x) can be factorized to the product of the Lie transformation on the individual functions. This property is due to the exponential structure of the Lie transformation along with application of the Leibnitz rule of differentiation, and the relation is valid provided that f and g are infinitely differentiable functions. The penetration

property in Eq. (II.9) also follows due to the particular structure of the Lie transformation and the assumed infinitely differentiable nature of the function f. Finally, one additional well known property of Lie transformations concerns the special case of the translation operator

$$e^{t\underset{\sim}{a}\cdot\underset{\sim}{\nabla}} f(\underset{\sim}{x}) = f(\underset{\sim}{x} + t\ \underset{\sim}{a}) \qquad \text{(II.10)}$$

which followed from a simple Taylor expansion of the right hand side.

We now desire to investigate the factorization of Lie transformations for one-dimensional systems. Although the one-dimensional nature of the problem makes it formally rather simple, this case also provides the best means to develop the factorization scheme presented here. In this case the Lie transformation can be written as

$$Q = \exp\left[tf(x)\ \frac{\partial}{\partial x}\right] \qquad \text{(II.11)}$$

where f(x) may have a number of zeros with one assumed to exist at the origin of the complex x-plane. This assumption about the location of a zero of f(x) at the origin does not create any loss of generality since a simple translation can bring one of the zeros of f(x) to the origin. The assumption about the existence of at least one zero of f(x) is more restrictive. However, in problems where f(x) forms the right hand side of an ordinary differential equation, there will usually be at least one stationary point for the solution. Therefore, the assumption about the existence of a zero of the function f(x) may be regarded as a minor loss of generality.

We may now make the additional assumption that the function f(x) may be expanded in a Taylor series

$$f(x) = \sum_{j=1}^{\infty} f_j x^j \qquad |x| < \rho \qquad \text{(II.12)}$$

where the expansion coefficients f_j are taken as known from the definition of the system. The expansion above implies that the system is well characterized, at least in a restrictive domain around the origin of the complex x-plane. We seek the factorization of the evolution operator Q such that every factor has an independent contribution in a fashion analogous to each term in the Taylor series of Eq. (II.12). To this end we define the flexible factorized structure

$$Q = \exp\left[tf(x)\frac{\partial}{\partial x}\right] = \prod_{j=1}^{\infty} \exp\left[\sigma_j(t)x^j\frac{\partial}{\partial x}\right] \qquad \text{(II.13)}$$

where $\sigma_j(t)$ are arbitrary at this point and yet to be determined. Equation (II.13) is the factorization formula for the one-dimensional case.

For the one-dimensional case the factorization in Eq. (II.13) may seem to be unnecessary due to the fact that the equation $\dot{x} = f(x)$ can be solved by the usual techniques of numerical analysis. However, in order to gain insight into the more interesting multi-dimensional case, the present reduced case presents the best way to understand the theory. Despite the existence of some attempts to factorize Q by time ordering techniques with respect to t, to our knowledge there has been no factorization of Q along the lines presented in Eq. (II.13) except Dragt's work for a different purpose and in a different framework.

Assuming that (II.13) holds and the coefficients σ_j are known, it is a simple matter to determine the effect of the operator Q on x. For this purpose we can investigate the individual effects of the factors in Eq. (II.13)

$$Q^{(j)}x = \exp\left[\sigma_j(t)x^j\frac{\partial}{\partial x}\right] x \tag{II.14}$$

By using a simple variable transformation

$$y = x^{-(j-1)} \tag{II.15}$$

we may write

$$Q^{(j)}x = \exp\left[-(j-1)\sigma_j(t)\frac{\partial}{\partial y}\right] y^{-1/(j-1)}, \quad j=2$$
$$Q^{(1)}x = \exp\left[\sigma_1(t)x\right] \tag{II.16}$$

and employ the translation operator property of Eq. (II.10) on the y-coordinate

$$Q^{(j)}x = \left[y - (j-1)\sigma_j(t)\right]^{-1/(j-1)} \tag{II.17}$$

or equivalently in terms of the x-coordinate

$$Q^{(j)}x = \frac{x}{\left[1 - (j-1)\sigma_j(t)x^{j-1}\right]^{1/(j-1)}} \tag{II.18}$$

where x and t are considered to be independent variables as we shall do so henceforth. In this formula the positive branch of the root has been taken. This is the fundamental formula of our factorization and it is valid provided the argument of the root appearing in Eq. (II.18) remains positive. We are now able to evaluate the effects of the individual factors in Eq. (II.13). In applications of Eq. (II.13) an approximation to Q would consist of truncating the infinite product involved.

At this point we need to determine the coefficient functions σ_j. To this end we can use the following relation

$$\frac{\partial Q}{\partial t} = \left[\sum_{j=1}^{\infty} f_j \, x^j \, \frac{\partial}{\partial x} \right] Q \quad ; \; Q(0) = I \tag{II.19}$$

which follows from Eqs. (II.12) and (II.13). If we now write

$$Q = Q^{(1)} Q_1 = \exp\left[\sigma_1(t) x \, \frac{\partial}{\partial x}\right] \, Q_1 \tag{II.20}$$

we may arrive at

$$\frac{\partial Q_1}{\partial t} = \exp\left[-\sigma_1(t) x \frac{\partial}{\partial x}\right] \left[f(x)\frac{\partial}{\partial x} - \dot{\sigma}_1 x \, \frac{\partial}{\partial x} \right] \exp\left[\sigma_1(t) x \, \frac{\partial}{\partial x}\right] \, Q_1 \tag{II.21}$$

using the properties in Eqs. (II.8) and (II.9). This result may be re-expressed as

$$\frac{\partial Q_1}{\partial t} = \left\{ \left[\frac{f\left[\exp\left[-\sigma_1 x \frac{\partial}{\partial x}\right] \, x\right]}{\exp\left[-\sigma_1 x \, \frac{\partial}{\partial x}\right] \, x} - \dot{\sigma}_1 \right] \, x \, \frac{\partial}{\partial x} \right\} \, Q_1 \tag{II.22}$$

The following formula

$$\exp\left[-\sigma_1(t) \; x \, \frac{\partial}{\partial x}\right] \, x = \exp(-\sigma_1(t)) \; x \tag{II.23}$$

allows for a rewriting of Eq. (II.22) utilizing the expansion in Eq. (II.12)

$$\frac{\partial Q_1}{\partial t} = \left\{ (f_1 - \dot{\sigma}_1) + (f_2 \exp(-\sigma_1(t)) \; x) + \ldots \right\} \, x \, \frac{\partial}{\partial x} \, Q_1 \tag{II.24}$$

The operator acting on Q_1 on the right hand side of Eq. (II.24) is a power series in x. Each of the terms of this series is independent and in the vicinity of the origin the first term will be dominant. We desire to make Q_1 as slowly varying as possible and therefore demand that the leading term in the series vanish for this purpose

$$\dot{\sigma}_1(t) = f_1 \; ; \; \sigma_1(0) = 0 \qquad \text{(II.25)}$$

The initial condition has been taken as zero to make the simple evolution operator $Q^{(1)}$ unitary. Equation (II.24) now has the form

$$\frac{\partial Q_1}{\partial t} = f^{(1)}(x)\, x^2 \frac{\partial}{\partial x} Q_1 \quad ; \quad Q_1(0) = I$$

where $f^{(1)}(x)$ can be identified from the remaining series of terms in the brackets of Eq. (II.24) and $f^{(1)}(0)$ is finite. Exactly this same logic may be put forth to evaluate $\sigma_2(t)$ by successively eliminating higher order powers of x in the differential equation. To construct a general recursion we assume knowledge of the first n of the σ_j's and write

$$Q = \left\{ \prod_{j=1}^{n} Q^{(j)} \right\} Q_n \qquad \text{(II.26)}$$

which suggests the equation

$$\frac{\partial Q_n}{\partial t} = f^{(n)}(x)\, x^{n+1} \frac{\partial}{\partial x} Q_n \quad ; \quad Q_n(0) = I \qquad \text{(II.27)}$$

The time-dependence of $f^{(n)}(x)$ is not explicitly shown and the function $f^{(n)}(x)$ is regular at the origin of the x-plane and is to be determined. We now may write

$$Q_n = Q^{(n+1)}\, Q_{n+1} = \exp\left[\sigma_{n+1}(t)x^{n+1}\frac{\partial}{\partial x}\right]\, Q_{n+1} \qquad \text{(II.28)}$$

and obtain

$$\frac{\partial Q_{n+1}}{\partial t} = \left[f^{(n)}\left[\exp\left[-\sigma_{n+1}(t)x^{n+1}\frac{\partial}{\partial x}\right]x\right] - \dot{\sigma}_{n+1}\right]x^{n+1}\,\frac{\partial}{\partial x}\,Q_{n+1} \qquad \text{(II.29)}$$

by utilizing again the properties in Eqs. (II.8) and (II.9). Employing the action of the factorization operator in (II.18) gives

$$\frac{\partial Q_{n+1}}{\partial t} = \left[f^{(n)}\left[\frac{x}{(1+n\sigma_{n+1}(t)\,x^n)^{1/n}}\right] - \dot{\sigma}_{n+1}\right]x^{n+1}\,\frac{\partial}{\partial x}\,Q_{n+1} \qquad \text{(II.30)}$$

We now apply logic analogous to that leading to Eq. (II.25) and eliminate the dominant contribution to the bracketed quantity multiplying the operator $x^{n+1}\,\frac{\partial}{\partial x}$ yielding

$$\dot{\sigma}_{n+1} = f^{(n)}(0) \quad ; \quad \sigma_{n+1}(0) = 0 \qquad \text{(II.31)}$$

where the initial value is again chosen as zero to make $Q^{(n+1)}$ unitary. Therefore we conclude

$$\frac{\partial Q_{n+1}}{\partial t} = f^{(n+1)}(x)\; x^{n+2}\,\frac{\partial}{\partial x}\,Q_{n+1} \qquad \text{(II.32)}$$

where

$$f^{(n+1)}(x) = \frac{1}{x}\left[f^{(n)}\left[\frac{x}{(1+n\sigma_{n+1}\,(t)\,x^n)^{1/n}}\right] - f^{(n)}(0)\right] \qquad \text{(II.33)}$$

This is a first order recursion relation with the initial condition

$$f^{(1)}(x) = \frac{1}{x^2}\left[f(\exp(-\sigma_1(t))\, x)\, \exp(\sigma_1(t)) - f_1 x\right] \qquad \text{(II.34)}$$

All of the σ-functions can be evaluated analytically in principle, however this is a tedious task and the use of a symbolic programming language such as MACSYMA or REDUCE is recommended. The first five of the σ-functions are given below.

$$\sigma_1(t) = f_1 t \qquad \text{(II.35)}$$

$$\sigma_2(t) = f_2\, g_1(t) \qquad \text{(II.36)}$$

$$\sigma_3(t) = f_3\, g_2(t) \qquad \text{(II.37)}$$

$$\sigma_4(t) = \left[f_4 + \frac{f_2 f_3}{f_1}\right] g_3(t) - \frac{f_2 f_3}{f_1}\, g_2(t) \qquad \text{(II.38)}$$

$$\sigma_5(t) = \left[f_5 + \frac{f_2 f_4}{f_1} + \frac{f_2^2 f_3}{2 f_1^2}\right] g_4(t) +$$

$$+ \left[\frac{f_2 f_4}{f_1} + \frac{f_2^2 f_3}{f_1^2}\right] g_3(t) + \frac{f_2^2 f_3}{2 f_1^2}\, g_2(t) \qquad \text{(II.39)}$$

where

$$g_n(t) = \frac{1-\exp(-n f_1 t)}{n f_1} \qquad \text{(II.40)}$$

We are now at a point to implement the factorization scheme. The essential approximation is to truncate Eq. (II.13) to a finite order thereby producing the following approximant.

$$\xi_n(x,t) = \left\{\prod_{j=1}^{n} Q^{(j)}\right\} x \qquad \text{(II.41)}$$

If the infinite product representation of Q given by (II.13) converges, then the following result will hold.

$$\xi(x,t) = Qx = \exp\left[tf(x)\frac{\partial}{\partial x}\right] x = \lim_{n=\infty} \xi_n \qquad \text{(II.42)}$$

Since the action of Q on x defines the fundamental operations of concern, we now focus our attention on the ξ-approximants. A recursion relation for these approximants can be obtained by first noting that

$$\xi_{n+1} = \left\{\prod_{j=1}^{n} Q^j\right\} \exp\left[\sigma_{n+1}(t)\ x^{n+1}\frac{\partial}{\partial x}\right] x \qquad \text{(II.43)}$$

An application of Eq. (II.18) yields

$$\xi_{n+1} = \left\{\prod_{j=1}^{n} Q^j\right\} \frac{x}{\left[1 - n\sigma_{n+1}\ x^n\right]^{\frac{1}{n}}} \qquad \text{(II.44)}$$

Since a product of Lie transformations is again a Lie transformation, we may use the property in Eq. (II.9) along with Eq. (II.41) to conclude that

$$\xi_{n+1}(x,t) = \frac{\xi_n(x,t)}{\left[1 - n\ \sigma_{n+1}(t)\ \xi_n^n(t)\right]^{1/n}} \qquad \text{(II.45)}$$

This is a rather simple first order recursion (difference equation) whose initial member is evaluated as follows

$$\xi_1(x,t) = \exp\left[\sigma_1(t)\ x\ \frac{\partial}{\partial x}\right] x = x\ \exp(\sigma_1(t)) = x\ \exp(f_1 t) \qquad \text{(II.46)}$$

Although this is a simple recursion relation, it is not typically suitable for numerical purposes. Numerical instabilities will

occur if f_1 is negative resulting in excessively small quantities for large times t or also under the conditions that x tends to zero. In these cases, error accumulations may occur due to the truncated arithmetic on the computer. To prevent this error we may renormalize the ξ-approximants and define a new recursion relation

$$\xi_{n+1} = \frac{\xi_n}{\left[1 - \bar{\sigma}_{n+1} x^n \xi_n^n\right]^{\frac{1}{n}}} \quad ; \quad \xi_1 = 1 \tag{II.47}$$

where

$$\bar{\sigma}_{n+1} = n\sigma_{n+1} \exp(nf_1 t) \tag{II.48}$$

The relation between the new approximants and the previous ones is

$$\tilde{\xi}_n(x,t) = \xi_n(x,t) \; x \; \exp(f_1 t) \tag{II.49}$$

which also implies

$$\tilde{\xi}(x,t) = \exp\left[tf(x)\frac{\partial}{\partial x}\right] x = \xi(x,t) \; x \; \exp(f_1 t) \tag{II.50}$$

Since the term $x \exp(f_1 t)$ characterizes the linear response of the system, we can consider $\xi(x,t)$ as a function measuring the deviations of the system from its linear response. We will accordingly refer to ξ as a "deviation function". As can be easily seen, the $\tilde{\xi}$ and ξ-approximants have branch points which move on trajectories in the x-plane. The location of these trajectories determines the convergence regions of the approximants. We shall leave the discussion of this issue and a comparison of the ξ-approximants with Padé approximants to a companion paper.

III. GENERALIZATION OF THE FACTORIZATION SCHEME TO THE MULTIDIMENSIONAL CASE

The logic put forth in section II for a systematic factorization of one dimensional evolution operators may now be generalized to multidimensional cases. In this situation, the evolution operator acting in a space of dimension N has the form

$$Q = \exp(t\, \underset{\sim}{f}(\underset{\sim}{x})\cdot\underset{\sim}{\nabla}) \tag{III.1}$$

where

$$\underset{\sim}{x}^{T} = [x_1,\ldots,x_N] \tag{III.2}$$

$$\underset{\sim}{\nabla}^{T} = \left[\frac{\partial}{\partial x_1},\ldots,\frac{\partial}{\partial x_N}\right] \tag{III.3}$$

$$\underset{\sim}{f}^{T} = [f_1(\underset{\sim}{x}),\ldots, f_N(\underset{\sim}{x})] \tag{III.4}$$

The function $\underset{\sim}{f}(\underset{\sim}{x})$ is assumed to have a zero at the origin

$$\lim_{|\underset{\sim}{x}|\to 0} \underset{\sim}{f}(\underset{\sim}{x}) = 0 \tag{III.5}$$

and it is also assumed to be expandable in a multidimensional Taylor series in the variable $x_1,\ldots x_N$. This latter expansion can be written in tensor form as

$$f_i = f_{ij}^{(1)}\, x_j + f_{ijk}^{(2)}\, x_j x_k + f_{ijkl}^{(3)}\, x_j x_k x_l + \ldots \tag{III.6}$$

where the convention of the explicit summation over repeated indices is used for convenience.

In the one-dimensional case the operator $\exp(\sigma_1 x\frac{\partial}{\partial x})$ played a fundamental role in the first step of establishing a recursion relation for the approximants. The same situation occurs again

here and we shall denote this first degree operator Q_L as taking on the following form

$$Q_L = \exp(\underset{\sim}{x}^T \underset{\sim}{\sigma}^{(1)} \underset{\sim}{\nabla}) \tag{III.7}$$

where $\underset{\sim}{\sigma}^{(1)}$ is a square matrix or equivalently a second degree tensor. The effect of this operator on the position vector $\underset{\sim}{x}$ is

$$Q_L \underset{\sim}{x} = \exp(\underset{\sim}{\sigma}^{T(1)}) \quad \underset{\sim}{x} \tag{III.8}$$

Since $Q_L \underset{\sim}{x}$ must be the system linear response we can conclude that

$$\sigma_{jk}{}^{(1)} = t\, f_{kj}^{(1)} \quad j,k = 1,\ldots,N \tag{III.9}$$

Henceforth, we shall denote the linear response of the system evolution by S,

$$S = \exp(t \underset{\sim}{f}^{(1)}) \tag{III.10}$$

Using the definition of the scalar product of two tensors of the same order,

$$A_{j_1 \ldots j_n} \; B_{j_1 \ldots j_n} = A \odot B \tag{III.11}$$

we can write the evolution operator in Eq. (III.1) as the infinite order factorized product

$$Q = Q_L \prod_{k=2}^{\infty} \exp[\sigma^{(k)} \odot L^{(k)}] \tag{III.12}$$

where $\sigma^{(k)}$ is a k-th order tensor to be be determined and $L^{(k)}$ is a tensor valued operator

$$L^{(k)}_{j_1 j_2, \ldots j_k} = x_{j_2} x_{j_3} \cdots x_{j_k} \; \frac{\partial}{\partial x_{j_1}} \tag{III.13}$$

The tensor product in the argument of each exponential term in Eq. (III.12) is itself a sum of operators which would be difficult to deal with in practice. Therefore, we have further factorized each individual term in Eq. (III.12) (except Q_L) to obtain

$$Q = Q_L \prod_{k=2}^{\infty} \prod_{j}^{*} \exp(\sigma^{(k)}_{j_1 j_2 \ldots j_k} \; L^{(k)}_{j_1 \ldots j_k}) \qquad \text{(III.14)}$$

where it is understood that the coefficient functions $\sigma^{(k)}$ are now distinct from the set in Eq. (III.12). The starred product in this formula means that the product operation is performed over the entire domain of the j-indices. There is no unique ordering to the factorization in Eq. (III.14) for a multi-dimensional case. However, if we define the following operators

$$\bar{Q}^{(n)} = \exp[\sigma^{(n)} \circledcirc L^{(n)}] \qquad \text{(III.15)}$$

$$Q^{(n)} = \prod_{j}^{*} \exp[\sigma^{(n)}_{j_1 j_2 \ldots j_n} \; L^{(n)}_{j_1 \ldots j_n}] \qquad \text{(III.16)}$$

one can prove that

$$\left\{\bar{Q}^{(n)} x\right\} - \left\{Q^{(n)} x\right\} = O\left[x^{2n-1}\right] \qquad \text{(III.17)}$$

Therefore, within this level of approximation the expressions in Eqs. (III.12) and (III.14) may be considered equivalent. The form given by Eq. (III.14) is practical since each of the sequence of evolution operators acts on a particular coordinate and degree of freedom.

The procedure for determining the σ - tensor is the same as in the previous section, however all scalars, (except time) must be replaced with tensor quantities and the conventional algebra must be replaced with tensor algebra. The details of these operations will not be dealt with further here, but the use of symbolic programming languages would be most helpful in practice. The second degree σ - tensor is given below as an example.

$$\sigma^{(2)}_{jkl} = \int_0^t S_{j\alpha}(\tau)\, f^{(2)}_{\alpha n_1 n_2} \left[S^{-1}(\tau)\right]_{n_1 k} \left[S^{-1}(\tau)\right]_{n_2 l} \, d\tau \qquad \text{(III.18)}$$

The evaluation of the ξ - approximants can again be accomplished by using the consecutive effects of the individual factors of the evolution operator. Symbolic programming techniques would likely be the best procedure for determining the $x_1,\ldots,x_N$ and t dependence of the ξ - approximants.

IV. ILLUSTRATIVE APPLICATIONS

In this section, five problems are considered, each of which exhibits different types of behavior. For the sake of comparison with the techniques introduced in the previous sections, we have chosen analytically soluble problems as described below.

1) $f(x) = 1 - e^x$ (IV.1)

From traditional linear stability analysis arguments this system is stable for $x>0$ and unstable for $x<0$. There is also only one steady state point located at the origin. An analytic expression for the effect of the Lie transformation on x can be written as

$$\xi(x,t) = \exp\left[tf(x)\frac{\partial}{\partial x}\right] x = -\ln\left[1-(1-e^{-x})\,e^{-t}\right] \qquad \text{(IV.2)}$$

A careful examination of the structure of $\xi(x,t)$ reveals that its branch point traverses the path from $-\infty$ to $+\infty$ along the horizontal axes $\mp i\Pi$ as time evolves. Figure 1a plots the exact deviation function $\xi(x,t)$ and its first five approximants $\xi_n(x,t)$ as defined in Eqs. (II.50) and (II.49) respectively for the case $x=0.1$. It is apparent that the approximants uniformly converge to the true deviation function as n increases. The error between the true deviation function and the $n=5$ approximant is shown in Figure 1b where it is apparent that the error decreases monotonically to an asymptotic value for large times. A similar pair of plots is shown in Figure 2

for x=5.0. At this larger value of x qualitatively similar behavior occurs but the rate of convergence of the approximants is slower and the peak in the error function may be a signal of the loss of global convergence. The situation for negative values of x is different as shown in Figure 3. Figure 3 presents the case for x=-1.0. The approximants in this case seem to show oscillatory nonmonotonic behavior with regard to true deviation function. The error of each of the approximants is qualitatively similar to that of Figures 1 and 2. At a sufficiently large negative value of x singular behavior shows up resulting in apparent non-convergence.

ii) $f(x) = 1 - e^{-x}$ (IV.3)

This system is unstable for positive x values due to the first Taylor coefficient being positive. It has only one steady state point located at the origin of the x-plane. The analytic expression of the Lie transformation effect on x is

$$\xi(x,t) = \ln\{1+(e^{x}-1)e^{t}\} \quad \text{(IV.4)}$$

The branch point trajectory of this system matches with the negative portion of the real axis of the x-plane. The branch point moves on this line towards the origin as time evolves and reaches there in the limit that $t\to\infty$. Figure 4 shows the deviation approximants and the error of the fifth member for x=0.1. Apparent convergence failure is observed, however during a finite time interval starting from t=0 there is temporary convergence.

iii) $f(x) = \sin x$ (IV.5)

This system is unstable around $x=0$ for $x>0$, however it has infinitely many steady state points and they alternatively make the system either stable or unstable. Figures 5 depicts the approximant behavior for the case $x=1.0$. There is apparent convergence behavior in the figure, however a peak in the error function may again be a signal of the loss of global convergence. It is difficult to prove this point from only a finite number of approximants. An additional calculation is shown in Figure 6 for $x = 5.0$ which is beyond the second stationary point of sinx. This well behaved nature of the approximants is probably due to the fact that all the branch points of this system are purely imaginary.

iv) Stakgold problem[7]

This problem is associated with the consideration of two coupled nonlinear differential equations with system coefficients given by

$$f_1(x_1,x_2) = \lambda x_1 - x_2 - x_1(x_1^2+x_2^2)$$

$$f_2(x_1,x_2) = \lambda x_2 - x_1 - x_2(x_1^2+x_2^2) \quad \text{(IV.6)}$$

The analytic expression for the effect of the Lie transformation on the position vector is

$$\xi_1(t)=(x_1\cos t - x_2\sin t)\, e^{-|\lambda|t}\, \eta(\underset{\sim}{x},t)$$

$$\xi_2(t)=(x_1\sin t + x_2\cos t)\, e^{-|\lambda|t}\, \eta(\underset{\sim}{x},t) \quad \text{(IV.7)}$$

where λ is assumed to be negative and η is defined as follows

$$\eta(x,t) = \left[1 + \frac{x_1^2 + x_2^2}{|\lambda|}\left[1 - e^{-2|\lambda|t}\right]\right]^{-\frac{1}{2}} \qquad \text{(IV.8)}$$

This system is stable as long as λ remains negative. In the case of positive λ the same condition again holds but the system does not have a steady state solution and a limit cycle appears.

In applying the method of Section III to Eq. (IV.6) we will find that the system has only the following non zero tensor coefficients

$$f_{11}^{(1)} = \lambda \qquad f_{12}^{(1)} = -1$$
$$f_{21}^{(1)} = 1 \qquad f_{22}^{(1)} = \lambda \qquad \text{(IV.9)}$$

$$f_{1111}^{(3)} = f_{1122}^{(3)} = f_{2211}^{(3)} = f_{2222}^{(3)} = -1 \qquad \text{(IV.10)}$$

Accordingly, the linear response term would be expressed by the tensor

$$\underset{\sim}{S} = \exp(t\underset{\sim}{f}^{(1)}) \qquad \text{(IV.11)}$$

and elements of this matrix and its inverse are given by the following expressions

$$S_{11} = e^{\lambda t}\cos t \qquad S_{12} = -e^{\lambda t}\sin t$$
$$S_{21} = e^{\lambda t}\sin t \qquad S_{22} = e^{\lambda t}\cos t \qquad \text{(IV.12)}$$

$$S_{11}^{(-1)} = e^{-\lambda t}\cos t \qquad S_{12}^{(-1)} = e^{-\lambda t}\sin t$$
$$S_{21}^{(-1)} = e^{-\lambda t}\sin t \qquad S_{22}^{(-1)} = e^{-\lambda t}\cos t \qquad \text{(IV.13)}$$

Since the second degree Taylor expansion coefficients are zero it follows that

$$\underset{\sim}{\sigma}^{(2)} = 0 \tag{IV.14}$$

The third order terms are non zero and $\underset{\sim}{\sigma}^{(3)}$ may be shown as

$$\sigma^{(3)}_{l_1 l_2 l_3 l_4} = \int_0^t S_{l_1 m_1} f^{(3)}_{m_1 m_2 m_3 m_4} S^{(-1)}_{m_2 l_2} S^{(-1)}_{m_3 l_3} S^{(-1)}_{m_4 l_4} \, d\tau \tag{IV.15}$$

where the explicit summation rule over repeated indices is employed. After some tedious algebra, one can show that all elements of the $\sigma^{(3)}$-tensor vanish except for the following four members

$$\sigma^{(3)}_{1111}(t) = \sigma^{(3)}_{1122}(t) = \sigma^{(3)}_{2211}(t) =$$

$$= \sigma^{(3)}_{2222}(t) = \frac{e^{-2\lambda t} - 1}{2\lambda} \equiv \sigma_3(t) \tag{IV.16}$$

This result immediately yields the tensor product

$$\underset{\sim}{\sigma}^{(3)} \odot \underset{\sim}{L}^{(3)} = \sigma_3(t) \left[x_1^2 + x_2^2\right] \left[x_1 \frac{\partial}{\partial x_1} + x_2 \frac{\partial}{\partial x_2}\right] \tag{IV.17}$$

As can be easily observed the operators $\underset{\sim}{\sigma}^{(3)} \odot \underset{\sim}{L}^{(3)}$ and $\underset{\sim}{f}^{(3)} \odot \underset{\sim}{L}^{(3)}$ commute and therefore there will be no contribution from higher degree terms of the remainder during the elimination of the operator $\underset{\sim}{\sigma}^{(3)} \odot \underset{\sim}{L}^{(3)}$ from the structure of L. In addition, there are no higher order terms than these already coming from the structure of L itself. Hence, we may conclude that the factorization exactly truncates at its second order terms if we retain $\underset{\sim}{\sigma}^{(3)} \odot \underset{\sim}{L}^{(3)}$ as a global second degree

Lie operator. Indeed, if we write

$$\underset{\sim}{Q}\underset{\sim}{x} = \exp\left[\underset{\sim}{x}^T \underset{\sim}{f}^{T(1)} \cdot \underset{\sim}{\nabla}\right] \exp\left[\underset{\sim}{\sigma}^{(3)} \odot \underset{\sim}{L}^{(3)}\right] \underset{\sim}{x}$$

$$= \exp\left[\underset{\sim}{x}^T \underset{\sim}{f}^{T(1)} \cdot \underset{\sim}{\nabla}\right] \frac{r}{\left[1 - 2\sigma_3(t)r^2\right]^{\frac{1}{2}}} \begin{bmatrix}\cos\theta \\ \sin\theta\end{bmatrix}, \qquad \text{(IV.18)}$$

then it follows that

$$\underset{\sim}{Q}\underset{\sim}{x} = \frac{e^{\lambda t}}{\left[1+(e^{2\lambda t}-1)\ \frac{x_1^2 + x_2^2}{\lambda}\right]^{\frac{1}{2}}} \begin{bmatrix}\cos t & -\sin t \\ \sin t & \cos t\end{bmatrix} \begin{bmatrix}x_1 \\ x_2\end{bmatrix} \qquad \text{(IV.19)}$$

and the exact result is obtained. In this result we have first used polar coordinates

$$r = \left[x_1{}^2 + x_2^2\right]^{\frac{1}{2}} \quad \cos\theta = x_1\left[x_1^2+x_2^2\right]^{-\frac{1}{2}} \quad ;$$

$$\underset{\sim}{\sigma}^{(3)} \odot \underset{\sim}{L}^{(3)} = \sigma_3(t) r^3 \frac{\partial}{\partial r} \qquad \text{(IV.20)}$$

and then returned to the cartesian representation in Eq. (IV.19).

The result in Eq. (IV.19) is just a confirmation of the operator algebra introduced earlier in the paper. We may still go a step further and factorize the evolution operator involving $\underset{\sim}{\sigma}^{(3)} \odot \underset{\sim}{L}^{(3)}$ to obtain

$$Q^{(2)} = \exp\left[\sigma_3(t)\ x_1^3 \frac{\partial}{\partial x_1}\right] \quad \exp\left[\sigma_3(t)\ x^2{}_1 x_2 \frac{\partial}{\partial x_2}\right]$$

$$\exp\left[\sigma_3(t)\ x_2^2\ x_1 \frac{\partial}{\partial x_1}\right] \quad \exp\left[\sigma_3(t)\ x_2^3 \frac{\partial}{\partial x_2}\right] \qquad \text{(IV.21)}$$

This different factorization creates an error which is of the order of magnitude of fifth degree terms. In this case, an infinite product appears which converges about the origin.

v) space extension

Consider the system defined by the function

$$f(x) = \sqrt{1-x^2} \qquad \text{(IV.22)}$$

This function has two zeros located at the points $x=+1$ and $x=-1$, however it does not fulfill the requirements for our method. In particular, it cannot be expanded in a Taylor series around these points. Nevertheless, the problem may still be approached by extending the space to two dimensions through the introduction of a new variable in addition to x as follows

$$y = \sqrt{1-x^2} \qquad \text{(IV.23)}$$

We can now define a new system with the descriptive functions

$$f_1(x,y) = y$$
$$f_2(x,y) = -x \qquad \text{(IV.24)}$$

This new system satisfies all of the necessary conditions for factorization. Therefore, in cases such as these, the technique of space extension may make it possible to factorize Lie transformations which otherwise might not admit to direct treatment.

V. CONCLUDING REMARKS

The basic thrust of this paper is the development of a new sequence of approximants appropriate for time evolution operators with Lie generator arguments. Although we have not given convergence theorems for the ξ-approximants sequences, the results in Section IV are encouraging. Rapid and highly accurate convergence seems to be obtained at least in a sufficiently closed vicinity to the origin. The next step in this work, examined in a companion paper, is the investigation of the ξ-approximant singularities and some convergence theorems.

Actual implementation of the factorization scheme, especially for multidimensional cases can involve a considerable amount of algebra. The use of symbolic programing on the computer would likely be a necessity in these cases, and this issue also needs further investigation for its practical implementation. A number of applications of the factorization may be envisioned as suggested in the introduction. Evolution operators of the type studied in this paper occur in a wide variety of problems, but perhaps the most obvious and simple application would be to the solution of ordinary differential equations. The possible attraction here follows from the fact that the approximants provide a global solution in time rather than the usual sequential time stepping procedures. A number of numerical issues need to be addressed for this case as well as other applications before the optimal realm of utility of the scheme may be established.

Acknowledgment

The authors would like to thank Professor Carl Wulfman for helpful comments. We also acknowledge support from the Office of Naval Research and the Air Force Office of Scientific Research.

Figure Captions

Figure 1 Plot of the exact deviation function $\xi(x,t)$ and its first five approximants $\xi_n(x,t)$, n=1,...5 for the characteristic function in Eq.(IV.1) where x=0.1 In figure a, the last three approximants are indistinguishable from the exact result. Figure b shows the deviation function for the approximant n=5. The same line masks in figure a will be used in the remaining ξ-approximant plots.

Figure 2. The same as figure 1, except x=5.0. These results are qualitatively similar to those of figure 1 except now the convergence rate is slower and there is a peak in the error function.

Figure 3. The same as figure 1, except now x = -1.0. Apparent oscillatory nonmonotonic behavior is exhibited with respect to the true deviation function in figure a.

Figure 4. Figure a exhibits the exact deviation function $\xi(x,t)$ and the first five approximants $\xi_n(x,t)$, n=1,...5 corresponding to the fundamental function in Eq. (IV.3) at x = 0.1. Figure b shows the error function for n=5 approximant. Apparent divergence behavior is observed at long time; however, during a finite interval around the origin there is temporary convergence. Here, only the fifth approximant goes to infinity. However, the first four approximants also have branch points

close to zero but they do not make the denominator in the corresponding transformation from the previous approximant zero when $t<10$. Some of the branch points are not even on the positive real axis.

Figure 5. The exact deviation function $\xi(x,t)$ and its first five approximants $\xi_n(x,t)$, $n=1,\ldots 5$ for the characteristic function in Eq. (IV.5) at $x=1.0$. The first and second approximants coincide as well as the third and fourth approximants. There is apparent convergence behavior in Figure a, however a peak in the error function in Figure b may signal a loss of global convergence.

Figure 6. The same as Figure 5, except that now $x=5.0$.

References

1. M. Demiralp, Bull. Tech. Univ., Instanbul, 37, 424-445, 1984.
2. A.J. Dragt and J.M. Finn, J. Math Phys., 17, 2215-2227, (1976).
3. A.J. Dragt and E. Forest, J. Math. Phys. 29, 2734-2744, (1983).
4. C. Wulfman and H. Rabitz, J. Phys. Chem, 90, 2269-2272 (1986).
5. L.M. Hubbard, C. Wulfman and H. Rabitz, J. Phys. Chem, 90, 2173-2280 (1986).
6. S. Shi and H. Rabitz, "Functional Transformation and Its Application in Study of Group Properties of Integro-Differential Equations", to be published.
7. R. Larter, H. Rabitz, M. Kramer, J. Chem. Phys., 80, 4120-4128, (1984).
8. I. Stakgold, Green's Functions and Boundary Value Problems" (Wiley, New York, 1979), p. 629.

Figure 1a

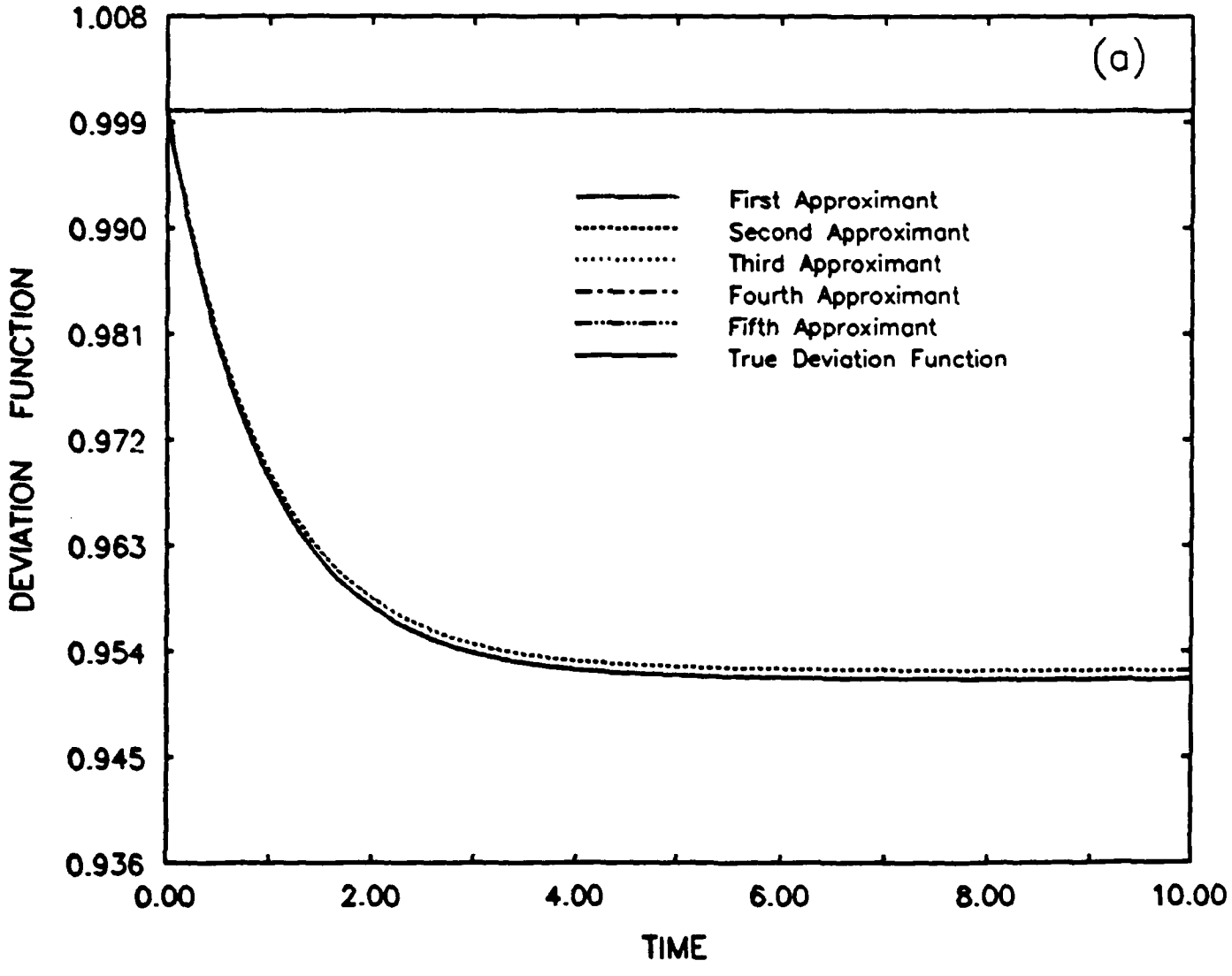

(a)
1.008
0.999
0.990
0.981
0.972
0.963
0.954
0.945
0.936
DEVIATION FUNCTION
0.00
2.00
4.00
6.00
8.00
10.00
TIME
First Approximant
Second Approximant
Third Approximant
Fourth Approximant
Fifth Approximant
True Deviation Function

Figure 1b

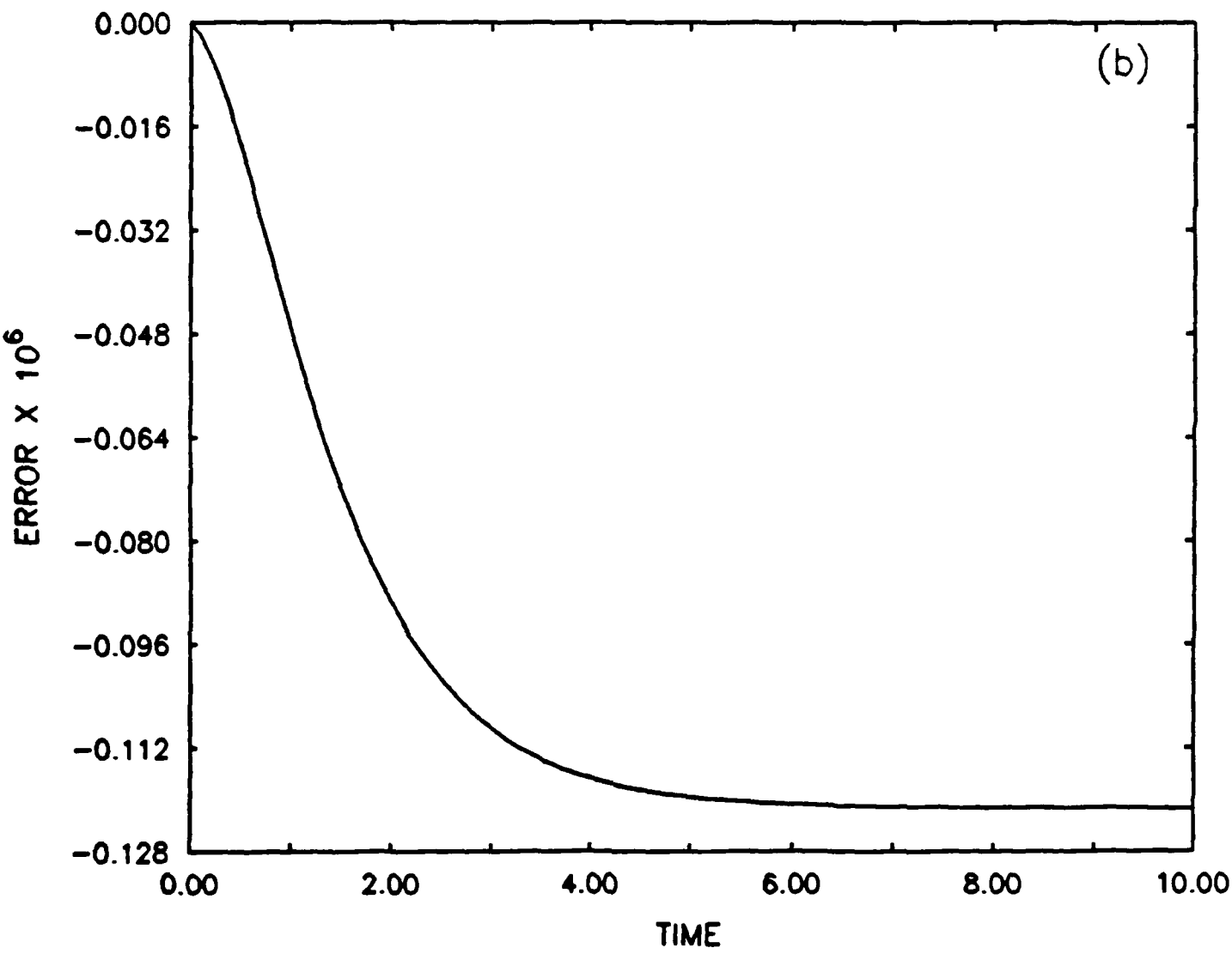

(b)
0.000
-0.016
-0.032
-0.048
-0.064
-0.080
-0.096
-0.112
-0.128
ERROR X 10^6
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 2a

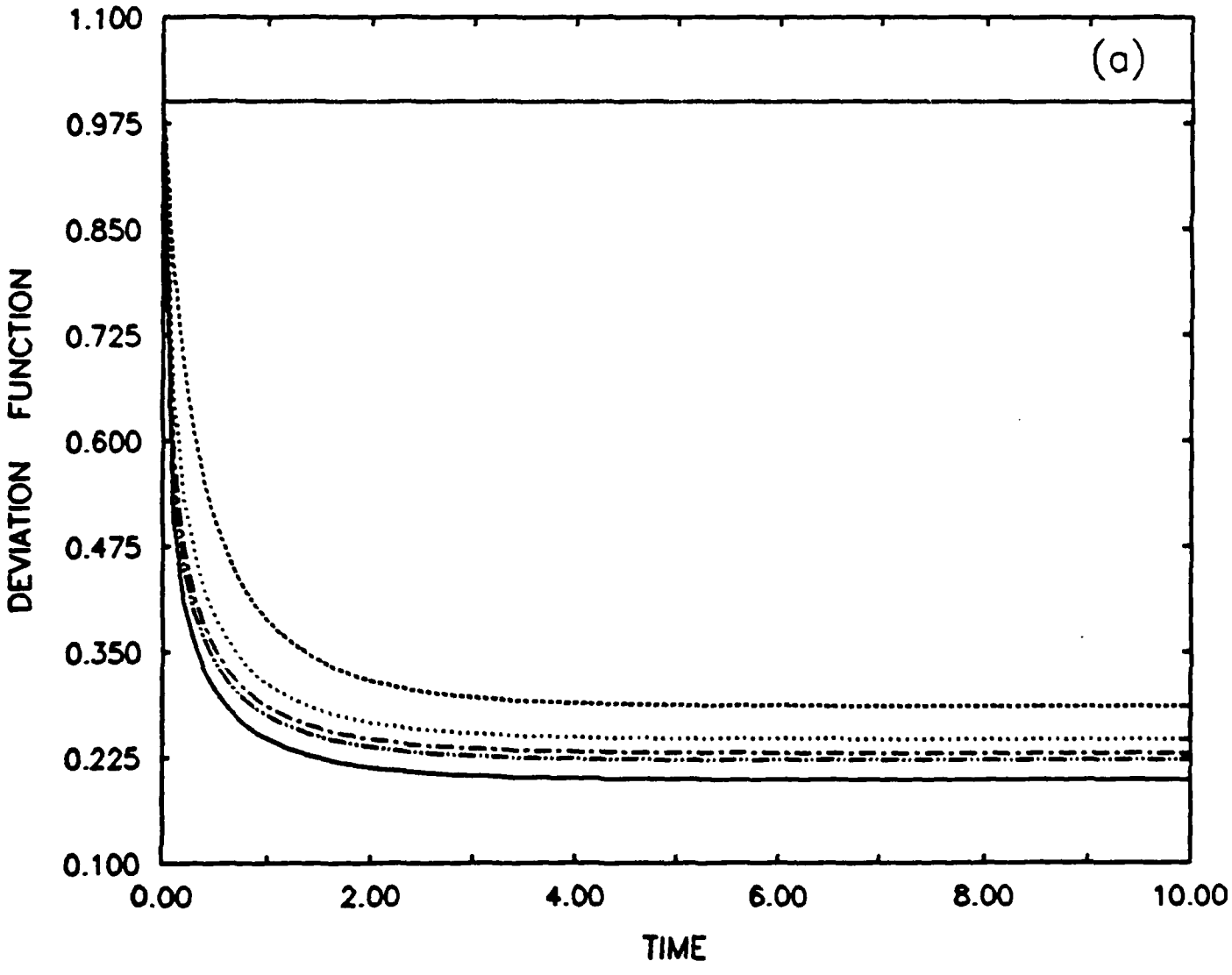

(a)
1.100
0.975
0.850
0.725
0.600
0.475
0.350
0.225
0.100
DEVIATION FUNCTION
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 2b

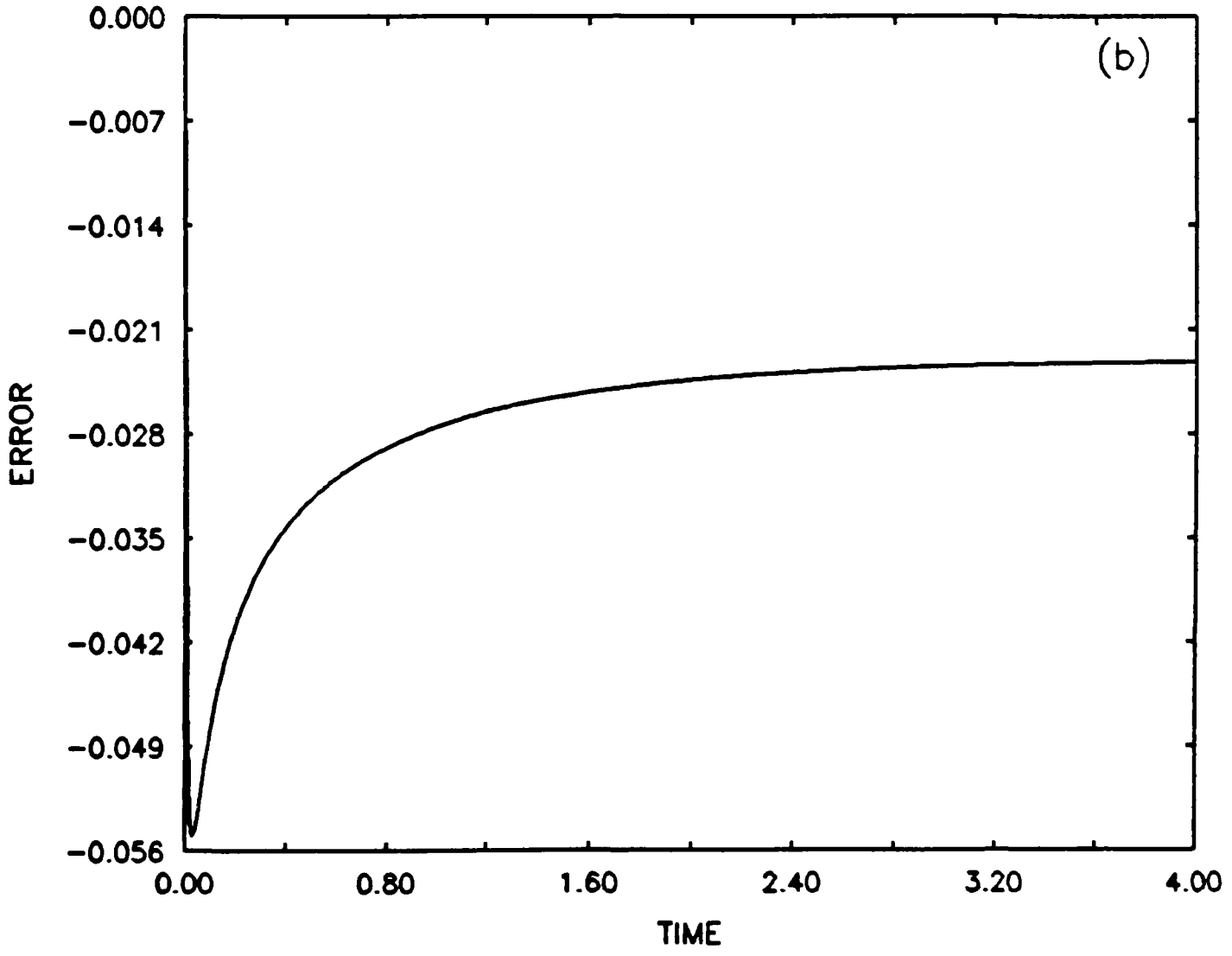

(b)
0.000
-0.007
-0.014
-0.021
-0.028
-0.035
-0.042
-0.049
-0.056
ERROR
0.00
0.80
1.60
2.40
3.20
4.00
TIME

Figure 3a

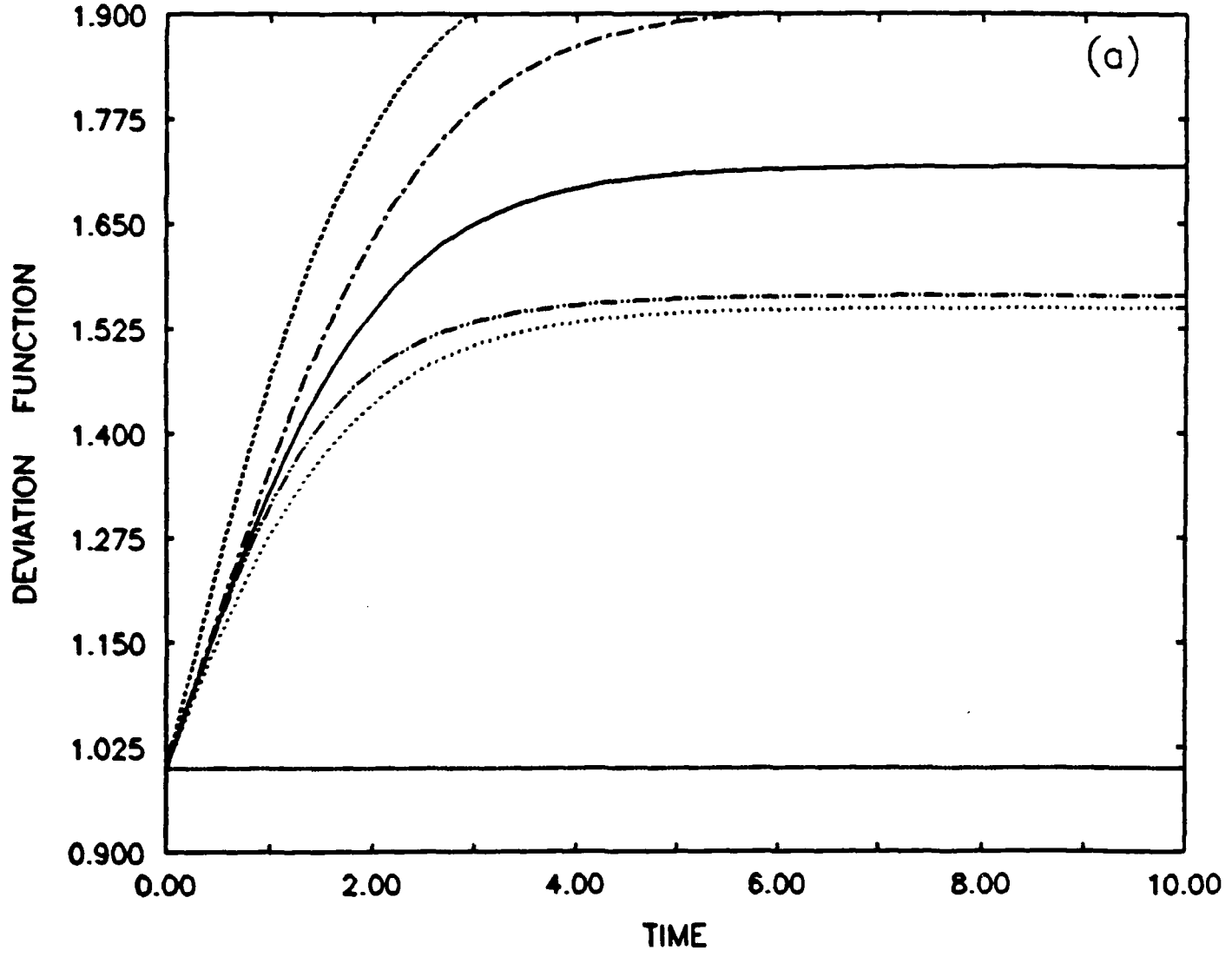

(a)
1.900
1.775
1.650
1.525
1.400
1.275
1.150
1.025
0.900
DEVIATION FUNCTION
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 3b

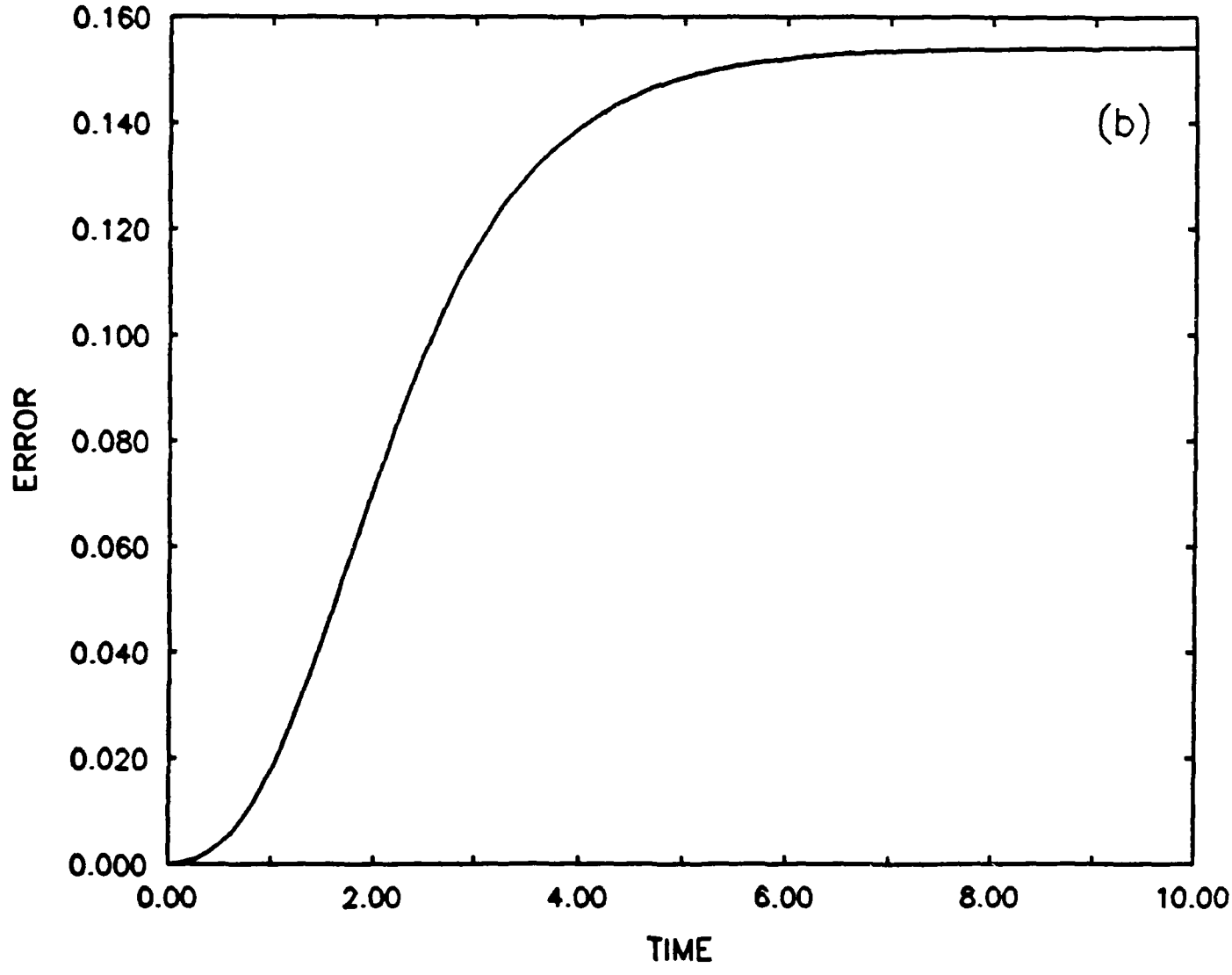

0.160
0.140
0.120
0.100
0.080
0.060
0.040
0.020
0.000
ERROR
(b)
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 4a

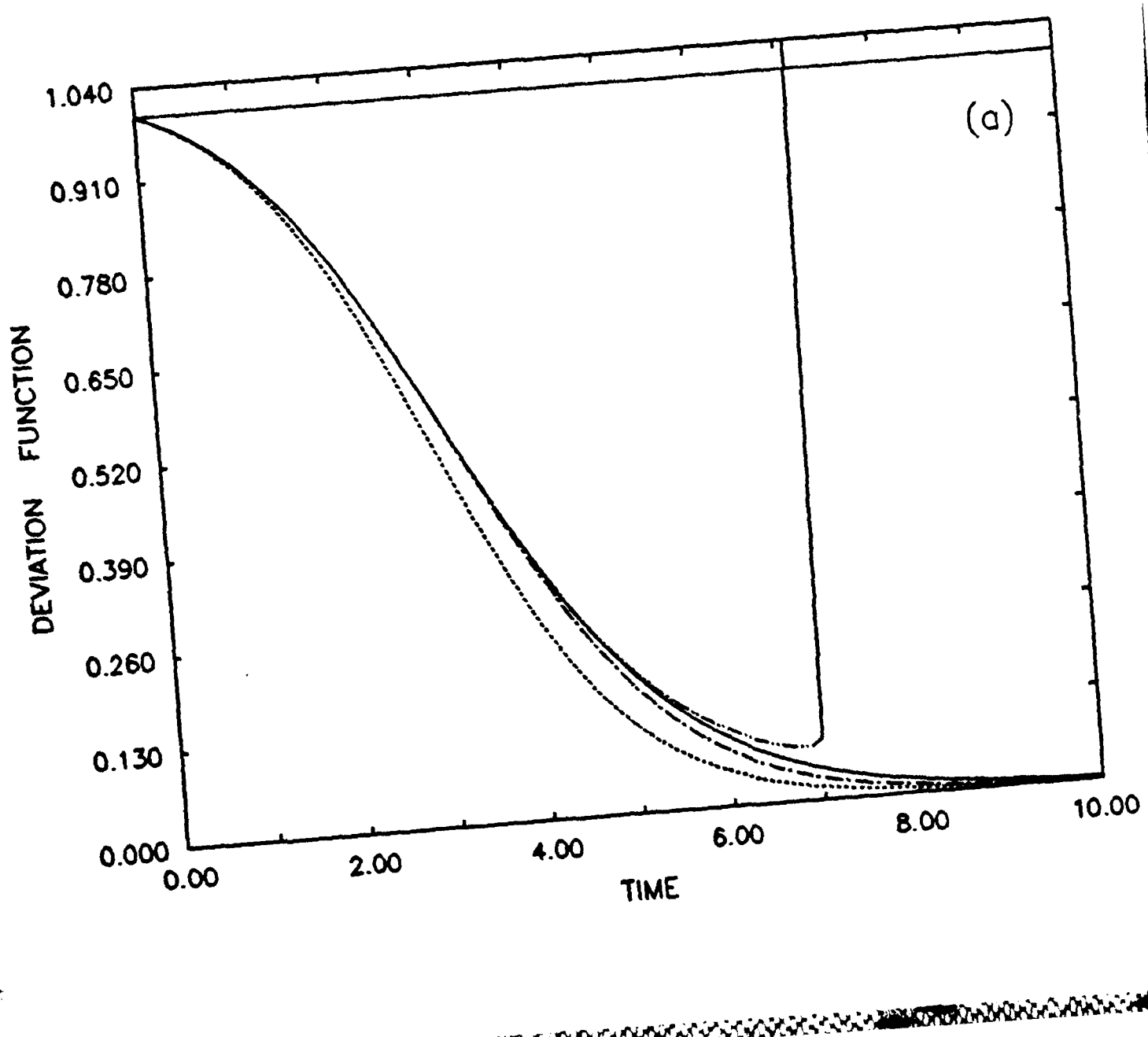

1.040
0.910
0.780
0.650
0.520
0.390
0.260
0.130
0.000
DEVIATION FUNCTION
(a)
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 4b

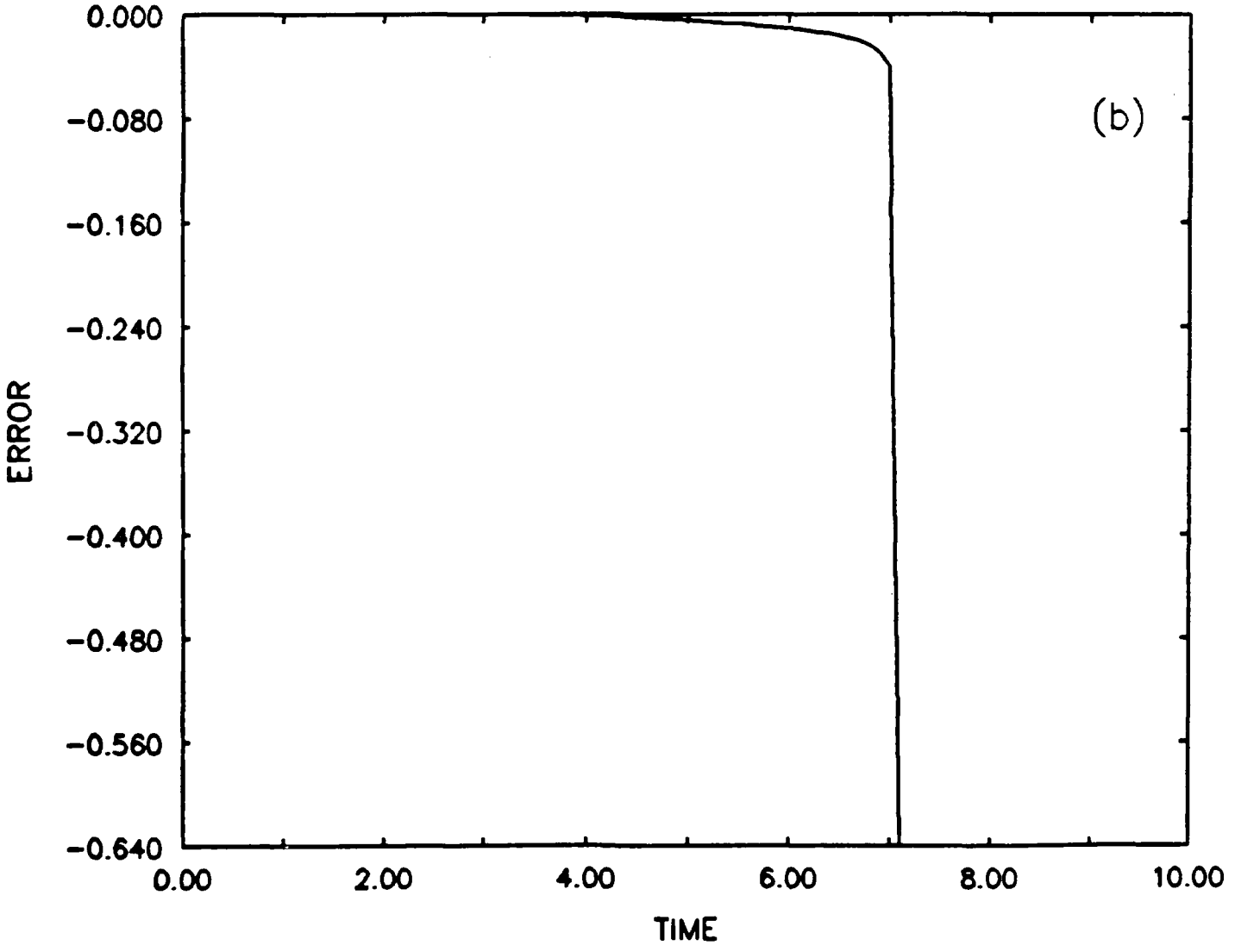

0.000
-0.080
-0.160
-0.240
-0.320
-0.400
-0.480
-0.560
-0.640
ERROR
(b)
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 5a

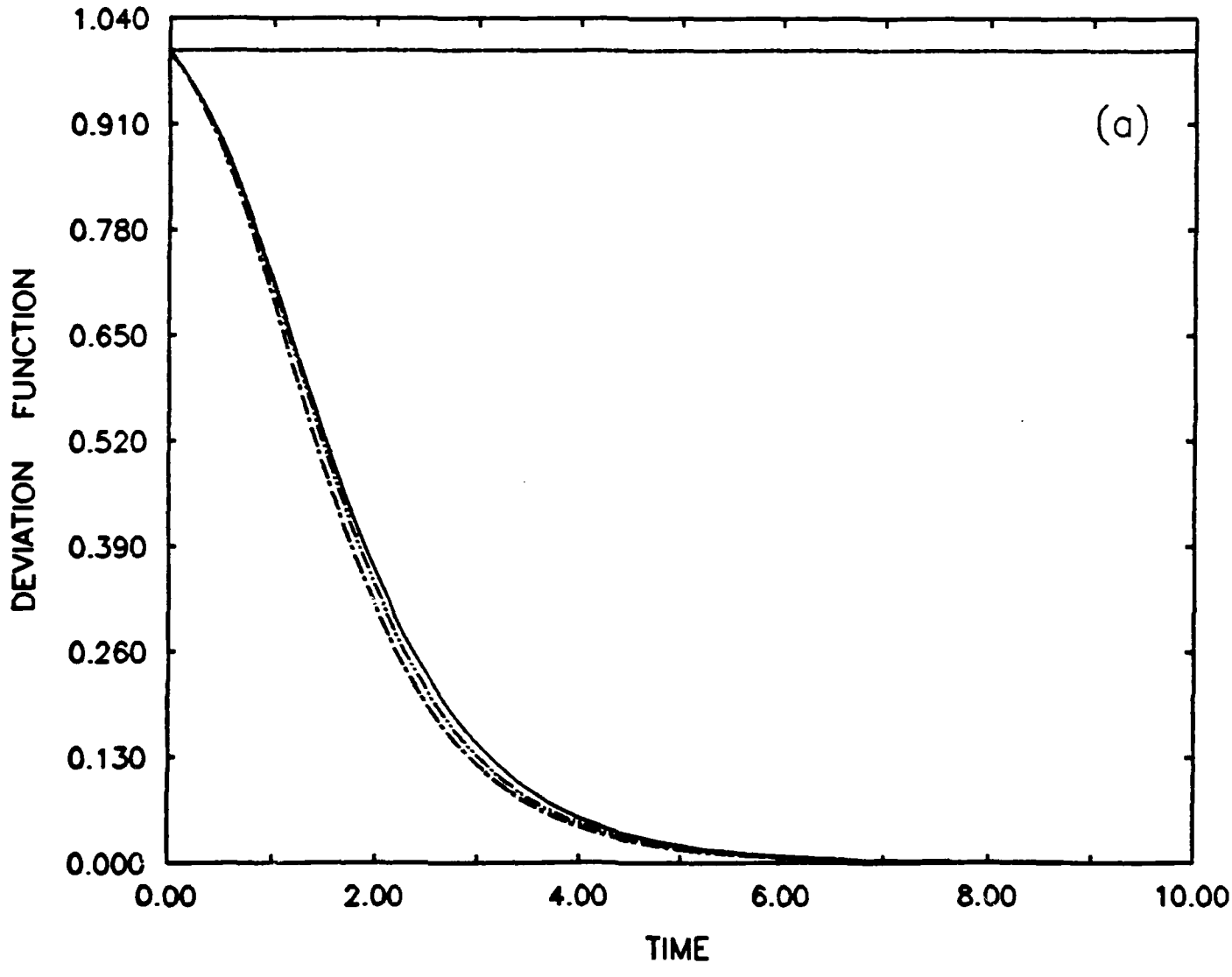

1.040
0.910
0.780
0.650
0.520
0.390
0.260
0.130
0.000
DEVIATION FUNCTION
(a)
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 5b

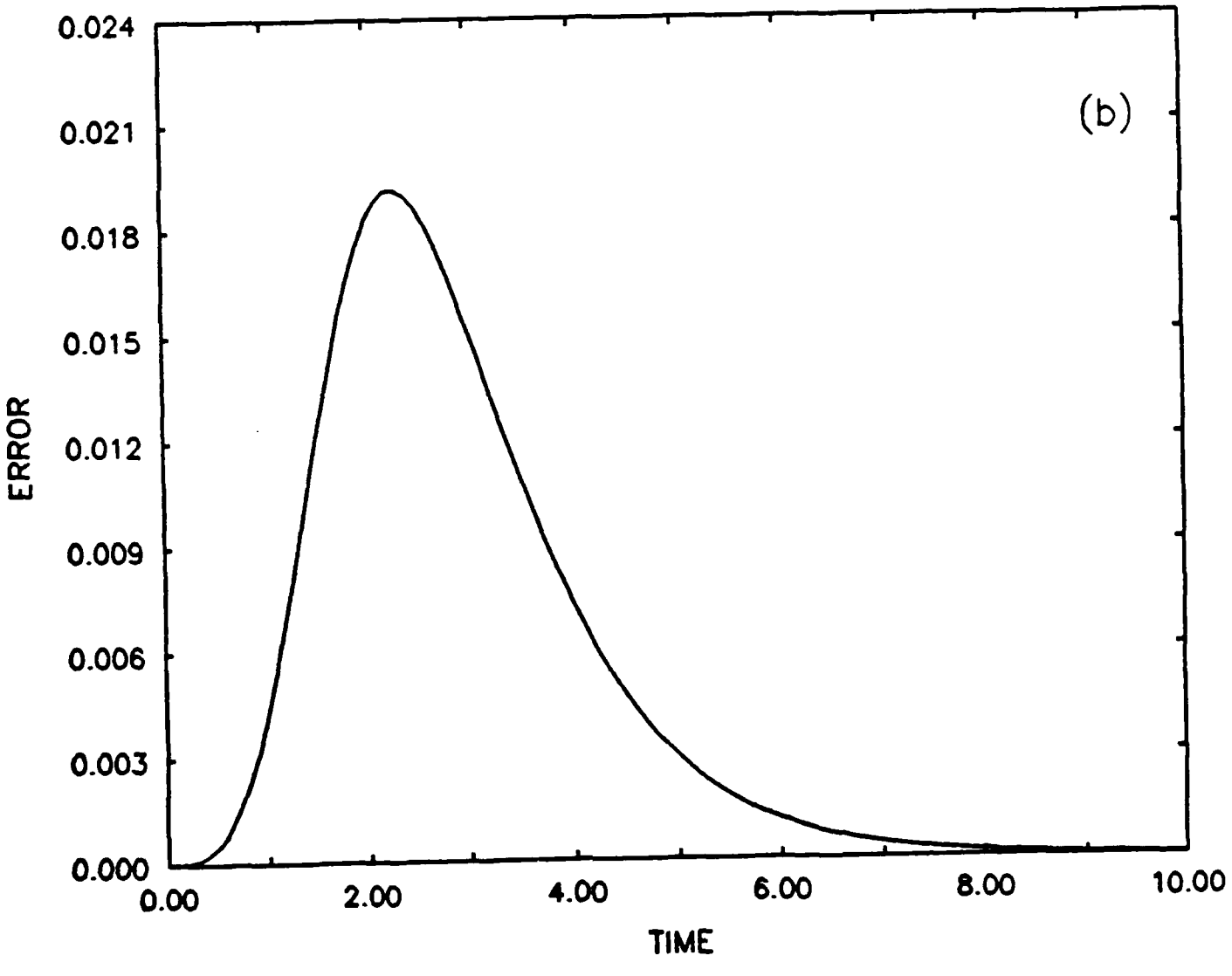

(b)
0.024
0.021
0.018
0.015
0.012
0.009
0.006
0.003
0.000
ERROR
0.00
2.00
4.00
6.00
8.00
10.00
TIME

Figure 6a

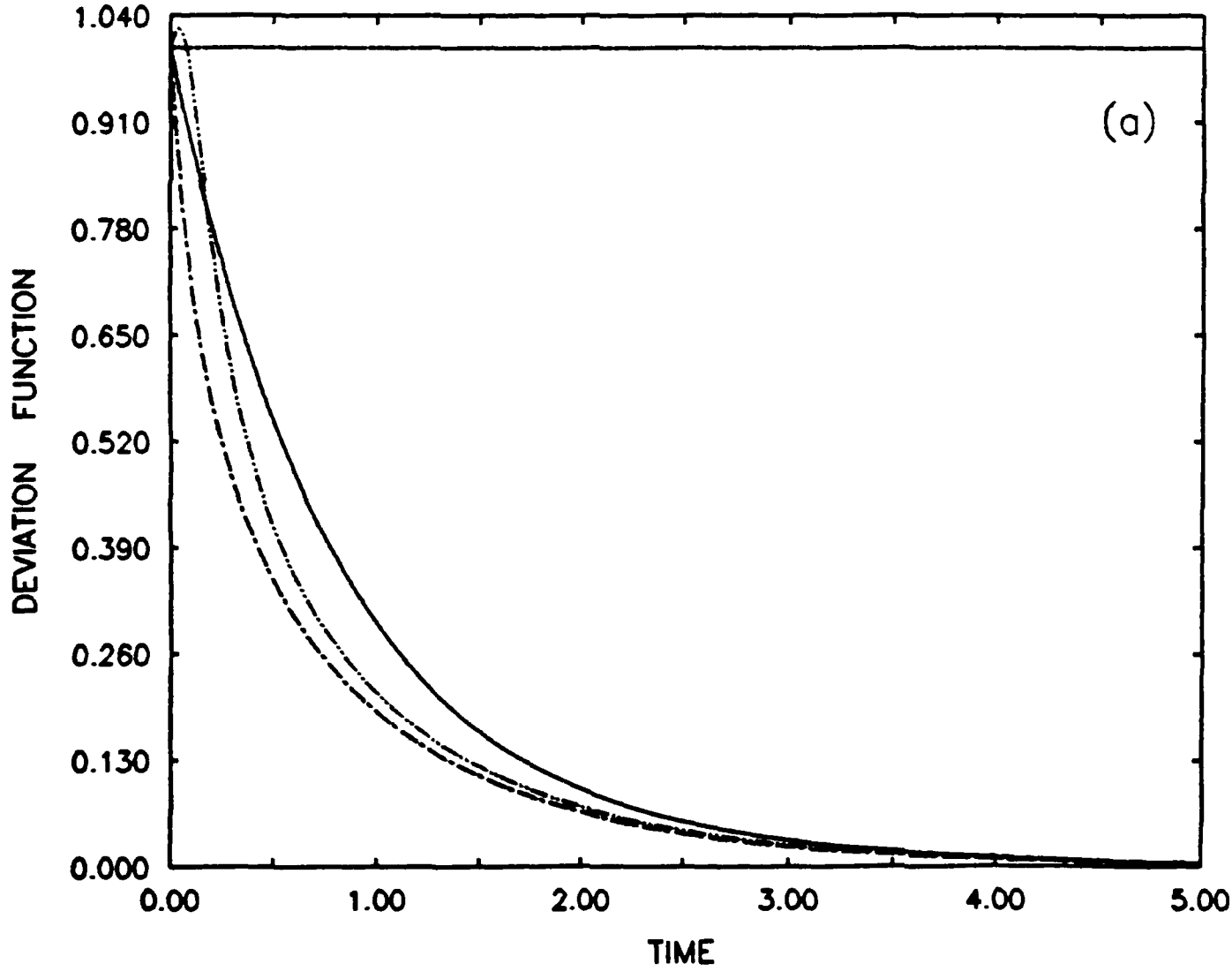
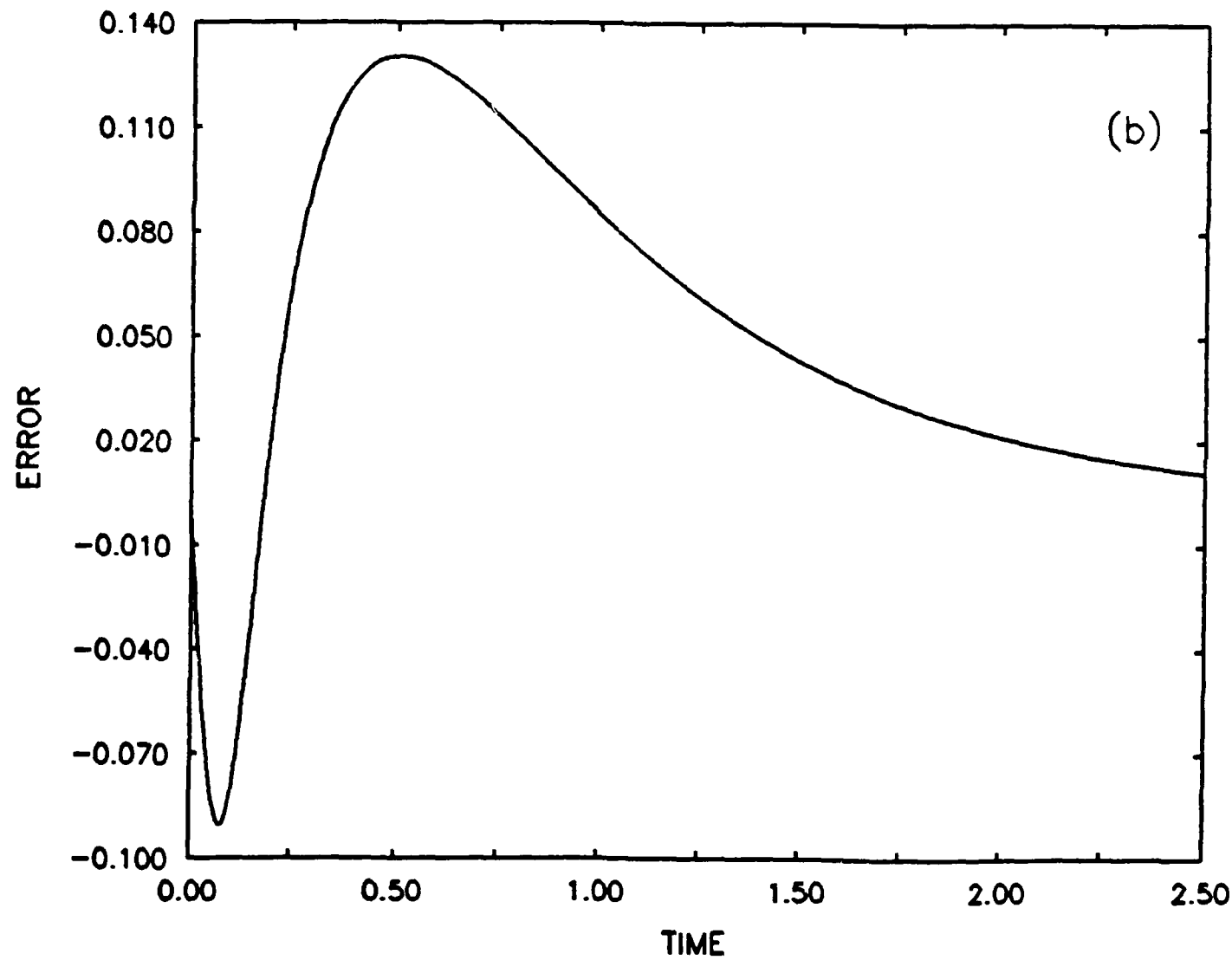

1.040
0.910
0.780
0.650
0.520
0.390
0.260
0.130
0.000
DEVIATION FUNCTION
0.00
1.00
2.00
3.00
4.00
5.00
TIME
(a)

Figure 6b


0.140
0.110
0.080
0.050
0.020
-0.010
-0.040
-0.070
-0.100
ERROR
(b)
0.00
0.50
1.00
1.50
2.00
2.50
TIME

STATE ISOMORPHISM APPROACH TO GLOBAL IDENTIFIABILITY OF NONLINEAR SYSTEMS

S. Vajda and H. Rabitz

Department of Chemistry, Princeton University

Princeton, N.J. 08544

Abstract - Global deterministic identifiability of nonlinear systems is studied by constructing the family of local state isomorphisms that preserve the structure of the parametric system. The method is simplified for homogeneous systems, where such isomorphisms are shown to be linear, thereby reducing the identifiability problem to solving a set of algebraic equations. The known conditions for global identifiability in linear and bilinear systems are special cases of these results.

Mailing address: Prof. H. Rabitz, Department of Chemistry, Princeton University, Princeton, N.J. 08544

Phone: 609-452-3917

Keywords: *Nonlinear systems, identification*

Acknowledgment: The authors acknowledge support for this research from the Office of Naval Research and the Air Force Office of Scientific Research.

I. INTRODUCTION

Consider the parameterized nonlinear system

$$\Sigma_p^{x_0(p)} : \begin{cases} \dot{x}(t,p) = f(x(t,p),p) + u\, g(x(t,p),p) \\ y(t,p) = h(x(t,p),p), \quad x(0,p) = x_0(p) \end{cases} \tag{1}$$

Let M and Ω be bounded, connected, and open sets in R^n and R^q, respectively, such that $x \in M$ and $p \in \Omega$, where p represents the constant parameter vector. We assume that the vector fields $f(\cdot,p)$ and $g(\cdot,p)$, and the function $h(\cdot,p) : M \to R^m$ are real analytic on M for all $p \in \Omega$. The problem considered here is identifiability of (1) in the experiments specified by $(x_0(p), U[0,t_1])$, where $x_0(p)$ denotes the (possibly parameterized) initial state, and $U[0,t_1]$ is the set of bounded and measurable controls defined on $[0,t_1]$. Let $\Sigma_p^{x_0(p)}$ denote the input-output map of (1). Parameter values p , $\tilde{p} \in \Omega$ are said to be indistinguishable (denoted by $p \sim \tilde{p}$) in the experiments $(x_0(p), U[0,t_1])$ if $\Sigma_p^{x_0(p)}(u) = \Sigma_{\tilde{p}}^{x_0(\tilde{p})}(u)$ for all $u \in U[0,t_1]$. System (1) is globally identifiable at p if $\tilde{p} \sim p$, $\tilde{p} \in \Omega$, implies $\tilde{p} = p$. System (1) is locally identifiable at p if there exists an open neighborhood W of p in Ω such that $\tilde{p} = p$, $\tilde{p} \in W$, implies $\tilde{p} = p$.

A summary of results on local identifiability of (1) is given in [1]. These results are based on three factors: (i) the relationship between local identifiability and local observability of the system augmented with the parameters as additional state variables; (ii) the functional expansion of the input-output map of (1), and (iii) the local state isomorphism theorem of nonlinear realization theory. While the first approach is inherently local, functional expansions (e.g., Taylor and generating series) enable one to study also global identifiability by formulating a set of algebraic equations for the parameters[2,3]. Except linear[2], bilinear[4], and homogeneous polynomial[5] systems there exists, however, no a priori upper bound on the number of series coefficients to be considered, and hence the resulting conditions are sufficient but not neccessary from a practical point of view. The structure of nonlinear

equations is far from simple (see, e.g.,[6]), their number is large even for bilinear and polynomial systems, and hence global identifiability properties are difficult to establish in most applications.

The goal of this note is to extend the state isomorphism approach to the analysis of global identifiability in nonlinear systems. In addition to its analiticity we assume that system (1) is locally reduced at $x_0(p)$ for all $p \in \Omega$, i.e., it satisfies both the controllability rank condition (C.R.C) and the observability rank condition (O.R.C)[7].

The problem of global identifiability is stated as follows.

Problem statement: Given (1) and $p \in \Omega$, find all $\tilde{p} \in \Omega$ and systems of the form

$$\Sigma_{\tilde{p}}^{\tilde{x}_0(\tilde{p})} : \begin{cases} \dot{\tilde{x}}(t,\tilde{p}) = f(\tilde{x}(t,\tilde{p}),\tilde{p}) + u\, g(\tilde{x}(t,\tilde{p}),\tilde{p}) \\ y(t,\tilde{p}) = h(\tilde{x}(t,\tilde{p}),\tilde{p}), \quad \tilde{x}(0,\tilde{p}) = \tilde{x}_0(\tilde{p}) = x_0(\tilde{p}) \end{cases} \tag{2}$$

such that

$$\Sigma_p^{x_0(p)}(u) = \Sigma_{\tilde{p}}^{\tilde{x}_0(\tilde{p})}(u) \tag{3}$$

for all $u \in U[0,t_1]$.

It follows that we deal with a highly restricted problem of system equivalence. First, both (1) and (2) are locally reduced , and have the same subset M in R^n as their state spaces. Second, in addition to the input-output map, the known system structure is also invariant under the feasible class of local state isomorphisms. The analysis is based on the construction of all such transformations. This idea has been applied to linear systems[2,8], where equivalence transformations are linear. Though local state isomorphisms between (1) and (2) generally are solutions of a set of partial differential equations, their construction is relatively simple for locally identifiable systems. We will also show that any local state isomorphism, preserving the structure of a homogeneous system, is linear. Thus the method is very simple for this class of systems, and the known conditions for global identifiability of linear

and bilinear systems are special cases of the present results. The single-input case is considered for notational simplicity and the conditions can be readily extended.

II. IDENTIFIABILITY AND CONSTRAINED EQUIVALENCE

The following condition for identifiability is the immediate consequence of the local state isomorphism theorem ([1],[7],[9],[10]) and the constraint (2) on the form of the representations of (1).

Proposition 1: Consider $p, \tilde{p} \in \Omega$, an open neighborhood V of $x_0(\tilde{p})$ in R^n, and any analytic map $\lambda : V \to R^n$ defined on V such that

$$\text{(i)} \qquad \lambda(x_0(\tilde{p})) = x_0(p) \tag{4}$$

$$\text{(ii)} \qquad \operatorname{rank} \frac{\partial \lambda}{\partial \tilde{x}} = n \quad \text{at all } \tilde{x} \in V \tag{5}$$

$$\text{(iii)} \qquad f(\lambda(\tilde{x}), p) = \frac{\partial \lambda}{\partial \tilde{x}} f(\tilde{x}, \tilde{p}) \tag{6a}$$

$$g(\lambda(\tilde{x}), p) = \frac{\partial \lambda}{\partial \tilde{x}} g(\tilde{x}, \tilde{p}) \tag{6b}$$

$$h(\lambda(\tilde{x}), p) = h(\tilde{x}, \tilde{p}) \tag{6c}$$

for all $\tilde{x} \in V$. Then there exists $t_1 > 0$ such that (1) is globally identifiable at p in the experiments $(x_0(p), U[0, t_1])$ iff conditions (4)-(6) imply $\tilde{p} = p$.

Proof: (Necessity.) Assume that $\tilde{p} \neq p$, V, and λ satisfy (4)-(6). Introducing $\tilde{x} = \lambda^{-1}(x)$ into (1) gives

$$\tilde{\Sigma}_p^{\tilde{x}_0(p)} : \begin{cases} \dot{\tilde{x}} = \left(\partial\lambda/\partial\tilde{x}\right)^{-1} f(\lambda(\tilde{x}), p) + u \left(\partial\lambda/\partial\tilde{x}\right)^{-1} g(\lambda(\tilde{x}), p) \\ y = h(\lambda(\tilde{x}), p), \quad \tilde{x}_0(p) = \lambda^{-1}(x_0(p)). \end{cases} \tag{7}$$

Select $t_1 > 0$ such that $\tilde{x}(t, p) \in V$ for all $u \in U[0, t_1]$, where $\tilde{x}(t, p)$ is the solution of the differential equation in (7) with the initial state $\tilde{x}_0(p)$. By (ii) λ is a local state isomorphism defined on V and hence $\Sigma_p^{x_0(p)}(u) = \tilde{\Sigma}_p^{\tilde{x}_0(p)}(u)$ for all $u \in U[0, t_1]$. By (i) $\tilde{x}_0(p) = \lambda^{-1}(\lambda(x_0(\tilde{p}))) = x_0(\tilde{p})$, and by (iii) (7) is represented by (2).Therefore,$\Sigma_p^{x_0(p)}(u) = \Sigma_{\tilde{p}}^{\tilde{x}_0(\tilde{p})}(u)$ for all $u \in U[0, t_1]$, and $\tilde{p} \sim p$ follows.

(Sufficiency.) If $\tilde{p} \neq p$, $\tilde{p} \sim p$, then (2) is a local representation of (1) on some neighborhood V_1 of $\tilde{x}_0(p)$, and it is locally reduced. By the local state isomorphism theorem for any such representation S of (1) there exists an open neighborhood V_2 of $\tilde{x}_0(p)$ and a unique analytic diffeomorphism λ defined on V_2 such that S is of the form (7). Therefore, $\tilde{p}$ and λ satisfy the conditions (4)-(6) on $V = V_1 \cap V_2$. □

Remark 1: Let (1) be globally identifiable at p. It follows from the uniqueness of the local diffeomorphism λ (see,e.g.,[10]) that the only pair $(\tilde{p}, \lambda)$ that satisfies the conditions of Proposition 1 is (p, id_n), where $id_n : V \to R^n$ denotes the identity mapping. Conversely, $\lambda \neq id_n$ implies $\tilde{p} \neq p$.

Remark 2: If certain initial states are completely known, write $x_0(p) = ([x_0^{(1)}]^T, [x_0^{(2)}(p)]^T)^T$, where $x_0^{(1)}$ represents the parameter-independent components of x_0. Then (4) defines the constraints

$$\lambda\begin{pmatrix} x_0^{(1)} \\ x_0^{(2)}(\tilde{p}) \end{pmatrix} = \begin{pmatrix} x_0^{(1)} \\ x_0^{(2)}(p) \end{pmatrix} \tag{8}$$

on λ. In the limiting (i.e.,parameter-independent) case $x_0 = x_0^{(1)}$, and (8) is reduced to

$$\lambda(x_0) = x_0. \tag{9}$$

Example 1: Consider the system

$$\begin{aligned} \dot{x}_1 &= p_1 x_1^2 + p_2 x_1 x_2 + u \qquad & x_1(0,p) = x_2(0,p) = 0 \\ \dot{x}_2 &= p_3 x_1^2 + p_4 x_1 x_2 & y = x_1. \end{aligned} \tag{10}$$

With $\Omega = \{p \in R^4, p_i \neq 0, |p_i| < K \subset R^1, K > 0\}$ (10) is locally reduced at x_0 for all $p \in \Omega$. We construct all local transformations $\lambda = (\lambda_1, \lambda_2)$ that satisfy conditions (4)-(6). Since $h(\tilde{x}, \tilde{p}) = [1 \;\; 0]\tilde{x}$, by (6c)

$$\lambda_1(\tilde{x}_1, \tilde{x}_2) = \tilde{x}_1 \tag{11}$$

whereas (6b) implies $\partial\lambda_2/\partial\tilde{x}_1 = 0$. Then (6a) is reduced to

$$\begin{pmatrix} p_1\tilde{x}_1^2 + p_2\tilde{x}_1\lambda_2 \\ p_3\tilde{x}_1^2 + p_4\tilde{x}_1\lambda_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \partial\lambda_2/\partial\tilde{x}_2 \end{pmatrix} \begin{pmatrix} \tilde{p}_1\tilde{x}_1^2 + \tilde{p}_2\tilde{x}_1\tilde{x}_2 \\ \tilde{p}_3\tilde{x}_1^2 + \tilde{p}_4\tilde{x}_1\tilde{x}_2 \end{pmatrix}, \tag{12}$$

to be satisfied on an open neighborhood of the origin in R^2. By the first equation of (12), $\tilde{p}_1 = p_1$, and

$$\lambda_2(\tilde{x}_1, \tilde{x}_2) = \frac{\tilde{p}_2}{p_2}\tilde{x}_2. \tag{13}$$

Thus $\partial\lambda_2/\partial\tilde{x}_2 = \tilde{p}_2/p_2$. From the second equation $\tilde{p}_3 = p_2p_3/\tilde{p}_2$ and $\tilde{p}_2\tilde{p}_4/p_2 = \tilde{p}_2p_4/p_2$. Since p_2 is arbitrary and $\tilde{p}_2 \neq 0$ by condition (ii), we have $\tilde{p}_4 = p_4$. The initial states are known, but (9) does not further restrict the one-parameter family of feasible transformations given by (11) and (13), where $\tilde{p}_2 \neq 0$ is arbitrary. Thus (10) is nowhere locally identifiable on Ω. This result can be obtained also by the methods presented in [1]. In addition we show, however, that parameters $\tilde{p}_1$ and $\tilde{p}_4$ are unique, independently of the value of $\tilde{p}_2$, and at fixed $\tilde{p}_2 = p_2$ the system becomes globally identifiable at all $p \in \Omega$. As shown in [11], it is much more tedious to establish these properties by the generating series expansion approach.

Remark 3: This example illustrates two important, though not completely general properties of the present method. First, note that the general solution λ of the set (6a)-(6b) of first-order linear partial differential equation depends on arbitrary functions (see,e.g.,[12]). If (1) is locally or globally identifiable,then at most a finite number of these solutions satisfies the additional constraints (6c) and (8). Therefore,restricting consideration to diffeomorphisms satisfying (6c) and (8) one can expect that there will be no need for actually solving the differential equations. As Example 1 shows, this may be the case even for locally unidentifiable systems. Second, the use of the local state isomorphism theorem restricts the scope of Proposition 1 to some interval $[0, t_1]$. However, if λ satisfies (5) and (6) for all $\tilde{x} \in M$, then the transformation is global and $t_1 > 0$ is arbitrary.

Remark 4: For a time-invariant, structurally controllable and structurally observable parameterized linear system represented by $(A(p), B(p), C(p))$ we have $\lambda(\tilde{x}) = T\tilde{x}$, where $T : R^n \rightarrow R^n$ is a nonsingular linear transformation[12]. Then the further conditions of Proposition 1 are reduced to $T(\theta)x_0(\tilde{p}) = x_0(p)$, $A(p)T(\theta) = T(\theta)A(\tilde{p})$, $B(p) = T(\theta)B(\tilde{p})$, and $C(p)T(\theta) = C(\tilde{p})$, where θ denotes the entries of T and emphasizes that θ is to be determined in addition to $\tilde{p}$ in order to satisfy the equations.The linear system is globally identifiable at p iff the only solution is $\tilde{p} = p$, and then $T(\theta) = I$ follows as mentioned in Remark 1. This agrees with the results of Walter[2,3,8]. Since state diffeomorphisms are linear, a similar identifiability condition can be formulated for bilinear systems with a linear observation function[14].

III. HOMOGENEOUS SYSTEMS

We now show that there exists a more general class of systems such that considerations can be restricted to linear state transformations when solving the problem stated in Section I. The result is based on the following lemma.

Lemma 1: Assume that $f(\cdot, p)$ and $g(\cdot, p)$ are defined by homogeneous coordinate functions, i.e.,there exist integers k and ℓ such that

$$kf(x,p) = (\partial f(x,p)/\partial x)x \ , \qquad \ell g(x,p) = (\partial g(x,p)/\partial x)x \tag{14}$$

at all $x \in M$; and the observation function is linear, $h(x,p) = C(p)x$. If $\tilde{p} \sim p$ in the experiments $(0, U[0,t_1])$ for some $t_1 > 0$, then there exists a nonsingular linear transformation $T : R^n \rightarrow R^n$ such that $x(t,p) = T\tilde{x}(t,\tilde{p})$ for all $0 < t < t_1$, where $x(t,p)$ and $\tilde{x}(t,\tilde{p})$ denote the solutions of the differential equations in (1) and (2), respectively.

Proof: Introduce the notation $\varphi(x,u) = f(x,p) + ug(x,p)$ and $\tilde{\varphi}(x,u) = f(x,\tilde{p}) + ug(x,\tilde{p})$. By Proposition 1 there exists an open neighborhood V_1 of $\tilde{x}_0 = 0$ such that $x = \lambda(\tilde{x})$ on V_1. Thus we can write (1) and (2) as

$$\dot{\lambda} = \varphi(\lambda, u)\ , \qquad \lambda_0 = \lambda(x_0) = 0 \tag{15}$$

and

$$\dot{\tilde{x}} = \tilde{\varphi}(\tilde{x}, u)\ , \qquad \tilde{x_0} = x_0 = 0 \tag{16}$$

respectively.By the local weak controllability of (16) at $\tilde{x}_0 = 0$, there exists an open neighborhood V_2 of $\tilde{x}_0$ such that any $\tilde{x} \in V_2$ is reachable from $\tilde{x}_0$. Therefore,the equality

$$y = C(p)\lambda(\tilde{x}) = C(\tilde{p})\tilde{x} \tag{17}$$

holds for all $\tilde{x} \in V = V_1 \cap V_2$. Let c_j and $\tilde{c}_j$ denote the jth rows of $C(p)$ and $C(\tilde{p})$, respectively.By (17) for any $\tilde{x} \in V$, for any $i \geq 0$, any constant controls $u^1, \ldots, u^i$, and sufficiently small $s_1, \ldots, s_i \geq 0$ we have

$$c_j(\gamma^i_{s_i} \circ \ldots \circ \gamma^2_{s_2} \circ \gamma^1_{s_1}(\lambda(\tilde{x}))) = \tilde{c}_j(\tilde{\gamma}^i_{s_i} \circ \ldots \circ \tilde{\gamma}^2_{s_2} \circ \tilde{\gamma}^1_{s_1}(\tilde{x})) \tag{18}$$

for $j = 1, \ldots, m$. Here $\gamma^i_{s_i}$ and $\tilde{\gamma}^i_{s_i}$ denote the flows of $\varphi^i(\lambda) = \varphi(\lambda, u^i)$ and $\tilde{\varphi}^i(\tilde{x}) = \tilde{\varphi}(\tilde{x}, u^i)$, respectively. Differentiating with respect to $s_i, \ldots, s_1$, at 0 yields

$$L_{\varphi^1}(\ldots(L_{\varphi^i}(c_j\lambda))\ldots)_{\lambda(\tilde{x})} = L_{\tilde{\varphi}^1}(\ldots(L_{\tilde{\varphi}^i}(\tilde{c_j}\tilde{x}))\ldots)_{\tilde{x}} \tag{19}$$

where L_{φ^i} denotes Lie differentiation along the vector field φ^i [7]. Differentiating (19) with respect to $\tilde{x}$ and multiplying by $\tilde{x}$ gives

$$\langle d(L_{\varphi^1}(\ldots(L_{\varphi^i}(c_j\lambda))\ldots)_{\lambda(\tilde{x})}, \frac{\partial\lambda}{\partial\tilde{x}}\tilde{x}\rangle = \langle d(L_{\tilde{\varphi}^1}(\ldots(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x}))\ldots)_{\tilde{x}}, \tilde{x}\rangle \tag{20}$$

where the differential $d(L_{\tilde{\varphi}^1}(\ldots(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x})\ldots)_{\tilde{x}}$ is represented by a row m-vector valued function,and the vector field with the coordinate functions $(\tilde{x}_1, \ldots, \tilde{x}_n)$ is also denoted by $\tilde{x}$. Assume first that $k = \ell$. By (14) we have

$$L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x}) = \tilde{c}_j\tilde{\varphi}^i = \frac{1}{k}\langle d(\tilde{c}_j\tilde{\varphi}^i), \tilde{x}\rangle = \frac{1}{k}\langle d(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x})), \tilde{x}\rangle. \tag{21}$$

Then

$$(L_{\tilde{\varphi}^{i-1}}(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x}))) = \frac{1}{k}\langle d(L_{\tilde{\varphi}^{i-1}}(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x}))), \tilde{x}\rangle + \frac{1}{k}\langle d(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x})), [\tilde{\varphi}^{i-1}, \tilde{x}]\rangle \tag{22}$$

where $[\cdot,\cdot]$ is the Lie bracket of two vector fields (see,e.g.,[9]). By (14) $[\tilde{\varphi}^{i-1}, \tilde{x}] = (1-k)\tilde{\varphi}^{i-1}$, and hence the second term on the rhs of (22) is $(1-k)/k(L_{\tilde{\varphi}^{i-1}}(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x})))$. Therefore,rearranging (22) yields

$$(2k-1)(L_{\tilde{\varphi}^{i-1}}(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x}))) = \langle d(L_{\tilde{\varphi}^{i-1}}(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x})))_{\tilde{x}}, \tilde{x}\rangle \tag{23}$$

Since φ is a k-order homogeneous function of λ, analogously rearranging the lhs of (19) and continuing for $\varphi^{i-2},\ldots,\varphi^1$ yields

$$\langle d(L_{\varphi^1}(\ldots(L_{\varphi^i}(c_j\lambda))\ldots)_{\lambda(\tilde{x})}, \lambda(\tilde{x})\rangle = \langle d(L_{\tilde{\varphi}^1}(\ldots(L_{\tilde{\varphi}^i}(\tilde{c}_j\tilde{x}))\ldots)_{\tilde{x}}, \tilde{x}\rangle. \tag{24}$$

From (20) and (24)

$$\langle d(L_{\varphi_1}(\ldots(L_{\varphi^i}(c_j\lambda))\ldots)_{\lambda(\tilde{x})}, \frac{\partial\lambda}{\partial\tilde{x}}\tilde{x}\rangle = \langle d(L_{\varphi^1}(\ldots(L_{\varphi^i}(c_j\lambda))\ldots)_{\tilde{x}}, \tilde{x}\rangle. \tag{25}$$

Since the O.R.C. is satisfied at 0, by analiticity of (1) there exists an open neighborhood of 0, also denoted by V, such that the vectors $d(L_{\varphi^1}(\ldots(L_{\varphi^i}(c_j\lambda)\ldots)_{\lambda(\tilde{x})}$ span an n-dimensional space at all $\tilde{x} \in V$ [7]. Thus equations of the form (25) imply

$$\frac{\partial\lambda}{\partial\tilde{x}}(\tilde{x})\tilde{x} = \lambda(\tilde{x}) \tag{26}$$

for all $\tilde{x} \in V$. By (26), $\partial\lambda/\partial\tilde{x}$ is a zero-order homogeneouos function, i.e.,

$$\frac{\partial\lambda}{\partial\tilde{x}}(\alpha\tilde{x}) = \frac{\partial\lambda}{\partial\tilde{x}}(\tilde{x}) \tag{27}$$

for all $\tilde{x} \in V$ and all $\alpha \in R^1$ satisfying $\alpha\tilde{x} \in V$, and by $0 \in V$ it is defined at $\tilde{x} = 0$. Setting $\alpha = 0$ in (27) shows that $\partial\lambda/\partial\tilde{x}(0) = \partial\lambda/\partial\tilde{x}(\tilde{x})$ at all $\tilde{x} \in V$, and

$\lambda(\tilde{x}) = T\tilde{x}$. By (14) $\lambda(\tilde{x}) = T\tilde{x}$ satisfies (5) and (6) for all $\tilde{x} \in M$, and hence $x(t,p) = T\tilde{x}(t,\tilde{p})$ for all $0 < t < t_1$.

For $\ell < m$ introduce the additional state variable x_{n+1}, an $(m-\ell)$-order homogeneouos function $r : I_\beta \to R^1$, where I_β is an open subset in R^1 such that $r(s) \neq 0$ for all $s \in I_\beta$, and input $\tilde{u} = u/r(x_{n+1})$.Let $x^* = (x^T, x_{n+1})^T$, $\tilde{x}^* = (\tilde{x}^T, \tilde{x}_{n+1})^T$, and consider the augmented vector field $\varphi^* : M \times I_\beta \to R^{n+1}$ and matrix C^* defined by

$$\varphi^*(x^*, \tilde{u}) = \begin{pmatrix} f(x,p) + \tilde{u}r(x_{n+1})g(x,p) \\ 0 \end{pmatrix}, \qquad C^* = \begin{pmatrix} C(p) & 0 \\ 0 & 1 \end{pmatrix}. \tag{28}$$

Augmenting (1) and (2) we have

$$\dot{x}^* = \varphi^*(x^*, \tilde{u}), \quad y = C^* x^*, \quad x_0^* = (x_0^T, \beta)^T, \tag{29}$$

and

$$\dot{\tilde{x}}^* = \tilde{\varphi}^*(\tilde{x}^*, \tilde{u}), \quad y = \tilde{C}^* \tilde{x}^*, \quad \tilde{x}_0^* = (\tilde{x}_0^T, \beta)^T, \tag{30}$$

where $\tilde{\varphi}^*$ and $\tilde{C}^*$ are defined by (28) at the parameter value $\tilde{p}$. Let $\Sigma_p^{x_0^*(p)}$ and $\Sigma_{\tilde{p}}^{\tilde{x}_0^*(p)}$ denote the input-output maps of (29) and (30), respectively.Since

$$\Sigma_p^{x_0^*(p)}(\tilde{u}) = \begin{pmatrix} \Sigma_p^{x_0(p)}(u) \\ \beta \end{pmatrix}, \tag{31}$$

$\tilde{p} \sim p$ implies $\Sigma_p^{x_0^*(p)}(u) = \Sigma_{\tilde{p}}^{\tilde{x}_0^*(\tilde{p})}(u)$ for all $u \in U[0,t_1]$ and for all $\beta \in I_\beta$. Since $\tilde{x}_{n+1} = x_{n+1} = \beta$, in spite of uncontrollability of (29) and (30), the only isomorphism between $\tilde{x}^*$ and x^* is given by $\lambda^* : \tilde{x}^* \to (\lambda^T(\tilde{x}), \tilde{x}_{n+1})$, and then $y = C^*\lambda^*(\tilde{x}^*) = \tilde{C}^*\tilde{x}^*$ for all $\tilde{x}^* \in V \times I_\beta$. The augmented systems are homogeneouos,and the previous part of the proof applies and yields

$$\begin{pmatrix} \partial\lambda(\tilde{x})/\partial\tilde{x} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \tilde{x} \\ \tilde{x}_{n+1} \end{pmatrix} = \begin{pmatrix} \lambda(\tilde{x}) \\ \tilde{x}_{n+1} \end{pmatrix}. \tag{32}$$

This is valid for all $\tilde{x} \in V$, and thus (26) and (27) follow. For $m - \ell$ the proof is analogous with $\tilde{u} = ur(x_{n+1})$. □

Replacing $\lambda(\tilde{x})$ by $T\tilde{x}$ in (4)-(6) gives a simple condition for identifiability of homogenous systems in the experiments $(0, U[0,t_1])$, $t_1 > 0$ arbitrary. This result can be slightly extended by considering a set $I_1 \subseteq R^{n_1}$, $n_1 \leq n$ of feasible initial states $x_0^{(1)}$. Define $I_p = \{x_0(p); x_0^{(1)} \in I_1\}$. System (1) is globally identifiable at $p \in \Omega$ in the experiments $(I_p, U[0,t_1])$ if and only if there exists $x_0(p) \in I_p$ such that (1) is globally identifiable in the experiments $(x_0(p), U[0,t_1])$. From Proposition 1 and Lemma 1 we have the following result.

Proposition 2: Assume that system (1) satisfies the assumptions of Lemma 1, it is locally reduced at $x_0(p)$ for all $p \in \Omega$,and for all $x_0^{(1)} \in I_1$, and $0 \in I_p$. Consider $p, \tilde{p} \in \Omega$, and any linear transformation $T(\theta) : R^n \rightarrow R^n$ such that (i) $T(\theta)x_0(\tilde{p}) = x_0(p)$ for all $x_0^{(1)} \in I_1$, (ii) $T(\theta)$ is nonsingular, and (iii) $f(T(\theta)\tilde{x}, p) = T(\theta)f(\tilde{x}, \tilde{p})$, $g(T(\theta)\tilde{x}, p) = T(\theta)g(\tilde{x}, \tilde{p})$, and $C(p)T(\theta) = C(\tilde{p})$ for all $\tilde{x} \in M$. Then (1) is globally identifiable at p in the experiments $(I_p, U[0,t_1])$ for arbitrary $t_1 > 0$ iff conditions (i),(ii),and (iii) imply $\tilde{p} = p$.

Proof: As shown in the proof of Lemma 1 for $x_0 = 0$, $x = T\tilde{x}$ is a global transformation defined at all $\tilde{x} \in M$ if (1) is homogeneous. Assume that T satisfies the constraint (i). By analiticity of local diffeomorphisms, $\lambda(\tilde{x}) = T\tilde{x}$ for any $x_0(p)$ and for any local transformation λ defined on some open neighborhood of $x_0(p)$, and the proposition follows. □

The identifiability conditions for linear and bilinear systems, discussed in Remark 4 are particular cases of Corollary 1 with $m = 1$, $\ell = 0$, and $m = 1$, $\ell = 1$, respectively.A further application,particularly important in ecology and chemical reaction kinetics, is to the system

$$\begin{aligned} \dot{x}_i(t,p) &= x^T(t,p)A^{(i)}(p)x(t,p) + b_i(p)u \ , \quad i = 1,\ldots,n, \\ y(t,p) &= C(p)x(t,p) \ , \qquad\qquad x(0,p) = x_0(p). \end{aligned} \tag{33}$$

where $A^{(i)}$, $i = 1,\ldots,n$, are $n \times n$ symmetric matrices. Denote $a_{ij}^{(k)} = [A^{(k)}]_{ij}$ and $\theta_{ij} = [T(\theta)]_{ij}$, $i,j,k = 1,\ldots,n$, then we have the following result.

Corollary 1: Consider $p, \tilde{p} \in \Omega$ and any nonsingular linear transformation $T(\theta) : R^n \to R^n$ such that $T(\theta)x_0(\tilde{p}) = x_0(p)$ for all $x_0^{(1)} \in I_1$, and

$$b(p) = T(\theta)b(\tilde{p}) , \qquad C(p)T(\theta) = C(\tilde{p}) \tag{34}$$

$$\sum_{r=1}^{n}\sum_{s=1}^{n} a_{sr}^{(k)}(p)\theta_{rj}\theta_{si} = \sum_{r=1}^{n} a_{ij}^{(r)}(\tilde{p})\theta_{kr}, \quad i,j,k = 1,\ldots,n. \tag{35}$$

System (33) is globally identifiable at p in the experiments $(I_p, U[0,t_1])$, where $0 \in I_p$, and $t_1 > 0$ arbitrary, iff the above conditions imply $\tilde{p} = p$.

Remark 5: As mentioned in Remark 1,global identifiability also implies $T(\theta) = I$.

Example 2: Consider the system

$$\begin{aligned} \dot{x}_1 &= p_1 x_1 x_2 + u , \qquad & x_1(0,p) = x_2(0,p) = 0 , \\ \dot{x}_2 &= p_2 x_1 x_2 + p_3 x_2^2 + u \quad & y = x_2 , \end{aligned} \tag{36}$$

which is of the form (33) with

$$A^{(1)}(p) = \begin{pmatrix} 0 & p_1/2 \\ p_1/2 & 0 \end{pmatrix}, \qquad A^{(2)}(p) = \begin{pmatrix} 0 & p_2/2 \\ p_2/2 & p_3 \end{pmatrix}, \qquad b(p) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{37}$$

and $C = [0\ 1]$. It is easy to show that (36) is locally reduced at $x_0 = 0$ for all $p \in \Omega = \{p \in R^3; p_i \neq 0, p_1 - p_2 - p_3 \neq 0, |p_i| < K \in R^1, K > 0\}$. By Proposition 2 we consider the nonsingular linear transformations $T(\theta) : R^2 \to R^2$. Since $b(p) = b(\tilde{p})$ and $C(p) = C(\tilde{p})$, conditions (34) restrict $T(\theta)$ to the form

$$T(\theta) = \begin{pmatrix} 1-\theta_{12} & \theta_{12} \\ 0 & 1 \end{pmatrix} , \tag{38}$$

and $T(\theta)$ is nonsingular for any $\theta_{12} \neq 1$. Then the nontrivial equations in (34) are $\tilde{p}_1 - \theta_{12}(\tilde{p}_1 - \tilde{p}_2) = p_1 - \theta_{12}p_1$, $\theta_{12}\tilde{p}_3 = \theta_{12}p_1$, $\tilde{p}_2 = p_2 - \theta_{12}p_2$, and $\tilde{p}_3 = \theta_{12}p_2 + p_3$. At $\theta_{12} = 0$, $\tilde{p} = p$ and $T(\theta) = I$. Except $p_1 = p_3$ there exists, however,a second solution $\theta_{12} = (p_1 - p_3)/p_2$, which yields $\tilde{p}_1 = p_3$, $\tilde{p}_2 = p_2 + p_3 - p_1$, and $\tilde{p}_3 = p_1$. Therefore,the system is locally identifiable at all $p \in \Omega$, but it is globally identifiable

only on the subset $\{p \in \Omega; p_1 = p_3\}$ of zero measure in Ω. We note that by the lack of applicable necessary conditions for global identifiability, (36) is the first such nonlinear system presented in the literature (see,e.g.,[1],[2],[3],[6],[11],[15]).

By Corollary 1 we can study also identifiability of (36) with nonzero initial conditions.In the most general case $x_1(0,p) = p_4$ and $x_2(0,p) = p_5$ are additional parameters. By condition (i) of Proposition 2 the additional constraints on (38) are $(1-\theta_{12})\tilde{p}_4 + \theta_{12}\tilde{p}_5 = p_4$ and $\tilde{p}_5 = p_5$. It follows that $\tilde{p}_5$ is unique and there exist two solutions for $\tilde{p}_4$. The system becomes, however, globally identifiable at all $p \in \Omega$ if $p_4 = x_{0,1}$ is known and there exists a point $x_{0,1} \neq 0$ in I_1 (i.e., the constraint $\tilde{p}_4 = p_4$ implies $\theta_{12} = 0$).

REFERENCES

[1] E. T. Tunali and T. J. Tarn, "New results for identifiability of nonlinear systems," *IEEE. Trans. Automat. Contr.*, vol. AC-32, pp. 146-154, 1987.

[2] E. Walter, *Identifiability of State Space Models*. New York: Springer-Verlag, 1982.

[3] E. Walter and Y. Lecourtier, "Global approaches to identifiability testing for linear and nonlinear state space models," *Math. Comput. Simulation*, vol. 24, pp. 472-482, 1982.

[4] S. Vajda, "Structural identifiability of dynamical systems," *Int. J. Systems. Sci.*, vol. 14, pp. 1229-1247, 1983.

[5] S. Vajda, "Deterministic identifiability and algebraic invariants for polynomial systems," *IEEE Trans. Automat. Contr.*, vol. AC-32, pp. 182-184, 1987.

[6] K. R. Godfrey and W. R. Fitch, "The deterministic identifiability of nonlinear pharmacokinetic models," *J. Pharmacokin. Biopharm.*, vol. 12, pp. 177-190, 1984.

[7] R. Hermann and A. J. Krener, "Nonlinear controllability and observability," *IEEE Trans. Automat. Contr.*, vol. AC-22, pp. 728-740, 1977.

[8] E. Walter and Y. Lecourtier, "Unidentifiable compartmental models: what to do?," *Math. Biosci.*, vol.56, pp. 1-25, 1981.

[9] A. Isidori, *Nonlinear Control Systems: An Introduction*. New York: Spinger-Verlag, 1985.

[10] H. J. Sussmann, "Existence and uniqueness of minimal realizations of nonlinear systems," *Math. Syst. Theory,* vol. 10, pp. 263-284. 1977.

[11] S. Vajda, "Identifiability of polynomial systems: structural and numerical aspects," in *Identifiability of Parametric Models*, E. Walter, Ed., pp.42-49, New York: Pergamon, 1987.

[12] L. R. Ford, *Differential Equations.* New York: McGraw-Hill, 1955.

[13] L. M. Silverman, "Realization of linear dynamical systems," *IEEE Trans. Automat. Contr.*, vol. AC-16, pp. 554-567, 1971.

[14] P. D'Alessandro,A. Isidori, and A. Ruberti, "Realization and structure theory of bilinear dynamical system," *SIAM J. Control*, vol. 12, pp. 517-535, 1974.

[15] H. Pohjanpalo, "System identifiability based on the power series expansion of the solution," *Math. Biosci.*, vol. 41, pp. 21-33, 1978.

IDENTIFIABILITY AND DISTINGUISHABILITY OF FIRST-ORDER REACTION SYSTEMS

S. Vajda[1] and H. Rabitz

Department of Chemistry, Princeton University

Princeton, N. J. 08544, U.S.A.

Abstract

By following the kinetics of a reaction through the use of certain classes of measurable quantities instead of the concentrations of all species neither the parameter values nor the reaction scheme are necessarily unique. Identifiability deals with the problem of determining whether an experiment is able to supply the desired information on the parameters of an assumed kinetic model, whereas indistinguishability means that two different reaction schemes generate the same values for the observed quantities in any possible experiment. This paper examines these issues for the case of first-order reaction systems and both problems are solved by the same analytical tools. The method involving Laplace transforms is conceptually simple, easy to apply, and is also used to derive simple rules to test distinguishability of reaction schemes. Another approach based on similarity transformations is used to generate all the first-order reaction schemes that are indistinguishable from a given one.

I. Introduction

Kinetic experiments are often conducted under conditions such that the reactions are first-order or pseudo first-order, with rate coefficients proportional to the concentration of a reaction partner in large excess. Interpretation of experimental data by postulating a mechanism and adjusting the values of some unknown parameters has received due attention in the literature[2-6]. The problems usually considered are techniques of parameter estimation and statistical interpretation of the estimates in terms of confidence intervals or joint confidence regions. Kinetists are aware that there remain further fundamental questions to ask.[6] First, are the derived parameters unique, or are there further parameter sets generating the same values for the observed quantities? Second, is the selected model the only plausible one which will give an acceptable fit to the data? These questions of parameter and model uniqueness are not trivial even for very simple mechanisms if not all concentrations are directly observed.

For example, consider the consecutive reaction scheme

$$A \rightarrow B \rightarrow C \tag{1.1}$$

studied in several works[3-5] assuming that initially only A is present in the system and the reaction is followed by observing the single property

$$y = \epsilon_A[A] + \epsilon_B[B] + \epsilon_C[C) \tag{1.2}$$

which may represent absorbance, conductivity, pH, or ligand release. We regard y as absorbance and ϵ_A, ϵ_B and ϵ_C as molar extinction coefficients. Frequently the intermediate species B cannot be isolated and separately investigated, hence ϵ_B is an additional parameter to be estimated simultaneously with the rate coefficients k_1 and k_2 from the time-absorbance data. As is well known,[3-5] under these conditions the solution of the estimation problem is not unique because of the slow-fast ambiguity, thus for

any solution $k=(k_1, k_2, \epsilon_B)^T$ there exist a second solution $k=(\bar{k}_1, \bar{k}_2, \bar{\epsilon}_B)$ given in terms of k by

$$\bar{k}_1 = k_2;\ \bar{k}_2 = k_1;\ \bar{\epsilon}_B = \epsilon_A + k_1(\epsilon_B - \epsilon_A)/k_2. \tag{1.3}$$

In addition to nonuniqueness in parameter values there may be ambiguities in the model structure. As emphasized by Milligan et al.,[5] a qood fit does not necessarily mean that the model is correct, since there exist further reaction schemes generating the same absorbance curve. They mention the schemes

$$S_1:\ A \underset{k_1}{\overset{k_1}{\rightleftharpoons}} B \longrightarrow C\ ; \qquad S_2:\ \begin{array}{ccccc} A & \xrightarrow{k_1} & B & & \\ & \searrow^{k_2} & C & \nearrow_{k_3} & \end{array}, \tag{1.4}$$

whereas Jackson et al.[4] claim that the absorbance data can equally be described by adopting the reaction schemes

$$S_3:\ \begin{array}{ccc} A & \xrightarrow{k_1} & \\ & & C \\ B & \xrightarrow{k_2} & \end{array}, \qquad S_4:\ \begin{array}{ccccc} A & \xrightarrow{k_1} & B & & \\ & \searrow^{k_3} & & \searrow^{k_2} & \\ & & C & & \end{array}, \qquad S_5:\ A \xrightarrow{k_1} B \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} C. \tag{1.5}$$

The purpose of this paper is to present a systematic and rather general analysis of the problems of parameter uniqueness, called identifiability, and distinguishability of different first-order reaction schemes. While identifiability has received a fair amount of attention in application areas such as automatic control,[7] compartmental modeling,[8] and chemical engineering,[9] in chemical kinetics results have been restricted to discovering nonuniqueness of the parameters in particular reaction systems through the use of methods of limited applicability. Similarly, the general results available on distinguishability are rarely applied to kinetic models,[9c] though without systematic analysis mistakes can be made. We will

show that the schemes S_2, S_3, and S_5 in (1.4) and (1.5) are, in fact, distingishable from the one in (1.1), whereas there exist indistingushable schemes overlooked in previous studies of this simple system (for the illustrative purposes of this paper we shall assume a measurement of the form in (1.2) in most examples).

Identifiability and distinguishability are so closely related that all the required machinery is introduced by discussing the first and somewhat simpler problem. Identifiability concepts are also needed to properly understand some distinguishability results. For example, we show that (1.1) and the scheme S_1 in (1.4) are indistinguishable, but their identifiability properties are substantially different, since using the latter model the desired absorbance curve can be generated at infinitely many different parameter values.

We will regard two reaction schemes as indistinguishable if and only if they generate exactly the same values for the observed quantities (e.g., for absorbance in the case discussed here) and hence employ a deterministic framework by restricting considerations to idealized experiments with the ability of observing the measurable variables at any instant of time error-free. Deterministic identifiability is a fundamental property of a kinetic model, since unidentifiability in this idealized experiment implies unidentifiability in any realistic experiment with constraints on sampling and measurement accuracy. Similarly, models indistinguishable in the deterministic sense remain indistinguishable in any experiment involving the same measurable quantities. It should be emphasized that the deterministic analysis is only the first step in establishing uniqueness of parameter estimates or uniqueness of a kinetic model. In fact, inadequate design of the identification experiment and/or large measurement errors may result in

highly uncertain estimates even for the parameters of an identifiable model. Similarly, a set of noisy observations may be compatible with the responses of several models in spite of their deterministic distinguishability. Since the analysis of these problems requires assumptions on the experiment design, on the structure of measurement errors, and on the values of the parameters, it can usually be performed only a posteriori after carrying out the experiment and estimating the parameters. The deterministic analysis is, however, an a priori procedure for detecting a fundamental class of ambiguities, thereby assisting the selection of possible models and the variables to be observed in the intended experiments.

The paper is organized as follows. In Section II we introduce the concepts of deterministic identifiability and offer two general methods of analysis based on Laplace and similarity transformations, respectively. The Laplace transformation approach is also used to study distinguishability in Section III and enables us to formulate a number of propositions, thereby considerably facilitating the required algebraic manipulations. The similarity transformation approach is fully exploited in Section IV offering a procedure to generate all first-order reaction schemes that are indistingishable from a given one. In particular, results are peresented for the scheme in (1.1). The methods can be most easily understood by solving simple problems and we present a number of examples for this purpose.

II. Identifiability

A first-order reaction scheme under isothermal condition gives rise to the kinetic equations of the general form

$$\dot{x}(t,k) = A(k)x(t,k) \qquad x(0,k) = x_0(k) = \begin{bmatrix} x_0^{(1)} \\ x_0^{(2)}(k) \end{bmatrix}, \qquad (2.1)$$

where $x(t,k)$ is the n-vector of concentrations, depending on the p-vector $k \in \Omega$

of unknown parameters, $\Omega \subseteq R^p$ representing the set of possible parameter values. We assume that Ω is a bounded open set in R^p, thus the parameters are a priori independent and restricted only by inequality constraints (e.g., by nonnegativity of the rate constants). In the initial concentration vector x_o we distinguish between the components in $x_o^{(1)}$, selected to specify an experiment, and those in $x_o^{(2)}$, depending on unknown parameters (e.g., initial conditions in $x_o^{(2)}$ can be parameters themselves). In addition to the kinetic equations (2.1) our model consists of the linear observation function

$$y(t,k) = C(k)x(t,k) \tag{2.2}$$

where $y(t,k)$ is the m-vector of observable quantities, also called the response function of the model. As is seen, the observation matrix $C(k)$ may also depend on unknown parameters.

Consider a kinetic experiment specified by the initial concentrations $x_o^{(1)}$ and let $\tilde{y}(t)$ represent the response function observed over some time interval T. The basic assumption of deterministic analysis is the existence of a nominal parameter value $k \in \Omega$ such that $y(t,k) = \tilde{y}(t)$ and this function can be observed at all $t \in T$ error-free. Two parameter values k and $\bar{k} \neq k$ are indistinguishable in the considered experiment if

$$y(t,\bar{k}) = y(t,k) \tag{2.3}$$

for any $t \in T$. The analysis of identifiability is based on eq. (2.3) and the following situations can be encountered:

(i) if the solution $\bar{k} = k$ of (2.3) is unique, the model (2.1)-(2.2) is said to be uniquely identifiable at $k \in \Omega$;

(ii) if there exist at most a finite number of distinct solutions $\bar{k} \neq k$, the model is said to be identifiable at k;

(iii) finally, with an infinite number of solutions in (2.3) the model is said to be unidentifiable.

Since the nominal parameter value k is not known a priori, the above concepts should be generalized. It would be easy to require identifiability at every $k \in \Omega$. In most models, however, there exist exceptional points or lower dimensional surfaces in Ω where the model is unidentifiable in spite of its identifiability at the majority of points. Properties that hold at almost every point of the parameter set are usually called structural ones.[10] Therefore, the model is said to be structurally identifiable (uniquely structurally identifiable) if it is identifiable (uniquely identifiable) at almost every $k \in \Omega$, thus except at the points of a set of measure zero in Ω. As shown in our examples, the existence of such exceptional subsets does not decrease practical utility of the concepts.

In view of the extensive list of publications[7-9] on the identifiability problem we restrict considerations to two basic methods of analysis enabling one to test first-order reaction systems of moderate complexity without programming efforts. Both methods will also be needed when studying distinguishability of different schemes.

1. Laplace transformation approach

Taking the Laplace transform of the differential equations (2.1) we obtain

$$sX(s,k) = A(k)X(s,k) + x_o(k) \tag{2.4}$$

where $X(s,k)$ is the transform of the concentration vector $x(t,k)$ and s denotes the complex argument.[11] Taking also the transform of (2.2) and using (2.4) gives the Laplace transform

$$Y(s,k) = C(k)[sI-A(k)]^{-1} x_o(k) \tag{2.5}$$

of the response function $y(t,k)$. Note that in spite of the general formula (2.5) the linear equations (2.4) can be solved for $X(s,k)$ by the elimination technique and no matrix inversion is necessary to obtain $Y(s,k)$.

Since (2.3) is satisfied for all $t \in T$ if and only if

$$Y(s,\tilde{k}) = Y(s,k) \tag{2.6}$$

for all $s \in \mathbb{C}$, where $\mathbb{C}$ is the field of complex numbers, we can restrict considerations to (2.5) without solving the kinetic equations. Each component $Y_i(s,k)$ of the m-vector $Y(s,k)$ is a rational function of the form

$$Y_i(s,k) = \frac{\phi^i_{n+1} s^{n-1} + \ldots + \phi^i_{2n}}{s^n + \phi_1{}^i s^{n-1} + \ldots + \phi^i_n} \quad , \tag{2.7}$$

where the coefficients ϕ^i_j generally depend on k and $x_o^{(1)}$. After simplifying the possible common factors between the numerator and the denominator polynomials in (2.7), the vector ϕ of "moment" invariants is formed by all different coefficients in $Y_1(s,k),\ldots,Y_m(s,k)$. Since (2.6) holds if and only if

$$\phi(\tilde{k}) = \phi(k), \tag{2.8}$$

the analysis of identifiability is reduced to the problem of determining the number of solutions in the set (2.8) of polynomial equations.[10,12] The following examples demonstrate the simplicity of the method and the presence of exceptional subsets of zero measure, taken into account in our definitions.

Example 1.1. Consider the reaction scheme S_1 in (1.4) with the response function (1.2) and initial concentrations $[B]_o=[C]_o=0$. The Laplace transform of (1.2) is given by

$$Y(s,k) = \frac{\epsilon_A s^2 + [\epsilon_A(k_{-1}+k_2) + \epsilon_B k_1]s + \epsilon_C k_1 k_2}{s^3 + (k_1+k_{-1}+k_2)s^2 + k_1 k_2 s} [A]_o \tag{2.9}$$

Since $\tilde{\epsilon}_A = \epsilon_A$ and $\tilde{\epsilon}_C = \epsilon_C$ are known, the independent equations of the form (2.8) are

$$\epsilon_A(\tilde{k}_{-1}+\tilde{k}_2) + \tilde{\epsilon}_B\tilde{k}_1 = \epsilon_A(k_{-1}+k_2) + \epsilon_B k_1 \tag{2.10a}$$

$$\bar{k}_1 + \bar{k}_{-1} + \bar{k}_2 = k_1 + k_{-1} + k_2 \tag{2.10b}$$

$$\bar{k}_1\bar{k}_2 = k_1k_2 \tag{2.10c}$$

(a) Assume first that $\bar{\epsilon}_B = \epsilon_B \neq \epsilon_A$ is also known. Then (2.10a) and (2.10b) give $\bar{k}_1=k_1$ and $\bar{k}_{-1}+\bar{k}_2=k_{-1}+k_2$. Using now (2.10c) we obtain the unique solution $\bar{k}=k$. There are, however, an infinite number of solutions if $k_1=0$ or $k_2=0$. These exceptional points form two planes in R^3, and thus are sets of zero measure and hence the model in uniquely structurally identifiable if ϵ_B is known.

(b) Consider the more general case with parameters $k=(k_1, k_{-1}, k_2, \epsilon_B)^T$. Since (2.10) consists only of three equations to determine four parameters, the model is unidentifiable.

Example 2.2. It is easy to show that the scheme in (1.1) is structurally identifiable, but not uniquely. The Laplace transform of (1.2) is given by

$$Y(s,k) = \frac{\epsilon_A s^2 + (\epsilon_A k_2 + \epsilon_B k_1)s + \epsilon_C k_1 k_2}{s^3 + (k_1+k_2)s + k_1k_2 s}\ [A]_0 \tag{2.11}$$

and with known $\bar{\epsilon}_A = \epsilon_A$ and $\bar{\epsilon}_C = \epsilon_C$ the independent equations of the form (2.8) are

$$\epsilon_A\bar{k}_2 + \bar{\epsilon}_B\bar{k}_1 = \epsilon_A k_2 + \epsilon_B k_1$$

$$\bar{k}_1 + \bar{k}_2 = k_1 + k_2 \tag{2.12}$$

$$\bar{k}_1\bar{k}_2 = k_1k_2$$

which clearly admit the second solution (1.3). The exceptional subsets are again $k_1=0$ and $k_2=0$, where the model is unidentifiable. Notice that $\bar{\epsilon}_B$ given

by (1.3) may be negative and since this is clearly unphysical the ambiguity is resolved in certain cases, depending on the value of the parameters k.

In Example 2.1 with ϵ_B unknown we have more parameters than equations and hence unidentifiability follows immediately. Though (2.8) generally contains at least as many equations as parameters, the analysis of structural identifiability is very simple.As shown by Vajda using the implicit function theorem,[13] the model is structurally identifiable if and only if rank $J(k)=p$ for some $k\epsilon\Omega$, where $J=\partial\phi/\partial k$ denotes the Jacobian matrix of ϕ, and p is the number of parameters. The condition is met if and only if det J(k) (or its principal minors in case of a nonsquare matrix) do not identically vanish. If rank $J(k)=q<p$ for all $k\epsilon\Omega$, then one can select p-q parameters such that by fixing their values the model becomes structurally identifiable with respect to the remaining free parameters. Therefore, the integer q=rank J(k) is called the number of determinable parameters and will play an important role in further sections. As shown in Example 2.1, the number of determinable parameters is 3, since the model is identifiable with ϵ_B fixed.

Remark 2.1. Since the elements of J(k) are analytic functions of the parameters, rank J(k) achieves its maximum value at almost every $k\epsilon\Omega$. Throughout the paper rank J(k) denotes this maximum or "generic" rank of J(k).

Remark 2.2. Though identifiability properties have been defined for a single experiment specified by the initial concentrations $x_0^{(1)}$ the analysis can easily be extended to a set of experiments with initial conditions $x_0^{(1)}\epsilon I\subset R^n$. Indeed, the elements of J are also analytic functions of the components in $x_0^{(1)}$ and similarly to Remark 2.1, structural identifiability at a single $x_0\epsilon I$ implies identifiability at almost every $x_0^{(1)}\epsilon I$. There can exist, however, exceptional points where identifiability is lost, e.g., selecting a

stationary state as initial concentrations in the experiment.

Remark 2.3. The number of components in the "moment" invariant vector ϕ is at most 2mn, which is an upper bound on the number of determinable parameters. Following the reaction by the use of a single quantity and considering only rate coefficients as unknown parameters the upper bound is 2n-1.

Though generating the Laplace transform (2.5) of the response function is usually not very tedious, it can considerably be simplified by taking advantage of the specific method proposed by Bossi et al.[14] As discussed, for testing structural identifiability we also need the Jacobian matrix $\partial\phi/\partial k$ and its determinant (or principal minors), which can easily be evaluated. The analysis of unique structural identifiability requires, however, the solution of the polynomial equations (2.8). It should be emphasized that this step may be considerably more difficult than in Example 2.2, where nonuniqueness follows from interchangeability of the rate coefficients. As shown by Norton[15] in his exhaustive analysis of first-order reaction schemes (linear compartmental models) with 3 species, sources of nonuniqueness are generally more subtle and the functions relating the different solutions more complicated. While symbolic languages such as REDUCE and algebraic manipulation subroutines are valuable tools in solving the polynomial equations,[16] with some persistency the solution can also usually be obtained by hand.

2. Similarity transformation approach

This method is based on introducing the new variables $\tilde{x}$ defined by $x=T\tilde{x}$ into (2.1) and (2.2), where T is an n x n nonsingular matrix. The transformed system is then described by

$$\dot{\tilde{x}}(t,k) = T^{-1}A(k)T\tilde{x}(t,k), \qquad \tilde{x}_o(k) = T^{-1}x_o(k)$$

$$\tilde{y}(t,k) = C(k)T\tilde{x}(t,k). \tag{2.13}$$

By the algebraic equivalence theorem of linear system theory[17] a similarity transformation does not change the response function, thus $\tilde{y}(t,k)=y(t,k)$. Let f denote the vector formed by the entries in T and introduce the notation $T=T(f)$. Since T is arbitrary, the elements of f are a priori free with the only constraint $\det T(f)\neq 0$. While the response is invariant, the system matrices and initial conditions are changed according to

$$\tilde{A}(k,f) = T^{-1}(f)A(k)T(f), \tag{2.14a}$$

$$\tilde{C}(k,f) = C(k)T(f), \tag{2.14b}$$

and

$$\tilde{x}_o(k,f) = T^{-1}(f)x(k), \tag{2.14c}$$

where $\tilde{A}$, $\tilde{C}$, and $\tilde{x}_o$ depend on f in addition to the original parameters k. Now we check how the knowledge of the system structure restricts the possible values of f. For simplicity assume that C and x_o are completely known (i.e., do not depend on unknown parameters). Then $\tilde{C}=C$ and $\tilde{x}_o = x_o$, thus (2.14b) and (2.14c) imply the constraints

$$C = CT(f), \qquad x_o = T(f)x_o. \tag{2.15}$$

Further constraints follow from the structure of the matrix A. If $a_{ij}(k)=0$, then we also require $\tilde{a}_{ij}(k,f)=0$, where a_{ij} and $\tilde{a}_{ij}$ denote the corresponding entries in A and $\tilde{A}$, respectively. All constraints form a set of equations for the parameters f. This set always admits a nominal solution f^o such that $T(f^o)=I$ and the transformations (2.14) yield the original system matrices. The existence of a second solution $f\neq f^o$ means, however, that the knowledge of the response function y(t,k) and all the available structural constraints does not specify the transformation matrix T(f) and hence $\tilde{A}$, $\tilde{C}$, and $\tilde{x}_o$ uniquely; thus the model is not uniquely structurally identifiable.

Similarly, an infinite number of solutions for f shows unidentifiability of the model.

If C and/or x_o also depend on unknown parameters, constraints of the form (2.15) do not apply but we still have some constraints on f from the partial knowledge of C and x_o. A formal description of the method is rather lenghty[18] but it can easily be understood with the aid of the following example.

Example 2.3. We use the similarity transformation approach to solve the simple identifiability problem studied in Example 2.2. The reaction scheme in (1.1) is described by

$$A(k) = \begin{bmatrix} -k_1 & o & o \\ k_1 & -k_2 & o \\ o & k_2 & o \end{bmatrix}, \quad x_o(t) = \begin{bmatrix} x_{0,1} \\ o \\ o \end{bmatrix} \tag{2.16}$$

$$C = [\, \epsilon_A \quad \epsilon_B \quad \epsilon_C \,]$$

where $x_{0,1} = [A]_0$. The transformation matrix is a priori arbitrary, thus

$$T(f) = \begin{bmatrix} f_1 & f_2 & f_3 \\ f_4 & f_5 & f_6 \\ f_7 & f_8 & f_9 \end{bmatrix} \tag{2.17}$$

with the only constraint $\det T(f) \neq 0$. Since $x_0(k) = x_0$ is completely known, the constraint in (2.15) applies and yields $f_1 = 1$, $f_4 = 0$, and $f_7 = 0$. According to (2.14b) the transformed observation matrix $\tilde{C}(k,f) = (\tilde{\epsilon}_A \ \tilde{\epsilon}_B \ \tilde{\epsilon}_C)$ is given by the elements

$$\tilde{\epsilon}_A = \epsilon_A, \ \tilde{\epsilon}_B = f_2\epsilon_A + f_5\epsilon_B + f_8\epsilon_C, \ \tilde{\epsilon}_C = f_3\epsilon_A + f_6\epsilon_B + f_9\epsilon_C \tag{2.18}$$

Since $\tilde{\epsilon}_C = \epsilon_C$ is known, (2.18) gives $f_3 = 0$, $f_6 = 0$, and $f_9 = 1$, thus using the knowledge of x_0 and partial knowledge of C we end up with the transformation matrix

$$T(f) = \begin{bmatrix} 1 & f_2 & o \\ o & f_5 & o \\ o & f_8 & 1 \end{bmatrix} \tag{2.19}$$

Apply now (2.14a) to form the transformed matrix $\tilde{A}(k,f)$. Using the well known formula[19] $T^{-1}=(\text{adj } T)^T/(\det T)$ we obtain

$$T^{-1}(f) = \begin{bmatrix} 1 & -f_2/f_5 & o \\ o & 1/f_5 & o \\ o & -f_8/f_5 & 1 \end{bmatrix} \tag{2.20}$$

and hence

$$\tilde{A}(k,f) = \begin{bmatrix} -k_1(1+f_2/f_5) & -k_1 f_2(1+f_2/f_5) + k_2 f_2 & o \\ k_1/f_5 & k_1 f_2/f_5 - k_2 & o \\ -k_1 f_8/f_5 & -k_1 f_8 f_2/f_5 + k_2 f_8 + k_2 f_5 & o \end{bmatrix} \tag{2.21}$$

Since for the scheme (1.1) this matrix should be of the form (2.16), the constraints imposed on (2.21) are as follows:

$$\tilde{a}_{31}(k,f) = -k_1 f_8/f_5 = 0 \tag{2.22a}$$

$$\tilde{a}_{12}(k,f) = -f_2(k_1 - k_2 + k_1 f_2/f_5) = 0 \tag{2.22b}$$

$$\tilde{a}_{21}(k,f) = -\tilde{a}_{11}(k,f) = \tilde{k}_1 \tag{2.22c}$$

$$\tilde{a}_{32}(k,f) = -\tilde{a}_{22}(k,f) = \tilde{k}_2 \tag{2.22c}$$

Eq. (2.22a) implies $f_8=0$, whereas (2.22b) admits two solutions given by

$$f_2 = 0 \tag{2.23a}$$

and

$$f_2/f_5 = (k_2 - k_1)/k_1 \tag{2.23b}$$

Substituting (2.23a) into (2.22c) and (2.22d) gives $f_5=1$, thus $T=I$ and we find the nominal solution f^o that yields the original system. (2.23b) gives $f_5=k_1/k_2$ and using (2.22c), (2.22d), and (2.18) one obtains the second solution (1.3) for the parameters.

In this example the similarity transformation approach requires more calculations than the one based on Laplace transforms. Notice, however, that in Example 2.2 we had to solve a quadratic equation to obtain the two solutions, whereas $\tilde{a}_{12}$ in (2.22) is the product of two factors. In more complex cases, j solutions for the parameters frequently imply that (2.8) is reduced to a single polynomial equation of degree j and numerical methods may be required, whereas in the similarity transformation approach we may have fewer variables (i.e., the elements of f remaining free after requiring invariance of C and x_0) and equations of somewhat simpler structure. Furthermore, only the latter method enables one to generate all reaction schemes indistinguishable from a given one. In particular, we will use the matrix (2.21) to solve this problem in Section IV.

Remark 2.4. Assume that taking into account all the available constraints there remain r free variables $f_1,\ldots,f_r$ in $\tilde{A}(k,f)$. Then the model is unidentifiable and r further constraints (e.g., fixed values for r parameters) are required to render the model structurally identifiable. Therefore, p-r=q is the number of determinable parameters, defined previously by q=rank $\partial\phi/\partial k$. This simple rule will be frequently used in the following sections.

III. Distinguishability

In this section we consider two different reaction schemes denoted by S and $\tilde{S}$, respectively. Both are described by models of the form (2.1) - (2.2). Let $y(t,k)$ and $\tilde{y}(t,\tilde{k})$ represent their response functions with the initial conditions $x_0^{(1)}=\tilde{x}_0^{(1)}$ and parameters $k\epsilon\Omega$ and $\tilde{k}\epsilon\tilde{\Omega}$, respectively, where Ω and $\tilde{\Omega}$ are the parameter sets. Extending the concept of parameter indistinguishability introduced in Section II to two different models, values $k\epsilon\Omega$ and $\tilde{k}\epsilon\tilde{\Omega}$ are said to be indistinguishable if

$$\tilde{y}(t,\tilde{k}) = y(t,k) \tag{3.1}$$

for all $t\epsilon T$. Since the values of k and $\tilde{k}$ are not a priori known, as a further generalization of the concept the models S and $\tilde{S}$ are said to be indistinguishable if for almost every parameter value $k\epsilon\Omega$ of the model S there exists an indistinguishable value $\tilde{k}\epsilon\tilde{\Omega}$ of $\tilde{S}$ and vice versa, thus for almost every $\tilde{k}\epsilon\tilde{\Omega}$ of $\tilde{S}$ there is an indistinguishable parameter $k\epsilon\Omega$ of S. Then with appropriate selection of parameters both models generate the same family of response functions which corresponds to the usual meaning of indistinguishability. Since the property is an equivalence relation between the parameter sets Ω and $\tilde{\Omega}$, indistinguishable models have also been called structurally equivalent.[20]

Let Y(s,k) and $\tilde{Y}(s,\tilde{k})$ denote the Laplace transforms of y(t,k) and $\tilde{y}(t,\tilde{k})$, respectively. Eq. (3.1) is satisfied if and only if

$$\tilde{Y}(s,\tilde{k}) = Y(s,k) \tag{3.2}$$

for all $s\epsilon C$. The components of Y(s,k) and $\tilde{Y}(s,\tilde{k})$ are rational functions of the form (2.7) and for each nonzero coefficient $\phi_j^i(k)$ of s^j in the numerator (denominator) polynomial of $Y_i(s,k)$ there should be a corresponding nonzero coefficient $\tilde{\phi}_j^i(\tilde{k})$ of s^j in the numerator (denominator) polynomial of $\tilde{Y}_i(s,\tilde{k})$. In this case Y(s,k) and $\tilde{Y}(s,\tilde{k})$ are said to be of the same symbolic form.[21] The same symbolic form is a necessary, but not sufficient condition for indistinguishability.[22] It implies, however, that listing the corresponding coefficients in ϕ and $\tilde{\phi}$ in the same order we can proceed to the analysis of the polynomial equations

$$\tilde{\phi}(\tilde{k}) = \phi(k) \tag{3.3}$$

where ϕ and $\tilde{\phi}$ denote the vectors of "moment" invariants for S and $\tilde{S}$, respectively. To establish indistinguishability one has to solve (3.3) both for $\tilde{k}$ in terms of k and for k in terms of $\tilde{k}$. The models are

indistinguishable if and only if both solutions exist at almost every $k\epsilon\Omega$ and $\tilde{k}\epsilon\tilde{\Omega}$, respectively. Frequently these solutions exist only over some open sets $\Omega_1 \underline{c}\Omega$ and $\tilde{\Omega}_1 \underline{c}\tilde{\Omega}$, then S and $\tilde{S}$ should be restricted to these subsets to ensure their indistinguishability.

Example 3.1.(a) In (1.4) and (1.5) we listed reaction schemes claimed to be indistinguishable from (1.1). With initial concentrations $[B]_o=[C]_o=0$ the Laplace transform of the response function for the latter model is given by (2.11). Evaluating $Y(s,k)$ for model S_3 in (1.5) we can immediately conclude that it has a different symbolic form and hence the two models are distinguishable.

(b) Now we test distinguishability of the model S_2 in (1.4) from (1.1), denoted here by $\tilde{S}$. For S_2 we have the Laplace transform

$$Y(s,k) = \frac{\epsilon_A s^2+(\epsilon_A k_3+\epsilon_B k_1+\epsilon_C k_2)s+\epsilon_B k_3(k_1+k_2)}{s^3+(k_1+k_2+k_3)s^2+k_3(k_1+k_2)s}\ [A]_o \tag{3.4}$$

which has the same symbolic form as (2.11) for $\tilde{S}$. Therefore, we consider equations of the form (3.3), given here by

$$\epsilon_A\tilde{k}_2 + \tilde{\epsilon}_B\tilde{k}_1 = \epsilon_A k_3 + \epsilon_B k_1 + \epsilon_C k_2 \tag{3.5a}$$

$$\epsilon_C\tilde{k}_1\tilde{k}_2 = \epsilon_B k_3(k_1+k_2) \tag{3.5b}$$

$$\tilde{k}_1+\tilde{k}_2 = k_1+k_2+k_3 \tag{3.5c}$$

$$\tilde{k}_1\tilde{k}_2 = k_3(k_1+k_2) \tag{3.5d}$$

Substituting (3.5d) into (3.5b) reduces the latter to the equation $\epsilon_C=\epsilon_B$. Since ϵ_B is a free parameter of the model S_2, whereas ϵ_C is a known constant, (3.5) can be solved for $\tilde{k}=(\tilde{k}_1,\tilde{k}_2,\tilde{\epsilon}_B)^T$ in terms of $k=(k_1,k_2,k_3,\epsilon_B)^T$ only at this particular value of ϵ_B, otherwise the equations are contradictory. Therefore, the models are distinguishable. In practical terms, the two models generate the same response function if and only if $\epsilon_B=\epsilon_C$. This

particular choice is, however, meaningless since species B and C are then lumped and both models loose identifiability. For example, the common factor $s+k_2$ appears in the numerator and denominator polynomials of (2.11). Notice, however, that the models become indistinguishable if ϵ_C is an additional free parameter of (1.1).

Example 3.2. Consider the schemes (1.1) and S_1 in (1.4). By (2.9) and (2.11) the Laplace transforms are of the same symbolic form, whereas equations (3.3) are given by

$$\epsilon_A\tilde{k}_2 + \tilde{\epsilon}_B\tilde{k}_1 = \epsilon_A(k_{-1}+k_2) + \epsilon_B k_1 \tag{3.6a}$$

$$\tilde{k}_1 + \tilde{k}_2 = k_1 + k_{-1} + k_2 \tag{3.6b}$$

$$\tilde{k}_1\tilde{k}_2 = k_1 k_2 . \tag{3.6c}$$

We have a special situation here, since S is obtained by setting $k_{-1}=0$ in S and hence $\tilde{S}$ is called a submodel of S. Thus for any $\tilde{k}=(\tilde{k}_1,\tilde{k}_2,\tilde{\epsilon}_B)^T\epsilon\tilde{\Omega}$ the parameter value $k=(\tilde{k}_1,0,\tilde{k}_2,\tilde{\epsilon}_B)^T\epsilon\Omega$ satisfies (3.6) and we have to solve the equations only for $\tilde{k}$ in terms of k. The solution exists for all $k\epsilon R^4$ and the models are indistinguishable.

Requiring solution of polynomial equations makes the analysis rather tedious. In a number of cases, however, we can take advantage of simple conditions and avoid calculations. We list here the basic results with the underlying mathematical ideas .

Proposition 1. Let q and $\tilde{q}$ denote the numbers of determinable parameters in S and $\tilde{S}$, respectively. If $q\neq\tilde{q}$ then S and $\tilde{S}$ are distinguishable.[23]

Proof: Since rank $\partial\phi/\partial k=q$, the set $\phi(\Omega)$ is a q-dimensional manifold in some Euclidean space of dimension $r\geq q$. Indistinguishability implies $\phi(\Omega)=\tilde{\phi}(\tilde{\Omega})$ and hence $q=\tilde{q}$.

Proposition 2. If $\tilde{S}$ is a submodel of S and the number of determinable parameters is q in both models, then there exist open sets $\Omega_1\epsilon\Omega$ and $\tilde{\Omega}_1\subset\tilde{\Omega}$ such

that restricted to these sets S and $\tilde{S}$ are indistinguishable.[24] Proof: For simplicity let $k=(k_1,\ldots,k_q,k_{q+1})^T$ and $\tilde{k}=(\tilde{k}_1,\ldots,\tilde{k}_q)$, thus $\tilde{S}$ is obtained by setting $k_{q+1}=0$ and it is structurally identifiable. Since $\tilde{S}$ is a submodel of S, we solve (3.3) only for $\tilde{k}$ in terms of k. The solution is $\tilde{k}_i=k_i$, $i=1,\ldots,q$ if $k_{q+1}=0$. Since rank $\partial\tilde{\phi}/\tilde{k}=q$, by virtue of the general implicit function theorem this solution can be extended onto an open neighborhood Ω_1 of the point $(k_1,\ldots,k_q,0)$ in $\Omega\subset R^{q+1}$.

Proposition 3. If $\tilde{S}$ and $\tilde{\tilde{S}}$ are submodels of S and all three models have the same number of determinable parameters, then there exist open sets $\tilde{\Omega}_1\subset\tilde{\Omega}$ and $\tilde{\tilde{\Omega}}_1\subset\tilde{\tilde{\Omega}}$ such that restricted to these sets $\tilde{S}$ and $\tilde{\tilde{S}}$ are indistinguishable.[25] Proof: It follows from Proposition 2 and transitivity of equivalence.

Proposition 4. Consider structurally identifiable models S and $\tilde{S}$ with the same number p of parameters. If $Y(s,k)$ and $\tilde{Y}(s,\tilde{k})$ are of the same symbolic form and $\partial\phi/\partial k$ is a square matrix, then there exist open sets $\Omega_1\subset\Omega$ and $\tilde{\Omega}_1\subset\tilde{\Omega}$ such that restricted to these sets S and $\tilde{S}$ are indistinguishable.[26] Proof: Since $\partial\phi/\partial k$ and $\partial\tilde{\phi}/\partial\tilde{k}$ are p×p matrices of full rank, $\phi(\Omega)$ and $\tilde{\phi}(\tilde{\Omega})$ are open sets in R^p. Their intersection contains some open neighborhood of $0\in R^p$, and hence is not empty.

Using these propositions our examples can be solved without calculations. Since the number of determinable parameters is 3 for (1.1), 4 for S_2 in (1.4), and 2 for S_3 in (1.5), by Proposition 1 all these models are distinguishable as shown in Example 3.1. Both models considered in Example 3.2 have 3 determinable parameters and hence are indistinguishable by Proposition 2. By Proposition 4 considering ϵ_c as an additional free parameter implies that models (1.1) and S_2 become indistinguishable as noticed in Example 3.1, part (b).

IV. Exhaustive modeling

Given a reaction scheme S and experimental conditions specified by the initial condition $x_o^{(1)}$ and observation matrix C we use the similarity transformation approach to generate all the first-order reaction schemes $\tilde{S}_1, \tilde{S}_2, \ldots, \tilde{S}_r$ that are indistinguishable from S. As shown in Section II, we first construct the transformation matrix T(f) that preserves all known properties of x_o and C. With this T(f) the transformations (2.14) yield the most general linear system that generates the original response function at any $k \in \Omega$ and f. In identifiability analysis we imposed further constraints in order to preserve the structure of the reaction scheme and checked uniqueness of the corresponding transformation. Looking for different models we don't consider structural constraints here, but find the first-order reaction scheme that corresponds to the transformed system matrix $\tilde{A}(k,f)$ by introducing the parameters

$$\tilde{k}_{ji} = \tilde{a}_{ji} \ , \ i \neq j \ ; \quad \tilde{k}_{oi} = -\sum_{j=i}^{n} \tilde{a}_{ji} \ , \tag{4.1}$$

where $\tilde{k}_{ji}$ is the rate coefficient of a first-order reaction consuming species i and producing species j, with index j=0 denoting a species not taken into account among the n species of the original model. If no such species can exist, then the mass conservation condition is assured by the constraints

$$\sum_{j=i}^{n} \tilde{a}_{ji}(k,f) = 0 \tag{4.2}$$

thereby further reducing the free components of f. The transformation (2.14) and the reparametrization (4.1) then give rise to the most general (in terms of the number of nonzero rate coefficients) first-order reaction scheme $\tilde{S}$ such that with the parameters

$$\tilde{k}_{ji}(k,f) = \tilde{a}_{ji}(k,f) \tag{4.3}$$

we have $\tilde{y}(t,\tilde{k})=y(t,k)$. As will be shown, any model indistinqishable from the original S is a submodel of $\tilde{S}$ and hence this latter is said to be the "frame" model. In particular, at the nominal value $f=f^0$ the model $\tilde{S}$ reduces to S, thus S itself is a submodel of $\tilde{S}$. Let q and $\tilde{q}$ denote the numbers of determinable parameters in S and $\tilde{S}$, respectively. If $\tilde{q}=q$, then S and $\tilde{S}$ are indistinguishable by Proposition 2.$\tilde{S}$ is unidentifiable and by Proposition 3 its submodels with q determinable parameters form the set of reaction schemes indistinguishable from S. The next example illustrates this case.

Example 4.1. As shown in Example 2.3., the transformed matrix for the scheme (1.1) is given by (2.21). Apply the constraints (4.2) to the columns of (2.21). It can be readily verified that these are satisfied by

$$f_8 = 1-f_2-f_5$$

thus we eliminated f_8 from (2.21). Introduce the parameters

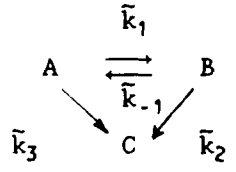
$$\tilde{k}_1 = k_1/f_5 \tag{4.5a}$$

$$\tilde{k}_{-1} = -k_1 f_2(1+f_2/f_5) + k_2 f_2 \tag{4.5b}$$

$$\tilde{k}_2 = k_1 f_2(f_2-1)/f_5 + f_2(k_1-k_2) + k_2 \tag{4.5c}$$

$$\tilde{k}_3 = k_1(1-f_2-f_5)/f_5 \tag{4.5d}$$

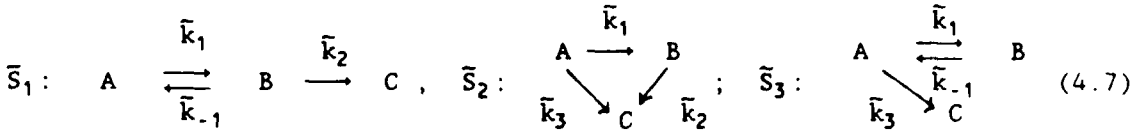
then the "frame" model obtained is given by



(4.6)

Since we have 5 parameters (i.e., 4 rate coefficients and the extinction coefficient ϵ_B), whereas f_2 and f_5 are free, by Remark 2.4 the number of determinable parameters $\tilde{q}$ in (4.6) is 3. Therefore, (1.1) and (4.6) are indistinguishable and the set of all reaction schemes indistinguishable from

(1.1) consists of the submodels

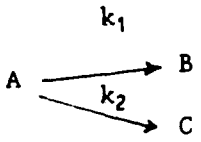


(4.7)

of (4.6) with 3 determinable parameters and (4.6) itself. This can also be verified by solving (4.5) for k and f in terms of $\tilde{k}$. Notice that (4.6) has no identifiable submodels with 3 determinable parameters, thus all models in (4.7), though indistinguishable from (1.1), are unidentifiable. The example gives the correct solution of the problem stated in Section I.

Consider now the case $\tilde{q}>q$. Then the original S and the "frame" model $\tilde{S}$ are distinguishable by Proposition 1. As discussed, for any $k\in\Omega$ (4.3) gives a parameter value for the "frame" model $\tilde{S}$ such that $\tilde{y}(t,\tilde{k})=y(t,k)$, but when selecting a point $\tilde{k}\in\tilde{\Omega}$ generally there exist no k and f that satisfy (4.3). In other words, the "frame" model is too large in the sense that it can generate all response functions of the original model, but the converse is not true. Therefore, the parameter set $\tilde{\Omega}$ should be restricted by considering the submodels of $\tilde{S}$ with $q < \tilde{q}$ determinable parameters and trying to solve (4.3) for k and f. Though these submodels are the only candidates for being indistinguishable from S, actual indistinguishability should be checked by direct calculation in each case as shown in the next example.

Example 4.2. We generate the indistinguishable schemes for the mechanism



(4.8)

where $[B]_0=[C]_0=0$ and the observed quantity is [B]. Since the rate expressions do not depend on [C], it is sufficient to consider the kinetic equations for $x_1=[A]$ and $x_2=[B]$:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -(k_1+k_2) & 0 \\ k_1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} ; \qquad x_o = \begin{bmatrix} x_{1,0} \\ 0 \end{bmatrix} \tag{4.9}$$

$$y = [\,0 \qquad 1\,] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

The parameter vector is $k=(k_1,k_2)^T$, and the model is structurally identifiable, thus $q=2$. The most general transformation matrix satisfying the constraints of the form (2.15) is

$$T(f) = \begin{bmatrix} 1 & f \\ 0 & 1 \end{bmatrix} \tag{4.10}$$

depending on a single parameter f. By (2.14a)

$$\tilde{A}(k,f) = \begin{bmatrix} (k_1+k_2) - k_1 f & -(k_1+k_2)f - k_1 f^2 \\ k_1 & k_1 f \end{bmatrix} \tag{4.11}$$

Because of the presence of species C in the system, constraints (4.2) are not imposed. Introducing the new parameters

$$\begin{aligned} \tilde{k}_1 &= \tilde{a}_{21} = k_1 \\ \tilde{k}_{-1} &= \tilde{a}_{12} = -f(k_1+k_2+k_1 f) \\ \tilde{k}_2 &= -\tilde{a}_{22}-\tilde{a}_{12} = f(k_2+k_1 f) \\ \tilde{k}_3 &= -\tilde{a}_{11}-\tilde{a}_{21} = k_2+k_1 f \ . \end{aligned} \tag{4.12}$$

we obtain the "frame" model with the structure (4.6), but only the 4 rate coefficients as parameters. Since these depend on a single f, by Remark 2.4 the number of determinable parameters is $q=3$. Therefore, (4.8) and the "frame" model are distinguishable. This can easily be proved also by trying to solve (4.12) for k and f in terms of $\tilde{k}$. Since $\tilde{k}_1 \neq 0$ and (4.8) has 2 determinable parameters, the candidate models for being indistingushable from (4.8) are only the submodels

$$\tilde{S}_1: \quad A \underset{\tilde{k}_{-1}}{\overset{\tilde{k}_1}{\rightleftharpoons}} B \qquad\qquad \tilde{S}_2: \quad A \xrightarrow{\tilde{k}_1} B \xrightarrow{\tilde{k}_2} C \tag{4.13}$$

of (4.6) with 2 determinable parameters. To obtain $\tilde{S}_1$ we assume $\tilde{k}_2=\tilde{k}_3=0$, which is satisfied if $f=-k_2/k_1$, and hence $\tilde{S}_1$ and (4.8) are indistinguishable. There exists, however, no f value that satisfies the equations $\tilde{k}_{-1}=\tilde{k}_3=0$ at all k, thus $\tilde{S}_2$ is distingushable from the other models. Notice that $\tilde{S}_1$ in (4.13), the only model indistinguishable from (4.8), is structurally identifiable in contrast to the models found in Example 4.1.

As will be shown in our last example, for a slightly more complex reaction scheme there may exist several identifiable models that are indistinguishable from the original one.

Example 4.3. Interpretation of growth and decay data through the use of the reaction scheme

$$\begin{array}{ccccc} & & k_1 & & k_2 \\ & & \nearrow & & \longrightarrow \\ A & & & B & & D \\ & & \searrow & & \\ & & k_3 & & \\ & & & C & \end{array} \tag{4.14}$$

has been discussed by Carrington.[6] In addition to a statistical analysis, he showed that with the initial conditions $[B]_o=[C]_o=[D]_o=0$ and observing only [B], (4.14) is not uniquely identifiable with 2 solutions for the parameters. We solve here the distinguishability problem also stated by Carrington in the introduction of his paper. As in Example 4.2, it is sufficient to consider the kinetic equations for the species A and B only. The transformation matrix T(f) is (4.10) and we obtain the "frame" model (4.6) as in the previous examples.Its parameters are, however,defined now by the relations

$$\begin{aligned} \tilde{k}_1 &= k_1 \\ \tilde{k}_{-1} &= -f(k_1 f + k_1 + k_3 - k_2) \\ \tilde{k}_2 &= k_1 f^2 + (k_3 - k_2) f + k_2 \\ \tilde{k}_3 &= k_1 f + k_3 \end{aligned} \tag{4.15}$$

Both (4.14) and the "frame" model have 3 determinable parameters and hence are indistinguishable. The further indistingiushable models are the two submodels

$$\tilde{S}_1: \quad C \xleftarrow{\tilde{k}_3} A \underset{\tilde{k}_{-1}}{\overset{\tilde{k}_1}{\rightleftharpoons}} B \qquad \tilde{S}_2: \quad A \underset{\tilde{k}_{-1}}{\overset{\tilde{k}_1}{\rightleftharpoons}} B \xrightarrow{\tilde{k}_2} C \tag{4.16}$$

of (4.6) with 3 determinable parameters.Notice that both latter models are structurally identifiable.

Though the above results follow immediately from the propositions in Section III, it is worthwhile to check solvability of (4.15). For example, indistinguishability of $\tilde{S}_1$ and (4.14) requires $\tilde{k}_2=0$. The solution for f is real if and only if the parameters of (4.14) satisfy the inequality constraint $D=(k_3-k_2)^2-4k_1k_2>0$. Thus the domain of indistinguishability is restricted to an open subset of the original parameter space $\Omega=R^3$. The calculation also shows that $\tilde{S}_1$ is not uniquely identifiable with two solutions. Since the solution of $\tilde{k}_3=0$ always exists and is unique, in the case of $\tilde{S}_2$ indistinguishability is unconstrained and the model is uniquely structurally identifiable.

V. Conclusions

Assuming a reaction scheme and following the reaction by observing the quantities accessible to measurements, the experiment does not necessarily provide sufficient information to derive unique values for the rate

coefficients or other unknown parameters included in the kinetic model. Similarly, there may exist further reaction schemes that are able to generate the same values for the observed varibles.

Both uniqueness problems are relatively easy to solve in the case of first-order reaction systems with observed quantities depending linearly on the concentrations. The problems of identifiability (i.e., uniqueness of the parameters) and distinguishability (i.e., uniqueness of the reaction scheme) are so closely related that the same analytical tools can be used to solve them. A very simple method is based on the application of a Laplace transform to the kinetic equations and results in a set of polynomial equations for the parameters. By solving these equations one can check identifiability and find the equivalent solutions for the parameters if the model is not uniquely identifiable. The approach can be extended to test distinguishability of two different first-order reaction schemes. The second method we used is based on state-space similarity transformations. It may be less convenient to study identifiability than with the Laplace transformation approch, but it can be used to solve the more general problem of exhaustive modeling, thus to generate all the first-order reaction schemes that are indistinguishable from a given one. Calculations can be considerably simplified by taking advantage of general conditions for indistinguishability, also formulated in the paper.

Acknowledgment

The authors acknowledge support for this research from the Office of Naval Research and the Air Force Office of Scientific Research.

References and Notes

(1) On leave from Laboratory for Chemical Cybernetics, Eotvos Lorand University, H-1088 Budapest, Hungary.

(2) Reilly, P. M.; Bajramov, R.; Blau, G. E.; Branson, D. R.; Sauerhof, M. W. Can. J. Chem. Eng., 1977, 55, 614-622.

(3) Alcock, N. W.; Benton, D. J.; Moore, P. Trans. Faraday Soc. 1970, 66, 2210-13.

(4) Jackson, W. G.; Harrowfield, J. M.; Vowles, P. D. Int. J. Chem. Kinet., 1977, IX, 535-47.

(5) Milligan, W. O.; Mullica, D. F.; Pennington, D. E.; Lok, C. K. C.; Kwong, D. W. J. Comp. Chem., 1984, 8, 285-98.

(6) Carrington, T. Int. J. Chem. Kinet., 1982, 14, 517-34.

(7) (a) Grewal, M. S.; Glover, K. IEEE Trans. Auto. Contr., 1976, 21, 833-837

(b) Reid, J. G. IEEE Trans. Auto. Contr., 1977, 22, 242-46.

(c) Vajda, S. Int. J. Syst. Sci., 1983, 14, 1229-47.

(d) Nguyen, V. V.; Wood, E. F. SIAM Rev., 1982, 24, 34-51.

(8) For reviews and recent results see

(a) Jacquez, J. A. Math. Comput. Simul., 1982, XXIV, 452-59.

(b) Walter, E. "Identifiability of State Space Models", Springer: Berlin, 1982.

(c) Eisenfeld, J. Math. Biosci., 1985, 77, 229-43.

(d) Eisenfeld, J. Math. Biosci., 1986, 79, 209-20.

(e) Delforge, J. Math. Biosci., 1984, 65, 51-68.

(f) Delforge, J. Math. Biosci., 1986, 81, 127-44.

(9) (a) Park, S.W.; Himmelblau, D. M. Chem. Eng. J., 1982, 25, 163-74.

(b) Happel, J.; Walter, E.; Lecourtier, Y. Ind. Eng. Chem. Fundamentals, 1986, 25, 704-712.

(c) Walter, E.; Lecourtier, Y.; Happel, J.; Kao, J.-Y. AIChE J. 1986, 32, 1360-1366.

(10) See, e. g., Cobelli, C.; Lepschy, A.; Romanin-Jacur, G. Math. Biosci., 1979, 44, 1-18.

(11) Spiegel, M. R. "Laplace Transforms", McGraw Hill: New York, 1968.

(12) Vajda, S. Math. Biosci., 1981, 55, 39-64.

(13) Lemma 2 in ref 12.

(14) Bossi, A.; Cobelli, C.; Colussi, L.; Romanin-Jacur, G. Math. Biosci., 1977, 43, 187-98.

(15) Norton, J. P. Math. Biosci., 1982, 60, 89-108.

(16) Raksanyi, A.; Lecourtier, Y.; Walter, E.; Venot, A. Math. Biosci. 1985, 77, 245-66.

(17) E. g., Chen, C. T. "Introduction to Linear System Theory", Holt, Rinehart, and Winston: New York, 1970.

(18) Walter, E.; Lecourtier, Y. Math. Biosci., 1981, 56, 1-25.

(19) See, e. g., Amundson, N. R. "Mathematical Methods in Chemical Engineering. Matrices and their Application", Prentice-Hall, Englewood Cliffs, 1966.

(20) For a detailed discussion of the concept see references 12 and 21.

(21) Vajda, S. Math. Biosci., 1984, 69, 57-75.

(22) Counterexamples are given in ref 21.

(23) Lemma 4 in ref 12.

(24) Corollary 1 in ref 12.

(25) Corollary 2 in ref 12.

(26) Theorem 1 in ref 21.

A GENERAL ANALYSIS ON EXACT LUMPING IN CHEMICAL KINETICS

Genyuan Li and Herschel Rabitz

Department of Chemistry

Princeton University

Princeton, New Jersey 08540

Abstract

A general analysis of exact lumping is presented. This analysis can be employed to any reaction system with n species described by a set of first order ordinary differential equations $d\mathbf{y}/dt = \mathbf{f}(\mathbf{y})$, where $\mathbf{y}$ is an n-dimensional vector; $\mathbf{f}(\mathbf{y})$ is an arbitrary n-dimensional function vector. Here we only consider lumping by means of a rectangular constant matrix M(i.e., $\hat{\mathbf{y}} = M\mathbf{y}$, where M is a row-full rank matrix and $\hat{\mathbf{y}}$ has lower dimension than $\mathbf{y}$). It is found that a reaction system is exactly lumpable if and only if the intersection of the invariant or the null subspaces of the Jacobian matrix $J(\mathbf{y})$ of $\mathbf{f}(\mathbf{y})$ for all values of $\mathbf{y}$ is nonempty. The intersection is the null space of the lumping matrix M. If the dimension of the intersection is less than n, nontrivial lumping schemes can be obtained. It is proved that the Jacobian matrix can be represented as a linear combination of certain constant matrices and the intersection of the invariant or the null subspaces of the constant matrices is just that of the Jacobian matrix. After the determination of the intersections, all possible lumping matrices can be obtained. The kinetic equations of the lumped system can be described as $d\hat{\mathbf{y}}/dt = M\mathbf{f}(\bar{M}\hat{\mathbf{y}})$, $\bar{M}$ is any generalized inverse of M satisfying $M\bar{M} = I_{\hat{n}}$. Several implications of these lumpability conditions are investigated as well as illustrated by some simple examples.

I. INTRODUCTION

A problem which frequently arises in the study of chemical kinetics is the high dimensionality and high degree of coupling of the reaction system. For example, in many realistic chemical processes, particularly those related to petrochemistry, industrial processes, combustion phenomena and atmospheric chemistry, the number of reacting species can often exceed 10^2 - 10^3. It is impractical to incorporate the kinetic equations for each species. Consequently, lumping, by which several species are treated as a single component, is a necessity. Thus one desires to reduce the reaction mixture into a small number of lumps in the kinetic study for practical purposes.

For different reaction systems the suitable ways of lumping will likely be different. Even for a given system, there could be many lumped models, depending on the objectives. However, one is not able to lump a system arbitrarily, because it is not always possible to find a model or a set of differential equations describing the behavior of the lumped species. For lack of theoretical guidance, researchers have often spent many years trying to find adequate lumping schemes by trial and error. The modelling of catalytic cracking for petroleum is a typical example.[1] Confounding this approach is the fact that the true lumped "species" may actually be a combination or function of the original physical species.

Prior research clearly suggests the need for a rigorous study of lumping which can give useful guidelines for choosing lumps. Wei and Kuo[2] gave a lumping analysis of unimolecular reaction system and their work was extended by Ozawa[3] and Bailey[4,5]. One of the authers[6] presented a lumping analysis for uni- and/or bimolecular reaction systems. Such research has been largely confined to the uni- and/or bimolecular reaction systems with the focus on establishing the necessary and sufficient conditions for "exact lumping". These analyses have shown that exact

lumping by a network of uni- and/or bimolecular reactions is feasible only under a very restrictive set of conditions. Studies of the pitfalls and magnitude of errors in the use of empirical rate expressions for lumping many independent single or consecutive reactions were presented by Luss and his co-workers.[7-10] Unfortunately until now lumping theory was not sufficiently developed to give useful guidelines as to which lumps to choose for many problems. There are still at least two important problems within exact lumping, which have not been solved yet.

1. There is no known a priori way to determine the lumping scheme.

2. The kinetic equations can have higher order nonlinearities than quadratic. For instance, this situation can arise in the presence of termolecular reactions or when one uses equilibrium or steady-state assumptions to omit the intermediates in reactions. In addition, nonisothermal processes or the use of empirical rate laws can lead to highly nonlinear kinetic equations. Therefore a general lumping analysis capable of treating arbitrary physical non-linearities is necessary.

Considering this situation, a general analysis of exact lumping is presented in this paper. It can be used for any reaction system and the previously studied lumping analyses of uni- and/or bimolecular reaction systems are special cases of this analysis. In addition, this analysis can also be applied in other problems described by a set of first order ordinary differential equations, such as problems arising in classical molecular mechanics, chemical engineering and control theory.

Section II of this paper presents the conditions under which a reaction system is exactly lumpable and the corresponding kinetic equations of the lumped system. In section III, the relationship between the Jacobian matrix and its intersection of the invariant or the null subspaces is discussed and the methods to determine the intersections are derived. Section IV provides some simple examples to which the general lumping method is applied. Section V presents a discussion of the results.

II. THE THEORY OF EXACT LUMPING

A . CONDITIONS UNDER WHICH A REACTION SYSTEM IS EXACTLY LUMPABLE

Suppose the kinetics of an n-component reaction system can be described by

$$d\mathbf{y}/dt = \mathbf{f}(\mathbf{y}), \tag{1}$$

where $\mathbf{y}$ is an n-composition vector; $\mathbf{f}(\mathbf{y})$ is an arbitrary n-function vector, which does not contain t explicitly.

Here we only consider a special class of lumping by means of an $\hat{n} \times n$ constant matrix M with rank $\hat{n}(\hat{n} < n)$. If a system can be exactly lumped by the matrix M, it means that for

$$\hat{\mathbf{y}} = M\mathbf{y} \tag{2}$$

we can find an $\hat{n}$-function vector $\hat{\mathbf{f}}(\hat{\mathbf{y}})$ such that

$$d\hat{\mathbf{y}}/dt = \hat{\mathbf{f}}(\hat{\mathbf{y}}). \tag{3}$$

If y_i is not lumped, row i of M is a unit vector $\mathbf{e}_i^T = (00...010...0)$, and $\hat{y}_i = y_i$. In this case, since the lumping is exact, the solutions for y_i and $\hat{y}_i$ by Equations 1 and 3 are the same. However, Equation 3 is simpler.

From Equations 1 and 2 we have

$$d\hat{\mathbf{y}}/dt = Md\mathbf{y}/dt = M\mathbf{f}(\mathbf{y}), \tag{4}$$

and upon comparing Equations 3 and 4 we have

$$M\mathbf{f}(\mathbf{y}) = \hat{\mathbf{f}}(\hat{\mathbf{y}}). \tag{5}$$

Differentiating both sides of Equation 5 with respect to $\mathbf{y}$, we obtain

$$MJ(\mathbf{y}) = \hat{J}(\hat{\mathbf{y}})M, \tag{6}$$

where $J(\mathbf{y})$ and $\hat{J}(\hat{\mathbf{y}})$ are the Jacobian matrices of $\mathbf{f}(\mathbf{y})$ and $\hat{\mathbf{f}}(\hat{\mathbf{y}})$, respectively. As the rank of M is $\hat{n}$, there must exist generalized inverses $\bar{M}$ of matrix M satisfying

$$M\bar{M} = I_{\hat{n}}, \tag{7}$$

where $I_{\hat{n}}$ is $\hat{n}$-identity matrix. Multiplying both sides of Equation 6 from the right by $\bar{M}$, we have

$$MJ(\mathbf{y})\bar{M} = \hat{J}(\hat{\mathbf{y}})M\bar{M} = \hat{J}(\hat{\mathbf{y}}). \tag{8}$$

Substituting Equation 8 into Equation 6, we obtain

$$MJ(\mathbf{y}) = MJ(\mathbf{y})\bar{M}M,$$

$$MJ(\mathbf{y})(I_n - \bar{M}M) = 0, \tag{9}$$

where I_n is n-identity matrix.

Equation 9 is the fundamental equation in exact lumping. Although it was derived by differentiation of Equation 5, it is not a local perturbation theory result. This comment follows from the fact that we demand that Equation 6 and subsequent ones be valid for all physical values of $\mathbf{y}$. It is easy to prove that the image $\bar{\mathbf{x}}$ of any vector $\mathbf{x}$ upon mapping by $(I_n - \bar{M}M)$ is in the invariant subspace of $(I_n - \bar{M}M)$. Since

$$\bar{\mathbf{x}} = (I_n - \bar{M}M)\mathbf{x},$$

and

$$(I_n - \bar{M}M)^2 = (I_n - \bar{M}M),$$

then

$$(I_n - \bar{M}M)\bar{\mathbf{x}} = \bar{\mathbf{x}}.$$

However, the invariant subspace of $(I_n - \bar{M}M)$ is the null space of M, since

$$M\bar{\mathbf{x}} = M(I_n - \bar{M}M)\mathbf{x} = (M - M\bar{M}M)\mathbf{x}$$
$$= (M - M)\mathbf{x} = \mathbf{0}. \tag{10}$$

For any vector $\mathbf{x}$ in n-dimensional space, Equation 9 shows that

$$MJ(\mathbf{y})(I_n - \bar{M}M)\mathbf{x} = \mathbf{0},$$

$$MJ(\mathbf{y})\bar{\mathbf{x}} = \mathbf{0}.$$

Let

$$\tilde{\mathbf{x}} = J(\mathbf{y})\bar{\mathbf{x}}.$$

Then

$$M\tilde{\mathbf{x}} = \mathbf{0}. \tag{11}$$

Comparing Equations 10 *and* 11, *we find that if* $\bar{\mathbf{x}}$ is in the null space of M, then so is $\tilde{\mathbf{x}}$. This is valid only if one of the following two conditions is satisfied: 1) The null space of M is the invariant subspace of $J(\mathbf{y})$. Therefore, after mapping by $J(\mathbf{y})$, the image of vector $\bar{\mathbf{x}}$ in the null space of M is still in the same space; 2) The null space of M is also the null subspace of $J(\mathbf{y})$. In this case, $\tilde{\mathbf{x}}$ is a null vector and Equations 10 and 11 trivially hold. Notice that these arguments are valid for any value of $\mathbf{y}$. Thus the conclusion is that there exists a nontrivial matrix M only if the intersection of the invariant or the null subspaces of the Jacobian matrix $J(\mathbf{y})$ for all values of $\mathbf{y}$ is nonempty. This condition is also sufficient. Let the intersection be spanned by the column vectors of matrix X. Then we can use X to represent the intersection. Since the intersection is a subspace of the n-dimensional space, the column number of X is less than its row number. If the intersection exists, one

can choose this intersection as the null space of the lumping matrix M. Then we have

$$MX = 0. \tag{12}$$

Transposing Equation 12, we obtain

$$X^T M^T = 0. \tag{13}$$

There are an infinite number of solutions of M for a given X. This equation can be considered as a set of linear algebric equations

$$X^T \mathbf{m} = \mathbf{0}. \tag{14}$$

All the linearly independent solutions of $\mathbf{m}$ compose the matrix M. In some situations we may desire to keep a number of species, say p, unlumped. Without loss of generality we can consider the first p species unlumped and all lumped species are composed of others. In this case the lumping matrix can be expressed as

$$M = \begin{pmatrix} I_p & 0 \\ 0 & M_1 \end{pmatrix}, \tag{15}$$

where I_p is the p-identity matrix; M_1 is an $(\hat{n}-p)\times(n-p)$ matrix. There is an extra restriction on determination of M_1 described below. Let the Jacobian matrices $J(\mathbf{y})$ and $\hat{J}(\hat{\mathbf{y}})$ be blocked as follows:

$$J(\mathbf{y}) = \begin{pmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{pmatrix},$$

$$\hat{J}(\hat{\mathbf{y}}) = \begin{pmatrix} \hat{J}_{11} & \hat{J}_{12} \\ \hat{J}_{21} & \hat{J}_{22} \end{pmatrix},$$

where J_{11}, J_{12}, J_{21} and J_{22} are $p \times p, p \times (n-p), (n-p) \times p$ and $(n-p) \times (n-p)$ matrices; $\hat{J}_{11}, \hat{J}_{12}, \hat{J}_{21}$ and $\hat{J}_{22}$ are $p \times p, p \times (\hat{n}-p), (\hat{n}-p) \times p$ and $(\hat{n}-p) \times (\hat{n}-p)$ matrices, respectively. Using Equation 6, we have

$$J_{11} = \hat{J}_{11}, \tag{16}$$

$$M_1 J_{21} = \hat{J}_{21}, \tag{17}$$

$$M_1 J_{22} = \hat{J}_{22} M_1, \tag{18}$$

$$J_{12} = \hat{J}_{12} M_1. \tag{19}$$

Let X_1 be the null space of M_1. Multiplying both sides of Equation 19 from right by X_1, one obtains the extra restriction

$$J_{12} X_1 = \hat{J}_{12} M_1 X_1 = 0. \tag{20}$$

If the intersection of the invariant or the null subspaces of $J_{22}(\mathbf{y})$ exists and satisfies Equation 20, then M_1 can be determined.

We can treat the general lumping problem in another way by considering the corresponding Green's functions of $J(\mathbf{y})$ and $\hat{J}(\hat{\mathbf{y}})$. For a given initial value of $\mathbf{y}$, $J(\mathbf{y})$ and $\hat{J}(\hat{\mathbf{y}})$ can be represented as $J(t)$ and $\hat{J}(t)$. The corresponding Green's functions $G(t,\tau)$ and $\hat{G}(t,\tau)$ satisfy the following relations:

$$dG(t,\tau)/dt - J(t)G(t,\tau) = 0, \qquad t > \tau. \tag{21a}$$

$$G(\tau,\tau) = I_n. \tag{21b}$$

$$d\hat{G}(t,\tau)/dt - \hat{J}(t)\hat{G}(t,\tau) = 0, \qquad t > \tau. \tag{22a}$$

$$\hat{G}(\tau,\tau) = I_{\hat{n}}. \tag{22b}$$

From Equations 21 and 22 we have

$$\begin{aligned} d(MG(t,\tau) - \hat{G}(t,\tau)M)/dt &= \\ &= MdG(t,\tau)/dt - (d\hat{G}(t,\tau)/dt)M \\ &= MJ(t)G(t,\tau) - \hat{J}(t)\hat{G}(t,\tau)M \\ &= \hat{J}(t)(MG(t,\tau) - \hat{G}(t,\tau)M) \end{aligned}$$

Let

$$K(t,\tau) = MG(t,\tau) - \hat{G}(t,\tau)M. \tag{23}$$

Considering Equations 21 and 22, we have

$$dK(t,\tau)/dt = \hat{J}(t)K(t,\tau), \qquad t > \tau. \tag{24a}$$

$$K(\tau,\tau) = 0. \tag{24b}$$

$$(dK(t,\tau)/dt)_{t=\tau} = \hat{J}(\tau)K(\tau,\tau) = 0.$$

Since $K(\tau,\tau)$ and $(dK(t,\tau)/dt)_{t=\tau}$ are all equal to zero, for $t \geq \tau$ we have

$$K(t,\tau) \equiv 0, \qquad t \geq \tau$$

or,

$$MG(t,\tau) = \hat{G}(t,\tau)M, \qquad t \geq \tau. \tag{25}$$

Equation 25 shows us that the corresponding Green's function has the same property as the Jacobian matrix. Therefore, all the results for the Jacobian matrix also hold for the corresponding Green's function. Since the treatment is the same, we will only consider the Jacobian matrix in the following sections. The Green's function has some advantage in numerical calculations, since we can use it to find the lumping scheme along a reaction path or a given region of initial conditions. This prospect also opens up the possibility of finding approximate lumping schemes valid only in a desired region of the composition space.

B. DETERMINATION OF THE KINETIC EQUATIONS OF THE LUMPED SPECIES

For the exactly lumped reaction system, after determing M we have

$$\hat{\mathbf{f}}(\hat{\mathbf{y}}) = M\mathbf{f}(\mathbf{y}), \tag{6}$$

or

$$\hat{\mathbf{f}}(M\mathbf{y}) = M\mathbf{f}(\mathbf{y}), \tag{26}$$

and this is an identity for any $\mathbf{y}$. Therefore let

$$\mathbf{y} = \bar{M}\hat{\mathbf{y}},$$

and substitute it into equation 26,

$$\hat{\mathbf{f}}(M\bar{M}\hat{\mathbf{y}}) = M\mathbf{f}(\bar{M}\hat{\mathbf{y}}),$$

$$\hat{\mathbf{f}}(\hat{\mathbf{y}}) = M\mathbf{f}(\bar{M}\hat{\mathbf{y}}). \tag{27}$$

Equation 27 does not place any restriction on $\bar{M}$ except that $M\bar{M} = I_{\hat{n}}$. This latter point is important in that the non-unique nature of $\bar{M}$ does not effect the form of the lumped equations (physical model) in the exact case. Thus the behavior of the lumped species can be described by

$$d\hat{\mathbf{y}}/dt = M\mathbf{f}(\bar{M}\hat{\mathbf{y}}), \tag{28}$$

where $\bar{M}$ is anyone of the generalized inverses satisfying $M\bar{M} = I_{\hat{n}}$. The kinetic equations of the lumped species for a given M are unique.

III. THE PROPERTIES OF THE JACOBIAN MATRIX AND ITS INTERSECTION OF THE INVARIANT OR THE NULL SUBSPACES

A. THE RELATIONSHIP BETWEEN THE JACOBIAN MATRIX AND ITS INTERSECTION OF THE INVARIANT OR THE NULL SUBSPACES

In above section it was shown that a system is exactly lumpable if and only if the intersection of the invariant or the null subspaces of $J(\mathbf{y})$ for all values of $\mathbf{y}$ is nonempty. The problem is how to determine the intersection. This task appears difficult, because $\mathbf{y}$ can take infinitely many values. Before presenting the method

to determine the intersections, we will first discuss the relationship between the Jacobian matrix of the kinetic equations and its intersection of the invariant or the null subspaces for all values of $\mathbf{y}$.

This intersection is first determined by the singular property of the Jacobian matrix, due to the conservation of the total mass. Let m_i, y_i be the mass and concentration of species i, and let $\mathbf{m}$ be the vector of m_i. According to the conservation of the total mass, we have

$$\mathbf{m}^T \mathbf{y} = \text{constant}. \tag{29}$$

$$\begin{aligned} d(\mathbf{m}^T \mathbf{y})/dt &= \mathbf{m}^T d\mathbf{y}/dt \\ &= \mathbf{m}^T \mathbf{f}(\mathbf{y}) = 0. \end{aligned} \tag{30}$$

$$\begin{aligned} d(\mathbf{m}^T \mathbf{f}(\mathbf{y}))/d\mathbf{y} &= \mathbf{m}^T d\mathbf{f}(\mathbf{y})/d\mathbf{y} \\ &= \mathbf{m}^T J(\mathbf{y}) = 0. \end{aligned} \tag{31}$$

This shows that at least one row of the Jacobian matrix is a certain linear combination of the others. Therefore the Jacobian matrix is singular for all values of $\mathbf{y}$.

Since $J(\mathbf{y})$ is singular, the image space X_i of the n-dimensional space upon mapping by $J(\mathbf{y}_i)$ is a subspace of it($\mathbf{y}_i$ is any given value of $\mathbf{y}$ and i can take infinitely many values).

$$J(\mathbf{y}_i) I_n = X_i, \tag{32}$$

where X_i is composed of the linearly independent columns of $J(\mathbf{y}_i)$ and has dimension less than n. It is easy to demonstrate that the union of all X_i also has dimension less than n.

Without loss of generality, let row n of $J(\mathbf{y})$ be a linear combination of the other rows. The linear combination is the same for any value of $\mathbf{y}$. Therefore, any

column vector of $J(\mathbf{y})$ for any value of $\mathbf{y}$ is located in the same $(n-1)$-dimensional subspace. Since X_i is composed of the linearly independent columns of $J(\mathbf{y}_i)$, X_i is in the $(n-1)$-dimensional subspace. Since all X_i are located in the same subspace, then the union of all X_i

$$X = X_1 \bigcup X_2 \bigcup \ldots \bigcup X_i \tag{33}$$

is also in the same $(n-1)$-dimensional subspace.

Similarly, if k rows of $J(\mathbf{y})$ are linear combinations of the others for any value of $\mathbf{y}$, the union X would be located in an $(n-k)$-dimensional subspace. This is true for any reaction system if we consider the mass conservation for each atom.

Now we can prove that the union X is an intersection of the invariant subspace of $J(\mathbf{y})$. We have

$$J(\mathbf{y}_i)X \in J(\mathbf{y}_i)I_n \in X_i \in X. \tag{34}$$

which is valid for any given $\mathbf{y}_i$ and X has dimension less than n. Therefore X is the intersection of the invariant subspaces of $J(\mathbf{y})$.

Although mass conservation is a "trivial" conservation property leading to lumping, the subspace X forms the starting point to determine other intersections of the invariant subspaces of $J(\mathbf{y})$. First we can demonstrate that any subspace in the n-dimensional space containing X is an intersection of the invariant subspaces of $J(\mathbf{y})$. Then we can prove that if any other intersection of the invariant subspaces of $J(\mathbf{y})$, say Z, with equal or lower dimension than X exists, the intersection of Z and X is nonempty and is a new intersection of the invariant subspaces of $J(\mathbf{y})$. These statements are proved below.

Let Y be a subspace containing X. For any $\mathbf{y}_i$, we have

$$J(\mathbf{y}_i)Y \in J(\mathbf{y}_i)I_n \in X_i \in X \in Y. \tag{35}$$

Then Y is an intersection of the invariant subspaces of $J(\mathbf{y})$.

Let Z be a subspace with dimension equal to or less than that of X and the intersection between Z and X is empty. It is easy to prove that Z can not be an intersection of the invariant subspaces of $J(\mathbf{y})$. Since

$$J(\mathbf{y}_i)Z \in J(\mathbf{y}_i)I_n \in X_i \in X, \tag{36}$$

and the intersection of Z and X is empty, the image of Z upon mapping by $J(\mathbf{y}_i)$ is out of Z. Therefore, Z is not an invariant subspace of $J(\mathbf{y}_i)$. Consequently it is not the intersection of the invariant subspaces of $J(\mathbf{y})$. If the intersection of Z and X, say W, exists and Z is an intersection of the invariant subspaces of $J(\mathbf{y})$, according to Equation 36 the image of Z must be in W. After mapping by $J(\mathbf{y}_i)$, the image of any vector in W is still in it. Therefore W is an intersection of the invariant subspaces of $J(\mathbf{y})$. This implies that X and its subspaces, which are invariant to $J(\mathbf{y})$, have a central role in constructing all intersections of the invariant subspaces of $J(\mathbf{y})$.

Suppose Z is a subspace of X, then Z is an intersection of the invariant subspaces of $J(\mathbf{y})$ if and only if the image of Z upon mapping by $J(\mathbf{y}_i)$ is the intersection of Z and X_i. To prove this point suppose that Z is an intersection of the invariant subspaces of $J(\mathbf{y})$, then

$$J(\mathbf{y}_i)Z \in Z.$$

However, we also have

$$J(\mathbf{y}_i)Z \in X_i.$$

This means that the image of Z upon mapping by $J(\mathbf{y}_i)$ is the intersection of Z and X_i. This condition is also sufficient. Suppose the image of Z upon mapping by $J(\mathbf{y}_i)$ is the intersection of Z and X_i. Therefore, the image of Z upon mapping by $J(\mathbf{y})$ for any value of $\mathbf{y}$ is in Z. Then Z is the intersection of the invariant subspaces of $J(\mathbf{y})$.

Let the intersection of Z and X_i be Z_i. Then we can represent Z as follows:

$$Z = Z_1 \bigcup Z_2 \bigcup \ldots \bigcup Z_i. \tag{37}$$

Equation 37 would be useful for determining the intersections of the invariant subspaces of $J(\mathbf{y})$ with dimension less than that of X.

We can consider this problem in other way. Suppose the intersection X has been determined, and a subspace Z of it is invariant to $J(\mathbf{y})$. Z can be described as

$$Z = XS, \tag{38}$$

where S is an $(n - \hat{n}) \times m(m < (n - \hat{n}))$ constant matrix. Notice that

$$J(\mathbf{y})X = XP(\mathbf{y}), \tag{39}$$

where $P(\mathbf{y})$ is an $(n - \hat{n}) \times (n - \hat{n})$ matrix, and

$$J(\mathbf{y})Z = J(\mathbf{y})XS = XSQ(\mathbf{y}), \tag{40}$$

where $Q(\mathbf{y})$ is an $m \times m$ matrix. Multiplying both sides of Equation 39 from the right by S, we have

$$J(\mathbf{y})XS = XP(\mathbf{y})S. \tag{41}$$

Comparing Equation 40 and 41, we obtain

$$XSQ(\mathbf{y}) = XP(\mathbf{y})S. \tag{42}$$

Considering that X is column-full rank matrix, one can always find a generalized inverse $\bar{X}$ satisfying

$$\bar{X}X = I_{n-\hat{n}}. \tag{43}$$

Multiplying both sides of Equation 42 from left by $\bar{X}$ gives

$$\bar{X}XSQ(\mathbf{y}) = \bar{X}XP(\mathbf{y})S,$$

$$SQ(\mathbf{y}) = P(\mathbf{y})S. \tag{44}$$

Similarly, we can determine a generalized inverse $\bar{S}$ satisfying

$$\bar{S}S = I_m. \tag{45}$$

Multiplying both sides of Equation 44 from left by $\bar{S}$, we have

$$\bar{S}SQ(\mathbf{y}) = \bar{S}P(\mathbf{y})S,$$

$$Q(\mathbf{y}) = \bar{S}P(\mathbf{y})S. \tag{46}$$

Substituting $Q(\mathbf{y})$ into Equation 44, one can obtain

$$S\bar{S}P(\mathbf{y})S = P(\mathbf{y})S,$$

$$(I_{n-\hat{n}} - S\bar{S})P(\mathbf{y})S = 0.$$

Transposing this equation gives

$$S^T P^T(\mathbf{y})(I_{n-\hat{n}} - \bar{S}^T S^T) = 0. \tag{47}$$

Equation 47 is exactly the same as Equation 9. This implies that X has subspaces invariant to $J(\mathbf{y})$ if and only if the intersection of the invariant or the null subspaces of $P^T(\mathbf{y})$ is nonempty. Therefore we can employ the same method for determining X to determine Z. In this way we can find out all intersections of the invariant subspaces of $J(\mathbf{y})$ with different dimensions.

The intersection of the null subspaces of $J(\mathbf{y})$ is the solution of the following linear algebric equation

$$J(\mathbf{y})\mathbf{x} = \mathbf{0}. \tag{48}$$

Moreover, the solution is independent of the value of $\mathbf{y}$. If we consider $\mathbf{y}$ symbolically, this shows that there is a nontrivial solution of Equation 48 if and only if

some columns of $J(\mathbf{y})$ are linear combinations of the other columns. All the linearly independent solutions of Equation 48 compose the largest intersection of the null subspaces of $J(\mathbf{y})$. Any subspace of this intersection is also an intersection of the null subspaces of $J(\mathbf{y})$.

B. DETERMINATION OF THE INTERSECTION OF THE INVARIANT OR THE NULL SUBSPACES OF $J(\mathbf{y})$ FOR ALL VALUES OF $\mathbf{y}$

The Jacobian matrix can be considered as an n^2 vector. Therefore, for any value of $\mathbf{y}$, $J(\mathbf{y})$ can be represented as a linear combination of $m(m \leq n^2)$ constant matrices:

$$J(\mathbf{y}) = \sum_{k=1}^{m} a_k(\mathbf{y}) A_k, \tag{49}$$

where $a_k(\mathbf{y})$ are parameters, which are the functions of $\mathbf{y}$; A_k are constant matrices, which are considered as a basis of $J(\mathbf{y})$. The problem is how to determine the basis A_k. There are several ways to do it, and one is as follows. The Jacobian matrix $J(\mathbf{y})$ can be represented as

$$J(\mathbf{y}) = \sum_{i,j=1}^{m} j_{ij}(\mathbf{y}) E_{ij}, \tag{50}$$

where $j_{ij}(\mathbf{y})$ is the (i,j)-entry of $J(\mathbf{y})$; E_{ij} is the elementary matrix, which is defined as the $(n \times n)$-matrix having unity in the (i,j)th position and all other elements are zero.

If j_{pq} is equal to $cj_{ij}(\mathbf{y})$, where c is a constant, we can combine these two terms as

$$a_k(\mathbf{y}) = j_{ij}(\mathbf{y}),$$

$$A_k = E_{ij} + cE_{pq}.$$

In this way one can combine the terms in Equation 50 as much as possible to obtain the following simplified formula

$$J(\mathbf{y}) = \sum_{k=1}^{m} a_k(\mathbf{y}) A_k, \tag{51}$$

where m is less than n^2.

It is easy to demonstrate that the intersection of the invariant or the null subspaces of all constant matrices A_k is that of $J(\mathbf{y})$. Let the $n \times (n - \hat{n})$ constant matrix X represent the intersection of the invariant subspaces for all A_k. Multiply both sides of Equation 51 from the right by X to obtain

$$\begin{aligned} J(\mathbf{y})X &= \sum_{k=1}^{m} a_k(\mathbf{y}) A_k X, \\ &= \sum_{k=1}^{m} a_k(\mathbf{y}) X P_k, \\ &= X \sum_{k=1}^{m} a_k(\mathbf{y}) P_k. \end{aligned} \tag{52}$$

where P_k are $(n - \hat{n}) \times (n - \hat{n})$ constant matrices. Equation 52 shows that X is the intersection of the invariant subspaces of $J(\mathbf{y})$.

Similarly, we can prove that the intersection X of the null subspaces of all A_k is also that of $J(\mathbf{y})$.

$$\begin{aligned} J(\mathbf{y})X &= \sum_{k=1}^{m} a_k(\mathbf{y}) A_k X, \\ &= \sum_{k=1}^{m} a_k(\mathbf{y}) 0 = 0. \end{aligned} \tag{53}$$

If Equation 49 satisfies the restriction that in each case we can choose an appropriate value of $\mathbf{y}$ such that all $a_k(\mathbf{y}_i)$ vanish except $a_i(\mathbf{y}_i)$, i.e.,

$$J(\mathbf{y}_i) = a_i(\mathbf{y}_i) A_i. \qquad (i = 1, 2, \ldots, m) \tag{54}$$

Then the intersection X of the invariant or the null subspaces of $J(\mathbf{y})$ is also that of all A_k. Multiplying both sides of Equation 54 by X from the right, when X is the intersection of the invariant subspaces of $J(\mathbf{y})$, we obtain

$$J(\mathbf{y}_i)X = a_i(\mathbf{y}_i)A_iX,$$

$$XP(\mathbf{y}_i) = a_i(\mathbf{y}_i)A_iX.$$

Since $a_i(\mathbf{y}_i)$ is not equal to zero, then

$$A_iX = XP(\mathbf{y}_i)/a_i(\mathbf{y}_i). \tag{55}$$

If X is the intersection of the null subspaces of $J(\mathbf{y})$, similarly we have

$$J(\mathbf{y}_i)X = a_i(\mathbf{y}_i)A_iX,$$

$$0 = a_i(\mathbf{y}_i)A_iX,$$

$$0 = A_iX. \tag{56}$$

Equations 55 and 56 show that X is the invariant or the null subspace of A_i. Since this is valid for all A_k, then X is the intersection of the invariant or the null subspaces for all A_k. Thus we can determine the intersections of $J(\mathbf{y})$ by only determining the intersections of all A_k.

When the reaction system is a uni- and/or bimolecular reaction system described by linear or quadratic first order ordinary differential equations, the elements of $J(\mathbf{y})$ are only linear functions of y_ks. In this case, Equation 51 will have a simple form, i.e., $a_k(\mathbf{y})$ is constant or y_k,

$$J(\mathbf{y}) = A_0 + \sum_{k=1}^{m} y_k A_k. \tag{57}$$

where m is equal to or less than n, and A_0 can be a null matrix. As Equation 57 satisfies the above restriction, we can determine all intersections of $J(\mathbf{y})$ by determining those of A_0 and all A_k.

Notice that the invariant subspace of a constant matrix contains at least one eigenvector of it. This property presents a method to establish the intersection of the invariant subspaces of A_0 and all A_k. First, the eigenvectors of A_0 and all A_k are determined. Then consider all possible combinations of these eigenvectors. Each combination contains at least one eigenvector of every constant matrix. The linearly dependent eigenvectors are cancelled in each combination. The resultant combinations are examined for A_0 and A_k whether they are invariant to all the matrices. If a combination is invariant to A_0 and all A_k, it is an intersection of the invariant subspaces of these matrices. We can also determine all intersections of the invariant subspaces of $J(\mathbf{y})$ by first determining X and then its subspaces Z invariant to $J(\mathbf{y})$.

The determination of the intersection of the null subspaces is easier. We need to find the common solutions for the following equations:

$$A_0\mathbf{x} = \mathbf{0}, \tag{58}$$

$$A_k\mathbf{x} = \mathbf{0}. \qquad (k = 1, 2, ..., m) \tag{59}$$

In order to obtain the common solutions, we put A_0 and all A_k together and omit the linearly dependent rows. Then we obtain a constant matrix A and solve the set of linear algebric equations

$$A\mathbf{x} = \mathbf{0}. \tag{60}$$

All the linearly independent solutions of $\mathbf{x}$ compose the largest intersection of the null subspaces of A_0 and all A_k. Most importantly any subspace of the largest

intersection is still an intersection of the null subspaces of these matrices. The procedure of determining the intersections will be illustrated by some simple examples below.

IV. APPLICATION TO UNI- AND/OR BIMOLECULAR REACTION SYSTEMS

As examples of the application of this analysis, we choose uni- and/or bimolecular reaction systems. As pointed out above, in this case the Jacobian matrix can be described as

$$J(\mathbf{y}) = A_0 + \sum_{k=1}^{m} y_k A_k. \tag{61}$$

For a unimolecular reaction system, the kinetic equations are

$$d\mathbf{y}/dt = K\mathbf{y}, \tag{62}$$

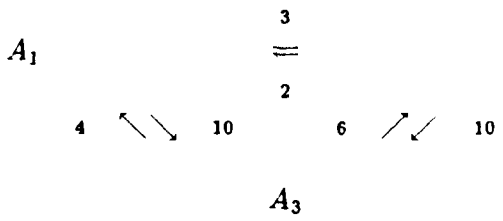
where K is the rate constant matrix. The Jacobian matrix for the unimolecular reaction system is just K,

$$J(\mathbf{y}) = K. \tag{63}$$

In this case, there does not exist the problem of intersections, because $J(\mathbf{y})$ is a constant matrix. The subspace spanned by a subset of eigenvectors is the invariant subspace of a constant matrix. If the matrix has full eigenvectors, the subspace spanned by the eigenvectors with zero eigenvalue is the null space of the matrix. Therefore any unimolecular reaction system is exactly lumpable and the lumping schemes are easy to obtain after determining the eigenvectors of the rate constant matrix. This behavior can be shown by a simple example.

Example I

A unimolecular reaction system with 3 species is described as follows:[2]



where A_1, A_2 and A_3 represent the three species; all numbers are unitless rate constants. Let y_i represent the concentration of species A_i. Then the corresponding kinetic equations can be described as

$$d\mathbf{y}/dt = K\mathbf{y}, \tag{64}$$

where $\mathbf{y}$ is the concentration vector; K is the rate constant matrix.

$$K = \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix} \tag{65}$$

$$J(\mathbf{y}) = d(K\mathbf{y})/d\mathbf{y} = K. \tag{66}$$

The eigenvector matrix X and the eigenvalue matrix Λ of $J(\mathbf{y})$ are

$$X = \begin{pmatrix} 0.2 & 0.2 & 1 \\ 0.3 & 0.3 & -1 \\ 0.5 & -0.5 & 0 \end{pmatrix} \tag{67}$$

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -20 & 0 \\ 0 & 0 & -15 \end{pmatrix} \tag{68}$$

Since any subspace spanned by a subset of eigenvectors is an invariant subspace of $J(\mathbf{y})$, we choose the 1-dimensional subspace spanned by the last eigenvector. Using Equation 14 we have

$$(1 \quad -1 \quad 0) \begin{pmatrix} m_{i1} \\ m_{i2} \\ m_{i3} \end{pmatrix} = 0. \qquad (i = 1, 2) \tag{69}$$

Solving this equation we obtain the solution for the lumping matrix M:

$$\mathbf{m}_1 = (\, c \quad c \quad 0\,)^T$$

$$\mathbf{m}_2 = (\, 0 \quad 0 \quad d\,)^T$$

$$M = \begin{pmatrix} c & c & 0 \\ 0 & 0 & d \end{pmatrix}, \tag{70}$$

where c, d are arbitrary constants. We want M to have a full row rank, thus requiring $c, d \neq 0$. A special case is $c = d = 1$, i.e.,

$$M = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{71}$$

For this M we can find an infinite number of $\bar{M}$ satisfying $M\bar{M} = I_2$. We arbitrarily choose two:

$$\bar{M}_1 = \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \bar{M}_2 = \begin{pmatrix} 0.4 & 0 \\ 0.6 & 0 \\ 0 & 1 \end{pmatrix}.$$

It is easy to show that the kinetic equations for the lumped system are the same in spite of using different $\bar{M}$. According to Equation 27

$$\hat{f}(\hat{\mathbf{y}}) = M\mathbf{f}(\bar{M}\hat{\mathbf{y}}),$$

and since

$$\mathbf{f}(\mathbf{y}) = K\mathbf{y},$$

then

$$\hat{\mathbf{f}}(\hat{\mathbf{y}}) = MK\bar{M}\hat{\mathbf{y}}. \tag{72}$$

For $\bar{M}_1$ we have

$$\begin{aligned} \hat{\mathbf{f}}(\hat{\mathbf{y}}) &= \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix} \begin{pmatrix} 0.5 & 0 \\ 0.5 & 0 \\ 0 & 1 \end{pmatrix} \hat{\mathbf{y}} \\ &= \begin{pmatrix} -10 & 10 \\ 10 & -10 \end{pmatrix} \hat{\mathbf{y}} \end{aligned}$$

Similarly for $\bar{M}_2$ we have

$$\hat{\mathbf{f}}(\hat{\mathbf{y}}) = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} -13 & 2 & 4 \\ 3 & -12 & 6 \\ 10 & 10 & -10 \end{pmatrix} \begin{pmatrix} 0.4 & 0 \\ 0.6 & 0 \\ 0 & 1 \end{pmatrix} \hat{\mathbf{y}}$$
$$= \begin{pmatrix} -10 & 10 \\ 10 & -10 \end{pmatrix} \hat{\mathbf{y}}$$

The reaction scheme of the lumped system can be described as

$$\hat{A}_1 \underset{10}{\overset{10}{\longrightarrow}} \hat{A}_2.$$

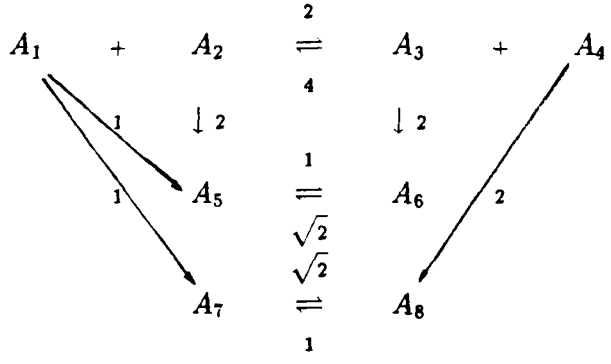
$$d\hat{\mathbf{y}}/dt = \hat{K}\hat{\mathbf{y}},$$

where $\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2)^T$ and

$$\hat{K} = \begin{pmatrix} -10 & 10 \\ 10 & -10 \end{pmatrix}.$$

EXAMPLE II

A uni- and bimolecular reaction system with 8 species is illustrated as follows:[13]



where the A_is are species; the numbers are unitless rate constants.

Leting y_i represent the concentration of A_i, it is easy to write out the kinetic equations and the corresponding Jacobian matrix $J(\mathbf{y})$.

$$\begin{aligned}
dy_1/dt &= -2y_1 - 2y_1y_2 + 4y_3y_4 \\
dy_2/dt &= -2y_2 - 2y_1y_2 + 4y_3y_4 \\
dy_3/dt &= -2y_3 - 4y_3y_4 + 2y_1y_2 \\
dy_4/dt &= -2y_4 - 4y_3y_4 + 2y_1y_2 \\
dy_5/dt &= -y_5 + y_1 + 2y_2 + \sqrt{2}y_6 \\
dy_6/dt &= -\sqrt{2}y_6 + 2y_3 + y_5 \\
dy_7/dt &= -\sqrt{2}y_7 + y_1 + y_8 \\
dy_8/dt &= -y_8 + 2y_4 + \sqrt{2}y_7
\end{aligned} \tag{74}$$

$$J(\mathbf{y}) = \begin{pmatrix}
-2(1+y_2) & -2y_1 & 4y_4 & 4y_3 & & & & \\
-2y_2 & -2(1+y_1) & 4y_4 & 4y_3 & & & & \\
2y_2 & 2y_1 & -2(1+2y_4) & -4y_3 & & 0 & & \\
2y_2 & 2y_1 & -4y_4 & -2(1+2y_3) & & & & \\
1 & 2 & 0 & 0 & -1 & \sqrt{2} & 0 & 0 \\
0 & 0 & 2 & 0 & 1 & -\sqrt{2} & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & -\sqrt{2} & 1 \\
0 & 0 & 0 & 2 & 0 & 0 & \sqrt{2} & -1
\end{pmatrix}$$

This matrix can be represented as

$$J(\mathbf{y}) = A_0 + \sum_{k=1}^{4} y_k A_k,$$

where

$$A_0 = \begin{pmatrix}
-2 & 0 & 0 & 0 & & & & \\
0 & -2 & 0 & 0 & & & & \\
0 & 0 & -2 & 0 & & 0 & & \\
0 & 0 & 0 & -2 & & & & \\
1 & 2 & 0 & 0 & -1 & \sqrt{2} & 0 & 0 \\
0 & 0 & 2 & 0 & 1 & -\sqrt{2} & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & -\sqrt{2} & 1 \\
0 & 0 & 0 & 2 & 0 & 0 & \sqrt{2} & -1
\end{pmatrix}$$

$$A_1 = \begin{pmatrix} 0 & -2 & 0 & 0 & & \\ 0 & -2 & 0 & 0 & & \\ 0 & 2 & 0 & 0 & & 0 \\ 0 & 2 & 0 & 0 & & \\ & & & & & \\ & 0 & & & & 0 \end{pmatrix} \qquad A_2 = \begin{pmatrix} -2 & 0 & 0 & 0 & & \\ -2 & 0 & 0 & 0 & & \\ 2 & 0 & 0 & 0 & & 0 \\ 2 & 0 & 0 & 0 & & \\ & & & & & \\ & 0 & & & & 0 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} 0 & 0 & 0 & 4 & & \\ 0 & 0 & 0 & 4 & & \\ 0 & 0 & 0 & -4 & & 0 \\ 0 & 0 & 0 & -4 & & \\ & & & & & \\ & 0 & & & & 0 \end{pmatrix} \qquad A_4 = \begin{pmatrix} 0 & 0 & 4 & 0 & & \\ 0 & 0 & 4 & 0 & & \\ 0 & 0 & -4 & 0 & & 0 \\ 0 & 0 & -4 & 0 & & \\ & & & & & \\ & 0 & & & & 0 \end{pmatrix}$$

The corresponding eigenvector matrices X_{A_0} and X_{A_k} with their eigenvalues are as follows:

$$\lambda_1 = \quad -2, \quad -2, \quad -2, \quad -2, \quad -(1+\sqrt{2}), \quad -(1+\sqrt{2}), \quad 0, \quad 0$$

$$X_{A_0} = \begin{pmatrix} 0 & 2-\sqrt{2} & 1 & 1-\sqrt{2} & & & & \\ (\sqrt{2}-1)/2 & (\sqrt{2}-3)/2 & -1 & \sqrt{2}-1 & & & & \\ (1-\sqrt{2})/2 & -1/2 & -1/2 & (1-\sqrt{2})/2 & & 0 & & \\ 0 & 1/\sqrt{2} & 1/2 & (\sqrt{2}-1)/2 & & & & \\ 1 & 1 & 1 & 1 & 1 & 0 & \sqrt{2} & 0 \\ -1 & 0 & 0 & -1 & -1 & 0 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 & 0 & 1 \\ 0 & 0 & -1 & 1 & 0 & -1 & 0 & \sqrt{2} \end{pmatrix}$$

$$\lambda_1 = \begin{matrix} -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$X_{A_1} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}$$

$$\lambda_2 = \begin{matrix} -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$X_{A_2} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}$$

$$\lambda_1 = \begin{matrix} -4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$X_{A_3} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}$$

$$\lambda_1 = \begin{matrix} -4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{matrix}$$

$$X_{A_4} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \end{pmatrix}$$

Using the methods presented in section III B, one can obtain all possible intersections with different dimensions of the invariant and the null subspaces of $J(\mathbf{y})$. First we use the singular property of the Jacobian matrix to determine the intersection of the invariant subspaces of $J(\mathbf{y})$. In this example, we have

$$\mathbf{1}^T J(\mathbf{y}) = \mathbf{0}^T.$$

where $\mathbf{1}^T = (1\ 1\ \ldots\ 1)$. Therefore, any column of $J(\mathbf{y})$ for any value of $\mathbf{y}$ is located in the same $(n-1)$-dimensional subspace. This subspace can be constructed by Equation 33.

$$X = X_0 \bigcup X_1 \bigcup \ldots \bigcup X_4,$$

where X_0 and X_k are images of n-dimensional space upon mapping by A_0 and A_k, which are composed of the linearly independent columns of A_0 and A_k. Then we have

$$X = \begin{pmatrix} -2 & -2 & 4 & 4 & -2 & 0 & 0 & 0 & 0 & 0 \\ -2 & -2 & 4 & 4 & 0 & -2 & 0 & 0 & 0 & 0 \\ 2 & 2 & -4 & -4 & 0 & 0 & -2 & 0 & 0 & 0 \\ 2 & 2 & -4 & -4 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 0 & -1 \end{pmatrix}$$

After omitting the linearly dependent columns the intersection X is obtained.

$$X = \begin{pmatrix} -2 & -2 & 0 & 0 & 0 & 0 & 0 \\ -2 & 0 & -2 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & -2 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 1 & 2 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 2 & 0 & -1 \end{pmatrix}$$

If we change the bases of this subspace, it can be described by a simpler form:

$$\tilde{X}_1 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{pmatrix}$$

It is easy to prove that any column of X is a linear combination of the columns of $\tilde{X}_1$. Then the two matrices are equivalent to represent the subspace.

Now we use the eigenvectors of A_0 and A_k to construct the intersection of the invariant subspaces of $J(\mathbf{y})$. After examining the eigenvector matrices we find that the first 6 columns of X_{A_0} and X_{A_2} to X_{A_4} are linear combinations of the same columns of X_{A_1}. Therefore the subspace $\tilde{X}_2$ spanned by the first 6 columns of X_{A_1} is an intersection of the invariant subspaces of $J(\mathbf{y})$.

$$\tilde{X}_2 = \begin{pmatrix} 1 & 1 & 1 & & & \\ 1 & 0 & 0 & & 0 & \\ -1 & -1 & 0 & & & \\ -1 & 0 & -1 & & & \\ & & & 1 & 1 & 1 \\ & & & -1 & 0 & 0 \\ & 0 & & 0 & -1 & 0 \\ & & & 0 & 0 & -1 \end{pmatrix}$$

After examination we can find that the subspaces $\tilde{X}_3$ of $\tilde{X}_2$

$$\tilde{X}_3 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \\ -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

is invariant to A_0 and A_k. Therefore it is also the intersection of the invariant subspaces of $J(\mathbf{y})$. Similarly the two subspaces $\tilde{X}_4$, $\tilde{X}_5$ of $\tilde{X}$ and the union $\tilde{X}_6$ of $\tilde{X}_4$ and $\tilde{X}_5$

$$\tilde{X}_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \qquad \tilde{X}_5 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ -1 \end{pmatrix} \qquad \tilde{X}_6 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ -1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}$$

are intersections of the invariant subspaces of $J(\mathbf{y})$.

The intersection of the null subspaces of $J(\mathbf{y})$ can be determined by Equation 48 and the solution is $\tilde{X}_7$.

$$\tilde{X}_7 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ \sqrt{2} & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & \sqrt{2} \end{pmatrix}$$

For these intersections the corresponding lumping matrices M_i are determined by Equation 13. All M_i are arranged below by dimension in increasing order except M_7, which is different from the others and will be discussed later.

$$M_1 = (1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1)$$

$$M_2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$M_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ 0 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$M_6 = \begin{pmatrix} 1 & 0 & 0 & 0 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ 0 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} 1 & 0 & 0 & 0 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ 0 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$M_5 = \begin{pmatrix} 1 & 0 & 0 & 0 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ 0 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}$$

$$M_7 = \begin{pmatrix} 1 & 0 & 0 & 0 & & & & \\ 0 & 1 & 0 & 0 & & & & \\ 0 & 0 & 1 & 0 & & 0 & & \\ 0 & 0 & 0 & 1 & & & & \\ 0 & 0 & 0 & 0 & -1 & \sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -\sqrt{2} & 1 \end{pmatrix}$$

We can determine the kinetic equations of the exactly lumped systems using Equation 28

$$d\hat{\mathbf{y}}/dt = M\mathbf{f}(\bar{M}\hat{\mathbf{y}}). \tag{28}$$

For M_1 we arbitrarily choose

$$\bar{M}_1 = (1 \quad 0 \quad \ldots \quad 0)^T.$$

Since $\mathbf{y} = \bar{M}\hat{\mathbf{y}}$, we have

$$y_1 = \hat{y}.$$

$$y_i = 0. \qquad (i = 2, 3, \ldots, 8)$$

Substituting these relations into Equation 28 and 74, the lumped kinetic equation is

$$d\hat{y}/dt = 0. \qquad (\hat{y} = \sum_{i=1}^{8} y_i) \tag{75}$$

For M_2 we arbitrarily choose

$$\bar{M}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}^T.$$

Similarly we have

$$y_1 = \hat{y}_1, \qquad y_5 = \hat{y}_2,$$

$$y_i = 0. \qquad (i = 2, 3, 4, 6, 7, 8)$$

Substituting them into Equation 28 and 74, the lumped kinetic equations are

$$\begin{aligned} d\hat{y}_1/dt &= -2\hat{y}_1, \\ d\hat{y}_2/dt &= 2\hat{y}_1. \end{aligned} \tag{76}$$

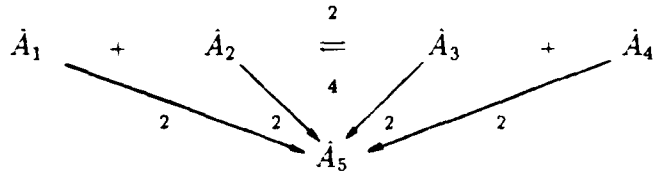
This kinetic equations can be described by a unimolecular reaction scheme:

$$\hat{A}_1 \xrightarrow{2} \hat{A}_2$$

$$\hat{A}_1 = \sum_{i=1}^{4} A_i, \qquad \hat{A}_2 = \sum_{i=5}^{8} A_i.$$

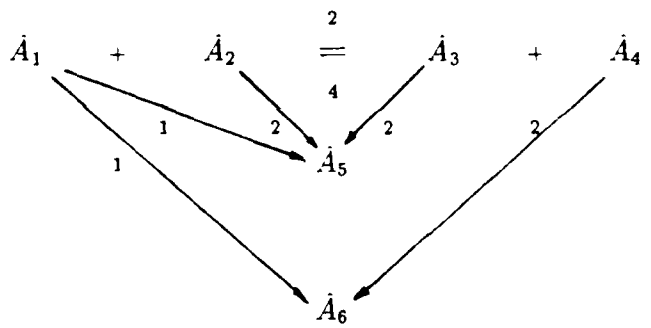
The first six lumped systems for M_1 to M_6 follow uni- and/or bimolecular reaction schemes. The corresponding reaction schemes of the lumped systems for M_3 to M_6 are illustrated as follows:

lumping scheme M_3



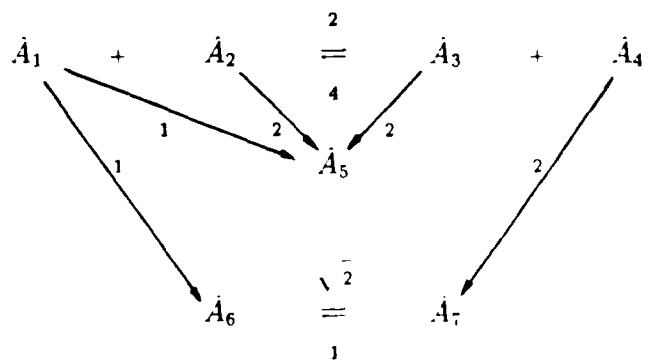
$$\hat{A}_i = A_i \quad (i = 1,2,3,4), \qquad \hat{A}_5 = \sum_{i=5}^{8} A_i.$$

lumping scheme M_6



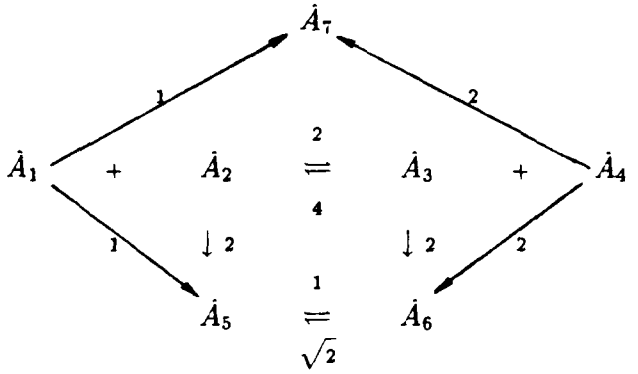
$$\hat{A}_i = A_i \quad (i = 1,2,3,4), \qquad \hat{A}_5 = A_5 + A_6, \qquad \hat{A}_6 = A_7 + A_8.$$

lumping scheme M_4



$$\hat{A}_i = A_i \quad (i = 1,2,3,4), \qquad \hat{A}_i = A_{i+1} \quad (i = 6,7), \qquad \hat{A}_5 = A_5 + A_6 .$$

lumping scheme M_5



$$\hat{A}_i = A_i \quad (i = 1,2,3,4,5,6), \qquad \hat{A}_7 = A_7 + A_8 .$$

The last lumping scheme M_7 can only be described by a simplified set of differential equations, since it does not follow a uni- and/or bimolecular reaction scheme. The condition under which a lumped system follows a physical uni- and/or bimolecular reaction scheme has been discussed in References [2] and [6]. The differential equations for the lumped system of M_7 is given as

$$\begin{aligned}
d\hat{y}_1/dt &= -2\hat{y}_1 - 2\hat{y}_1\hat{y}_2 + 4\hat{y}_3\hat{y}_4 \\
d\hat{y}_2/dt &= -2\hat{y}_2 - 2\hat{y}_1\hat{y}_2 + 4\hat{y}_3\hat{y}_4 \\
d\hat{y}_3/dt &= -2\hat{y}_3 - 4\hat{y}_3\hat{y}_4 + 2\hat{y}_1\hat{y}_2 \\
d\hat{y}_4/dt &= -2\hat{y}_4 - 4\hat{y}_3\hat{y}_4 + 2\hat{y}_1\hat{y}_2 \\
d\hat{y}_5/dt &= -\hat{y}_1 + 2\hat{y}_2 + 2\sqrt{2}\hat{y}_3 - (1+\sqrt{2})\hat{y}_5 \\
d\hat{y}_6/dt &= -\sqrt{2}\hat{y}_1 + 2\hat{y}_4 - (1+\sqrt{2})\hat{y}_6
\end{aligned} \tag{77}$$

where

$$\hat{y}_i = y_i, \qquad (i = 1,2,3,4)$$

$$\hat{y}_5 = -y_5 + \sqrt{2}y_6 .$$

$$\hat{y}_6 = -\sqrt{2}y_7 + y_8 .$$

This lack of a corresponding chemical mechanism for the lumped system may often arise in lumping, and there is no practical difficulty in this situation. Since the lumping scheme above is exact, the simplified model will give exactly the same values of the lumped species as those given by the original one. Notice that $\dot{y}_i = y_i$ for $i = 1, 2, 3, 4$. If one only considers y_1 to y_4, Equation 77 will give the same results as those given by Equation 74. However, Equation 77 is simpler, even though it does not follow uni- and/or bimolecular reaction schemes.

These examples are very simple, however, they illustrate the method which can be applied to other more complicated systems.

V. CONCLUSION AND DISCUSSION

In this paper a genaral analysis of exact lumping has been given, which can be used for any system described by a set of first order ordinary differential equations with any degree of nonlinearity. Uni- and/or bimolecular reaction systems are only special cases of this general analysis.

The kinetic properties and the coupling pattern of the reactions in the exactly lumpable system must satisfy some restrictions, which are reflected by the properties of their Jacobian matrix and the corresponding Green's function. The intersections of the invariant and the null subspaces of the Jacobian matrix represent possible lumpabilities of a given complex reaction system. A systematic method to determine the intersections of the invariant and the null subspaces of the Jacobian matrix and the corresponding lumping schemes was developed. Using the generalized inverse of the lumping matrix, the differntial equations of the lumped system can be readily obtained, and the non-unique nature of the generalized inverses does not effect the form of the lumped equations in the exact case.

Although some useful results about exact lumping have been obtained, there is still further work to do. Systematic application of this analysis to complex reaction systems needs to be considered. However, in the treatment of actual reaction systems, the first problem encountered will likely be their non-exact lumpability. Very few systems satisfy the restrictions for the existance of exact lumping. Sometimes, even if a system is exactly lumpable, the results may not meet practically desired goals. For example, in the $CO/H_2O/O_2$ combustion system we would like the easily measurable concentrations of CO, CO_2, O_2, H_2O to be unlumped[13]. With this constraint, the system likely can not be exactly lumped, and we have to lump the other species of the system approximately. Developing a general approach for approximate lumping is very important for realistic problems. The exact lumping analysis presented above should form a rigorous starting point for the development of approximate lumping.

NOTATION

Scalars

$a_k(\mathbf{y})$ = kth coefficient of a linear combination of matrices

A_i = ith species of a reaction system

c = constant

d = constant

$j_{i,j}(\mathbf{y})$ = (i,j)-entry of matrix $J(\mathbf{y})$

m = dimension of matrix $Q(\mathbf{y})$

m_i = mass of species A_i

$m_{i,j}$ = (i,j)-entry of matrix M

n = dimension of vector $\mathbf{y}$

$\hat{n}$ = dimension of vector $\hat{\mathbf{y}}$

p = dimension of identity matrix

t = time

y_k = kth element of vector $\mathbf{y}$

Vectors and Matrices

Capital letters represent matrices; bold-face lower case letters represent vectors.

A_0 = constant matrix

A_k = constant matrix

$\mathbf{e}_i$ = unit vector with 1 as its ith element, and 0 for the rest of the elements

E_{ij} = elementary matrix with 1 as its (i,j)-entry, and 0 for the rest of the elements

$\mathbf{f}(\mathbf{y})$ = n-dimensional function vector

$\hat{\mathbf{f}}(\hat{\mathbf{y}})$ = $\hat{n}$-dimensional function vector

$G(t,\tau)$ = Green's function of $J(\mathbf{y})$

$\hat{G}(t,\tau)$ = Green's function of $\hat{J}(\hat{\mathbf{y}})$

I = identity matrix

$J(\mathbf{y})$ = Jacobian matrix of $\mathbf{f}(\mathbf{y})$

J_{ij} = submatrix of $J(\mathbf{y})$

$\hat{J}(\hat{\mathbf{y}})$ = Jacobian matrix of $\hat{\mathbf{f}}(\hat{\mathbf{y}})$

$\hat{J}_{ij}$ = submatrix of $\hat{J}(\hat{\mathbf{y}})$

K = rate constant matrix

$K(t,\tau)$ = defined as $MG(t,\tau) - \hat{G}(t,\tau)M$

M = lumping matrix

M_1 = submatrix of M

$\mathbf{m}$ = n-dimensional vector which ith element m_i is the mass of species A_i or the row vector of lumping matrix M

$\bar{M}$ = generalized inverse of M satisfying $M\bar{M} = I_{\hat{n}}$

P = $(n - \hat{n}) \times (n - \hat{n})$ constant matrix

$P(\mathbf{y})$ = $(n - \hat{n}) \times (n - \hat{n})$ function matrix

$Q(\mathbf{y})$ = $m \times m$ function matrix

S = $(n - \hat{n}) \times m$ constant matrix

$\bar{S}$ = generalized inverse of S satisfying $\bar{S}S = I_m$

W = intersection of Z and X

$\mathbf{x}$ = n-dimensional vector

$\bar{\mathbf{x}}$ = image of $\mathbf{x}$ upon mapping by $(I_n - \bar{M}M)$

$\hat{\mathbf{x}}$ = image of $\bar{\mathbf{x}}$ upon mapping by $J(\mathbf{y})$

X = $n \times (n - \hat{n})$ constant matrix or intersection of the invariant subspaces of the Jacobian matrix $J(\mathbf{y})$

$\bar{X}$ = generalized inverse of X satisfying $\bar{X}X = I_{n-\hat{n}}$

$\bar{X}_i$ = intersection of the invariant subspaces of the Jacobian matrix $J(\mathbf{y})$

X_i = invariant subspace of $J(\mathbf{y}_i)$

X_{A_k} = eigenvector matrix of A_k

$\mathbf{y}$ = n-dimensional variable vector

$\hat{\mathbf{y}}$ = $\hat{n}$-dimensional variable vector

Y = subspace of n-dimensional space, which contains X

Z = $n \times m$ constant matrix or subspace of X, which is invariant to $J(\mathbf{y})$

Z_i = intersection between Z and X_i

Greek Letters

λ_i = ith eigenvalue of matrix A_k or K

Λ = diagonal eigenvalue matrix of matrix K with λ_i as its diagonal elements

Symbols

$\hat{}$ = any property related to the lumped system

$\mathbf{0}$ = null vector

0 = null matrix

REFERENCES

[1] Jacob, S.M., B. Gross, S.E. Voltz, and V.W. Weekman, Jr., AIChE J., **22**, 701(1976).

[2] Wei, J., and J.C.W. Kuo, Ind. Eng. Chem. Fundamentals, **8**, 114(1969).

[3] Ozawa, Y., Ind. Eng. Chem. Fundamentals, **12**, 191(1973).

[4] Bailey, J.E., Chem. Eng. J., **3**, 52(1972).

[5] Bailey, J.E., AIChE J., **21**, 192(1975).

[6] Li, G., Chem. Eng. Sci., **39**, 1261(1984).

[7] Golikeri, S.V., and D. Luss, AIChE J., **18**, 277(1972).

[8] Hutchinson, P., and D. Luss, Chem. Eng. J., **1**, 129(1970).

[9] Luss, D., and P. Hutchinson, Chem. Eng. J., **2**, 172(1971).

[10] Golikeri, S.V., and D. Luss, Chem. Eng. Sci., **29**, 845(1974).

[11] Isral, A.B., and T.N.E. Greville, Generalized Inverse: Theory and Applications, John Wiley & Sons, Inc., New York, 1974.

[12] Graham, A., Knonecker Products and Matrix Calculus: with Applications, Ellis Horwood Limited, New York, 1981.

[13] Yetter, R.A., F.F. Dryer, and H. Rabitz, Combustion and Flame, **59**, 107(1985).