

International Series of Numerical Mathematics
Internationale Schriftenreihe zur Numerischen Mathematik
Série internationale d'Analyse numérique
Vol. 70

ISNM 70

Numerical Methods for Bifurcation Problems

Edited by

**T. Küpper
H. D. Mittelmann
H. Weber**

Springer Basel AG

B

ISNM 70:

International Series of Numerical Mathematics

Internationale Schriftenreihe zur Numerischen Mathematik

Série internationale d'Analyse numérique

Vol. 70

Edited by

Ch. Blanc, Lausanne; A. Ghizzetti, Roma;

R. Glowinski, Paris; G. Golub, Stanford;

P. Henrici, Zürich; H. O. Kreiss, Pasadena;

A. Ostrowski, Montagnola; J. Todd, Pasadena

Springer Basel AG

Numerical Methods for Bifurcation Problems

**Proceedings of the Conference at the University of Dortmund,
August 22–26, 1983**

Edited by

**T. Küpper
H. D. Mittelmann
H. Weber**

Springer Basel AG 1984

Editors

T. Küpper
H. D. Mittelmann
Universität Dortmund
Abt. Mathematik
Postfach 50 05 00
D-4600 Dortmund 50

H. Weber
Johannes Gutenberg-Universität
Rechenzentrum
Postfach 3980
D-6500 Mainz 1

Library of Congress Cataloging in Publication Data

Numerical methods for bifurcation problems.

(International series of numerical mathematics ; v. 70)

1. Numerical analysis—Congresses. 2. Bifurcation theory—Congresses. I. Küpper, T. (Tassilo), 1947—
II. Mittelmann, H. D., 1945— , III. Weber, H., 1948— . IV. Series.
QA297.N864 1984 519.4 84-9286
ISBN 978-3-0348-6257-8 ISBN 978-3-0348-6256-1 (eBook)
DOI 10.1007/978-3-0348-6256-1

CIP-Kurztitelaufnahme der Deutschen Bibliothek

Numerical methods for bifurcation problems :
proceedings of the conference at the Univ. of
Dortmund, August 22 – 26, 1983 / ed. by
T. Küpper ...

(International series of numerical mathematics ; 70)
ISBN 978-3-0348-6257-8

NE: Küpper, Tassilo [Hrsg.]; Universität
(Dortmund); GT

All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

© 1984 Springer Basel AG

Originally published by Birkhäuser Verlag Basel in 1984
Softcover reprint of the hardcover 1st edition 1984

ISBN 978-3-0348-6257-8

Foreword

These are the Proceedings of the Conference on

Numerical Methods for Bifurcation Problems

which was held August 22 - 26, 1983 at the University of Dortmund as a sequel to a smaller Workshop held at Dortmund in 1980. The conference was organized in connection with the GAMM committees Diskretisierende Methoden in der Festkörpermechanik and Effiziente numerische Verfahren für partielle Differentialgleichungen.

The aim of this conference was to provide an up to date survey on the current state of the art in applying numerical methods to bifurcation problems as well as to inform about most recent developments and, last not least, to bring together numerical analysts and engineers working in the field. For that reason papers of both expository and research character were welcome. The published articles have been subject to a reviewing process.

The majority of the participants were numerical analysts, and the volume reflects this fact. A number of articles, however, deal with problems of direct relevance to the applied sciences, in particular to problems in biology and engineering.

It is a pleasure for us to thank several agencies for their help. The conference was mainly supported by a generous grant of the Deutsche Forschungsgemeinschaft. Participation of colleagues from Eastern European countries was made possible through a special program of the Deutscher Akademischer Austauschdienst. A donation of the Freundesgesellschaft der Universität Dortmund helped us to cover organizational costs. We further acknowledge the hospitality of the University of Dortmund and we like express our sincerest thanks to Mrs. U. Kriegelstein and Mrs. I. Speck for their helpful assistance. Finally we greatly appreciate the rapid publication of these proceedings and we express our gratitude to Birkhäuser-Verlag and to the editors of the ISNM-series.

December 1983

Tassilo Küpper

Hans Detlef Mittelmann

Helmut Weber

INDEX

Foreword.....	5
List of Participants.....	11
ALLGOWER, E.:	
Bifurcations Arising in the Calculation of Critical Points via Homotopy Methods.....	15
BERNARDI, Ch., RAPPAZ, J.	
Approximation of Hopf Bifurcation for Semilinear Parabolic Equations.....	29
BEYN, W.-J.:	
Defining Equations for Singular Solutions and Numerical Applications.....	42
BEYN, W.-J., BOHL, E.:	
Organizing Centers for Discrete Reaction Diffusion Models....	57
BOHL, E.:	
Discrete Versus Continuous Models for Dissipative Systems....	68
BREZZI, F., USHIKI, S., FUJII, H.:	
"Real" and "Ghost" Bifurcation Dynamics in Difference Schemes for ODEs.....	79
CALUWAERTS, R.:	
The Use of Extrapolation for the Solution of Bifurcation Problems.....	105
CHAN, T.F.:	
Techniques for Large Sparse Systems Arising from Continuation Methods.....	116
CLIFFE, K.A., SPENCE, A.	
The Calculation of High Order Singularities in the Finite Taylor Problem.....	129
DESCLOUX, J.:	
On Hopf and Subharmonic Bifurcations.....	145
GRIEWANK, A.:	
Quadratically Appended Linear Models for Locating Generalized Turning Points.....	162
HADELER, K.P.:	
Hysteresis in a Model for Parasitic Infection.....	171

HOLODNIOK, M., KUBÍČEK, M.	
Continuation of Periodic Solutions in Ordinary Differential Equations - Numerical Algorithm and Application to Lorenz Model.....	181
JEPSON, A.D., SPENCE, A.:	
Singular Points and their Computation.....	195
KEARFOTT, R.B.:	
On a General Technique for Finding Directions Proceeding from Bifurcation Points.....	210
KELLER, H.B., JEPSON, A.D.::	
Steady State and Periodic Solution Paths: their Bifurcations and Computations.....	219
KUBÍČEK M., HOLODNIOK, M.	
Numerical Determination of Bifurcation Points in Steady State and Periodic Solutions - Numerical Algorithms and Examples...	247
KUEPPER, T., KUSZTA, B.:	
Feedback Stimulated Bifurcation.....	271
LANGFORD, W.F.:	
Numerical Studies of Torus Bifurcations.....	285
MACKENS, W., JARAUSCH, H.::	
Numerical Treatment of Bifurcation Branches by Adaptive Condensation.....	296
MENZEL, R.:	
Numerical Determination of Multiple Bifurcation Points.....	310
MITTELMANN, H.D.::	
Continuation Near Symmetry-Breaking Bifurcation Points.....	319
MOORE, G.:	
The Numerical Buckling of a Visco-elastic Rod.....	335
MUNZ, H.:	
Asymptotic Error Expansion for Finite Difference Schemes for Elliptic Systems Near Turning Points.....	344
PEITGEN, H.-O., PRUEFER, M.:	
Global Aspects of Newton's Method for Nonlinear Boundary Value Problems.....	352
RAUGEL, G., GEYMONAT, G.:	
Finite Dimensional Approximation of Some Bifurcation Problems in Presence of Symmetries.....	369
REDDIEN, G.W., GRIEWANK, A.:	
Computation of Generalized Turning Points and Two-Point Boundary Value Problems.....	385

RHEINBOLDT, W.C.:		
On Some Methods for the Computational Analysis of Manifolds..		401
ROOSE, D., CALUWAERTS, R.:		
Direct Methods for the Computation of a Nonsimple Turning Point Corresponding to a Cusp.....		426
SCHEIDL, R.		
On the Axisymmetric Buckling of Thin Spherical Shells.....		441
SCHOLZ; R.:		
On the Rate of Convergence for the Approximation of Nonlinear Problems.....		452
SCHWETLICK, H.:		
Algorithms for Finite-Dimensional Turning Point Problems from Viewpoint to Relationships with Constrained Optimization Methods.....		459
SEYDEL, R.:		
A Continuation Algorithm with Step Control.....		480
SLJBRAND, J., DIPRIMA, R.C., EAGLES, P.M.:		
Bifurcation Near Multiple Eigenvalues for the Flow Between Concentric Counterrotating Cylinders.....		495
SPENCE, A., JEPSON, A.D.:		
The Numerical Calculation of Cusps, Bifurcation Points and Isola Formation Points in Two Parameter Problems.....		502
STEINRUECK, H., TROGER, H., WEISS, R.:		
Mode Jumping of Imperfect, Buckled, Rectangular Plates.....		515
TROGER, H.:		
Application of Bifurcation Theory to the Solution of Nonlinear Stability Problems in Mechanical Engineering.....		525
WEBER, H.:		
A Singular Multi-Grid Iteration Method for Bifurcation Problems.....		547
WERNER, B.:		
Regular Systems for Bifurcation Points with Underlying Symmetries.....		562
WEYER, J., ASTABURUAGA, M.A., FIGUEROA, J.:		
Newton Iterates for Positive Solutions of a Class of Nonlinear Eigenvalue Problems.....		575

List of Participants

Allgower, E.	Dept. of Math., Colorado State Univ., Fort Collins, Colorado 80523
Bernardi, Christine	C.N.R.S. and University P. et M. Curie, Analyse Numérique Université P. et M. Curie, 4 Place Jussieu 75230 Paris, Cedex 05, France
Beyn, Wolf-Jürgen	Fakultät für Mathematik, Universität Konstanz Postfach 5560, D-7750 Konstanz
Bischoff, Dieter	Univ. Hannover, I.f. Baumech. u. Numer. Mech. Callinstr. 32, 3000 Hannover 1
Bohl, Erich	Universität Konstanz, Fakultät für Mathematik Postfach 5560, 7750 Konstanz
Bolley, Catherine	I.N.S.A. Laboratoire d'Analyse Numérique 20 Av. des Bultes de Cœsmes 35043 Rennes Cedex, France
Börsch-Supan, W.	FB 17 Mathematik, Joh. Gutenberg-Univ. Mainz Postfach 3980, 6500 Mainz
Braess, Dietrich	Institut für Mathematik, Ruhr-Universität Bochum Universitätsstr. 150, Geb. NA, 463 Bochum
Brezzi, Franco	Dept. of Structural Mechanics and I.A.N. of C.N.R. Università di Pavia, C.S.O. Strada Nuova 65 27100 Pavia, Italy
Brown, Adrian	University of Bath, School of Mathematics Claverton Down, Bath BA2 7AY, England
Caluwaerts, Renaat	Katholieke Universiteit Leuven Dept. Computerwetenschappen Celestijnlaan 200 A, 3030 Heverlee, Belgien
di Carlo, Antonio	Università di Roma "La Sapienza" Istituto di Scienza delle Costruzioni, Facoltà di Ingegneria, Via Eudossiana 18, 00184 Roma, Italy
Chan, Tony F.	Yale University, Computer Science Dept. Box 2158, Yale Station, New Haven, CT 06520, USA
Cliffe, K.A.	Aere Harwell, Theoretical Physics Division Oxfordshire, OX11 ORA
Descloux, J.	EPFL Lausanne, Dept. Mathematik MA (Eublens), CH-1015 Lausanne
Döring, B.	Mathematisches Institut der Universität Düsseldorf Universitätsstr. 1, 4 Düsseldorf 1
Göthel, Rainer	Mathematisches Institut Universität Dortmund
Greenway, Philip	University of Bath, School of Mathematics Claverton Down, Bath, England
Griewank, Andreas	Dept. of Mathematics SMV Southern Methodist University Dallas, Tx 75275 USA

Hackbusch, W.	Institut für Informatik und Praktische Mathematik Universität Kiel, Olshausenstr. 40-60, 2300 Kiel 1
Hadeler, K. P.	Universität Tübingen, Lehrstuhl für Biomathematik Auf der Morgenstelle 28, 7400 Tübingen
Holodniok, Martin	Computing Centre, Prague Institute of Chemical Technology, Suchbatarova 3, 166 28 Praha 6, Czechoslovakia
Jepson, Allan D.	Dept. of Computer Science University of Toronto, Toronto Ontario, Canada M5S IA7
Kearfott Baker, R.	Dept. of Mathematics/Statistics University of Southwestern Louisiana U.S.L Box 4-1010, Lafayette, Louisiana 70504-1010
Keller, H. B.	Cal. Tech., Firestone Lab. Pasadena, CA - 91125
Kubicek, Milan	Dept. of Chem. Engng., Prague Institute of Chemical Technology Suchbatarova 3, 166 28 Praha 6, Czechoslovakia
Küpper, Tassilo	Mathematisches Institut, Universität Dortmund
Langford, William F.	Dept. of Mathematics and Statistics University of Guelph Guelph, Ontario, Canada NIG 2WI
Lorenz, Jens	Universität Trier, FB IV, Mathematik, 55 Trier (für Korrespondenz Fak. f. Math., Univ. Konstanz, 775 Konstanz)
Mackens, Wolfgang	Institut für Geometrie und Praktische Mathematik der RWTH Aachen, Templergraben 55, D - 5100 Aachen
Mittelmann, Hans	Mathematisches Institut, Universität Dortmund
Mooney, John	Mathematics Department Glasgow College of Technology, Cowcaddens Road Glasgow G4 0BA, Scotland
Moore, G.	National Institute for Higher Education Glasnevin, Dublin 9, Ireland
Munz, Harry	Universität Tübingen, Lehrstuhl für Biomathematik Auf der Morgenstelle 28, 7400 Tübingen
Paffrath, Meinhard	SFB 72, Universität Bonn Goetheallee 25, 53 Bonn 3
Pasquali, A.	Istituto Matematico "U. Dini" Viale Morgagni 67/A, 50134 Firenze, Italy
Peitgen, Heinz-Otto	Forschungsschwerpunkt "Dynamische Systeme" Universität Bremen, D-2800 Bremen 33
Prüfer, Michael	Forschungsschwerpunkt "Dynamische Systeme" Universität Bremen, D-2800 Bremen 33
Raugel, Geneviève	C.N.R.S. and Université de Rennes 6, boulevard Jourdan, 75014 Paris, France
Reddien, George W.	Mathematics Department, Southern Methodist Univ. Dallas TX 75275, USA

Rheinboldt, Werner C.	University of Pittsburgh Department of Mathematics and Statistics Pittsburgh, PA 15260
Roose, Dirk	Katholieke Universiteit Leuven Dept. of Computer Science Celestynenlaan 200A, B3030 Leuven, Belgien
Scheidl, Rudolf	Institut für Mechanik, TU Wien Karlsplatz 13, A - 1040 Wien
Scholz, Reinhard	Institut für Angewandte Mathematik, Univ. Freiburg Hermann-Herder-Str. 10, 7800 Freiburg i. Br.
Schrauf, G.	SFB 72 Universität Bonn, Institut für Angew. Math Beringstr. 4, 5300 Bonn 1
Schwertlick, H.	Martin-Luther-Universität Halle, Sektion Math. Weinbergweg 17, 4010 Halle, DDR
Schwichtenberg, Horst	Gesellschaft für Mathematik und Datenverarbeitung Postfach 1240, 5205 St. Augustin 1
Seydel, R.	Mathematisches Institut, TU München Arcisstr. 21, 8000 München 2
Sirl, David	University of Bath, School of Chemical Engineering Claverton Down, Bath, England
Söhnen, Andreas	Joh. Gutenberg Universität Mainz Blüssusstr. 3, 6500 Mainz
Sijbrand, J.	Shell Research BV Badhuisweg 3, 1031 CM Amsterdam, Netherlands
Spence, Alastair	University of Bath, School of Mathematics Claverton Down, Bath BA2 7AY, England
Tatone, Amabile	Università Dell'Aquila, Facoltà di Ingegneria Loc. Monteluco, 67100 L'Aquila, Italy
Thole, Clemens-August	Gesellschaft für Mathematik und Datenverarbeitung Postfach 1240, 5205 St. Augustin 1
Troger, Hans	Institut für Mechanik Karlsplatz 13, A - 1040 Wien
Verfürth, Rüdiger	Ruhr-Universität Bochum Universitätsstr. 150, 4630 Bochum
Voß, H.	FB Mathematik, Universität Essen - GHS Universitätsstr. 3, 4300 Essen 1
Weber, Helmut	Universität Mainz, Rechenzentrum Postfach 3980, 6500 Mainz 1
Weiβ, Richard	Technische Universität Wien, Institut für Angewandte und Numerische Mathematik Gussbausstr. 27-29, A - 1040 Wien
Werner, B.	Universität Hamburg Institut für Angewandte Mathematik Bundesstr. 55, D - 2000 Hamburg 13
Weyer, Jürgen	Mathematisches Institut, Universität zu Köln Weyertal 86-90, D - 5000 Köln 41
Zeman, Klaus	Institut für Mechanik/Mech. II, TU Wien Karlsplatz 13, 1040 Wien

BIFURCATIONS ARISING IN THE CALCULATION
OF CRITICAL POINTS VIA HOMOTOPY METHODS

Eugene L. Allgower
Colorado State University
Fort Collins, CO 80523

1. Introduction

Among the applications of numerical continuation methods is the determination of the critical points of a mapping $f: \mathbb{R}^N \rightarrow \mathbb{R}^1$. One may choose a mapping $e: \mathbb{R}^N \rightarrow \mathbb{R}^1$ such that the zero points a_j , $j = 1, \dots, m$ of $E := \nabla e$ are known and are regular points of E . One then formulates a smooth deformation map H such that $H(t, x) \in \mathbb{R}^N$ for $(t, x) \in \mathbb{R}^1 \times \mathbb{R}^N$ and

$$H(0, x) = E(x) \text{ and } H(1, x) = F(x) := \nabla f(x).$$

The numerical aspect consists of tracing the smooth 1-manifolds $c(a_j) \subset H^{-1}(0)$ with $(0, a_j) \in c(a_j)$, $j = 1, \dots, m$.

Of course, it may occur that $c(a_j) \cap (\{1\} \times \mathbb{R}^N) = \emptyset$, in which case $c(a_j)$ doesn't converge to any critical point of f . On the other hand, the following result indicates that the continuation method has an appealing property which might permit "targeting" critical points having a specific Morse index.

(1.1) THEOREM. [2] Let e, E, f, F, H be as described above and let a be a critical point of e . Let $c(a) = \{(t(s), x(s)) \in \mathbb{R}^{N+1} | H(t(s), x(s)) = 0, (t(0), x(0)) = (0, a)\}$ be a smooth 1-manifold such that

- (i) $t(s)$ is monotone increasing for $s \in [0, s^*]$,
- (ii) $(t(s^*), x(s^*)) = (1, b)$.

Then the critical points a (of e) and b (of f) have the same Morse index i.e. $E(a)=0$, $F(b)=0$ and the Hessian matrices $D^2e(a)$ and $D^2f(b)$ have the same number of negative eigenvalues.

One may also apply homotopy continuation methods to determine the zero points of complex analytic maps $F: \mathbb{C}^N \rightarrow \mathbb{C}^N$. Analogously, one may regard curves of the form

$$c(a) = \{(t(s), z(s)) \in \mathbb{R}^1 \times \mathbb{C}^N | H(t(s), z(s)) = 0, (t(0), z(0)) = (0, a)\}.$$

It has been observed [4] that if $H(t, z)$ is holomorphic in z for all t , and if $c(a) \subset H^{-1}(0)$ is a smooth 1-manifold, then $t(s)$ is monotone increasing for $s \in [0, \infty)$.

It is plain that if $H: \mathbb{R}^1 \times \mathbb{C}^N \rightarrow \mathbb{C}^N$, one may write

$$(1.2) \quad H(t, x + iy) = H^r(t, x, y) + iH^i(t, x, y) \quad \text{where } H^r, H^i \text{ are defined by}$$

$$(1.3) \quad H^r(t, x, y) := \frac{1}{2}(H(t, x+iy) + H(t, x-iy))$$

$$H^i(t, x, y) := \frac{-i}{2}(H(t, x+iy) - H(t, x-iy)).$$

Thus, one may reformulate the complex map $H: \mathbb{R}^1 \times \mathbb{C}^N \rightarrow \mathbb{C}^N$ equivalently as a real map $H^c: \mathbb{R}^1 \times \mathbb{R}^{2N} \rightarrow \mathbb{R}^{2N}$ where

$$(1.4) \quad H^c(t, x, y) = \begin{pmatrix} H^r(t, x, y) \\ H^i(t, x, y) \end{pmatrix}.$$

If H is holomorphic for all t , and if $c^c(a) \subset H^{-1}(0)$ is a smooth 1-manifold, then $t(s)$ is monotone increasing for $s \in [0, \infty)$.

It has been suggested [10] that the complex imbedding (1.4) be used to target critical points as in (1.1). Although this idea seems appealing and may in fact often work, there are certain difficulties. We show here that if (t^*, x^*) is a turning point of $c(a) \subset H^{-1}(0)$ (relative to t) then $(t^*, x^*, 0)$ is a bifurcation point of $c^c(a)$. Furthermore, the eigenvalues of H_x^r may change sign at regular points of the complex 1-manifold. Thus, any practical implementation of a continuation method for H^c must incorporate a test whether the Morse index has indeed been preserved on one of the branches when a bifurcation back onto a real curve in $H^{-1}(0)$ takes place.

2. Davidenko's Equation

In order to fix the setting for the subsequent discussion we assume that

$$(2.1) \quad f: \mathbb{R}^N \rightarrow \mathbb{R}^1 \text{ is a real analytic map and } F = \nabla f,$$

$$(2.2) \quad e: \mathbb{R}^N \rightarrow \mathbb{R}^N \text{ is a real analytic map, so chosen that } a_j, j = 1, \dots, m \text{ are known critical points (i.e. zero points of } E = \nabla e), \text{ which are also regular points of } E \text{ (i.e. rank } D^2 e(a_j) = N).$$

$$(2.3) \quad H(t, x) \in \mathbb{R}^N \text{ for } (t, x) \in \mathbb{R}^1 \times \mathbb{R}^N \text{ such that}$$

$$(i) \quad H(0, x) = E(x), \quad H(1, x) = F(x)$$

$$(ii) \quad H \text{ is analytic in } x \text{ for all } t \in \mathbb{R}^1 \text{ and } C^\infty \text{ with respect to } t.$$

By the implicit function theorem, there exist smooth curves

$$(2.4) \quad c(a_j) = \{(t(s), x(s)) \in \mathbb{R}^{N+1} \mid H(t(s), x(s)) = 0 \text{ and } (t(0), x(0)) = (0, a_j), s \in [0, s^*], s^* \leq \infty\}, \quad j = 1, \dots, m.$$

Furthermore, by a generalized version of Sard's theorem, the "generic situation" [1], [3] is that the $c(a_j)$ are diffeomorphic to either \mathbb{R}^1 or S^1 (the 1-dimensional unit ball). It is convenient for the subsequent discussion (and no loss of generality) to regard the curve $c(a_j)$ as being parametrized according to arc length viz. for $(t(s), x(s)) \in c(a_j)$, the equation

$$(2.5) \quad (\dot{t}(s))^2 + \|x(s)\|_2^2 = 1 \quad \text{holds for } s \in [0, s^*],$$

where $\cdot \equiv d/ds$.

If $c(a_j)$ is diffeomorphic to \mathbb{R}^1 or S^1 , just two occurrences are possible:

$$(A) \quad t(\bar{s}) = 1 \quad \text{for some } \bar{s} \in [0, \infty)$$

$$(B) \quad c(a_j) \cap (\{1\} \times \mathbb{R}^N) = \emptyset.$$

Case (A) represents a successful convergence to a critical point $(t(\bar{s}), x(\bar{s})) = (1, b)$ of f . If $\dot{t}(s) \geq 0$ for $s \in [0, \bar{s}]$, we say $c(a_j)$ is monotone. If there is a point \tilde{s} where

$$(2.6) \quad \dot{t}(\tilde{s} - \varepsilon)\dot{t}(\tilde{s} + \varepsilon) < 0 \text{ for } \varepsilon \in (0, \varepsilon_0), \varepsilon_0 > 0$$

we say $c(a_j)$ has a turning point at $(t(\tilde{s}), x(\tilde{s}))$ (with respect to t). There are two ways in which case (B) can occur if $c(a_j)$ is diffeomorphic to \mathbb{R}^1 or S^1 :

- $$(2.7) \quad \begin{aligned} (a) \quad & \sup_s t(s) < 1 \text{ (} c(a_j) \text{ "turns back")} \\ (b) \quad & \sup_s t(s) = 1, \text{ but } t(s) < 1 \text{ for all } s \in [0, \infty). \end{aligned}$$

In this case $\|x(s)\| \rightarrow \infty$ as $s \rightarrow \infty$.

To summarize, a curve $c(a) \subset H^{-1}(0)$ of the type described in (2.4) satisfies the equations

- $$(2.8) \quad \begin{aligned} (i) \quad & (\dot{t})^2 + \|\dot{x}\|_2^2 = 1 \\ (ii) \quad & H(t(s), x(s)) = 0 \\ (iii) \quad & (t(0), x(0)) = (0, a). \end{aligned}$$

If we differentiate (2.8)(ii) with respect to s , we obtain the Davidenko initial-value problem:

$$(2.9) \quad \begin{bmatrix} \dot{t} & \dot{x}^T \\ H_t & H_x \end{bmatrix} \begin{bmatrix} \dot{t} \\ \dot{x} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{on } c(a)$$

and $(t(0), x(0)) = (0, a)$.

If $c(a)$ is diffeomorphic to \mathbb{R}^1 or S^1 , then $DH(t(s), x(s))$ has rank $= N$ for $(t(s), x(s)) \in c(a)$ and hence the augmented Jacobian

$$A(s) := \begin{bmatrix} \dot{t}(s) & \dot{x}(s)^T \\ H_t(t(s), x(s)) & H_x(t(s), x(s)) \end{bmatrix}$$

is nonsingular for $s \in [0, \bar{s}]$ since $(\dot{t}(s), \dot{x}(s)^T)$ is orthogonal to the row space of $DH(t(s), x(s))$.

It is now possible to give a simple proof of Theorem (1.1):

Note that

$$A(s) \begin{bmatrix} \dot{t} \\ \dot{x}(s) \end{bmatrix} = \begin{bmatrix} 1 & \dot{x}^T \\ 0 & H_x(t(s), x(s)) \end{bmatrix}$$

Hence

$$(2.10) \quad \dot{t} \det A(s) = \det H_x(t(s), x(s)).$$

Since $A(s)$ is nonsingular on $c(a)$, we have:

$$(2.11) \quad \text{On } c(a), \det H_x(t(s), x(s)) \text{ changes sign exactly when } \dot{t}(s) \text{ changes sign.}$$

In particular, if $\dot{t}(s) \geq 0$ for $s \in [0, s^*]$, and if $(t(s^*), x(s^*)) = (1, b)$, then the Hessian matrices

$$H_x(t(0), x(0)) = D^2e(a) \text{ and } H_x(t(s^*), x(s^*)) = D^2f(b)$$

have the same number of negative eigenvalues.

The drawback in applying Theorem 1.1 is that in general it is difficult to choose the initial map e (or E) so that one knows a priori that the curve $c(a) \subset H^{-1}(0)$ will be monotone.

3. The Complex Analytic Case

Suppose $H(t, z) \in \mathbb{C}^N$ for $(t, z) = (t, x + iy) \in \mathbb{R}^1 \times \mathbb{C}^N$ such that $H(t, z)$ is holomorphic in z for all $t \in \mathbb{R}^1$ and C^∞ with respect to t . Instead of considering the real mapping H^C defined in (1.4), it is more convenient to study the map

$$(3.1) \quad \hat{H}(t, x, y) := \begin{pmatrix} H^r(t, x, y) \\ -H^i(t, x, y) \end{pmatrix}.$$

Clearly

$$(3.2) \quad H(t, x + iy) = 0 \text{ if and only if } H^C(t, x, y) = 0 \\ \text{if and only if } \hat{H}(t, x, y) = 0.$$

The Davidenko equation for \hat{H} assumes the form

$$(3.3) \quad \begin{pmatrix} \dot{t} & \dot{x}^T & \dot{y}^T \\ H_t^r & H_x^r & H_y^r \\ -H_t^i & -H_x^i & -H_y^i \end{pmatrix} \begin{pmatrix} \dot{t} \\ \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

From (1.3) we have

$$H_y^r = -H_x^i \quad \text{and} \quad H_y^i = H_x^r.$$

Hence (3.3) may be rewritten as

$$(3.4) \quad \begin{pmatrix} \dot{t} & \dot{x}^T & \dot{y}^T \\ H_t^r & H_x^r & -H_x^i \\ -H_t^i & -H_x^i & -H_x^r \end{pmatrix} \begin{pmatrix} \dot{t} \\ \dot{x} \\ \dot{y} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

Hence

$$\begin{pmatrix} \dot{t} & \dot{x}^T & \dot{y}^T \\ H_t^r & H_x^r & -H_x^i \\ -H_t^i & -H_x^i & -H_x^r \end{pmatrix} \begin{pmatrix} \dot{t} & 0^T \\ \dot{x} & I_{2N} \end{pmatrix} = \begin{pmatrix} 1 & \dot{x}^T & \dot{y}^T \\ 0 & H_x^r & -H_x^i \\ 0 & -H_x^i & -H_x^r \end{pmatrix}$$

yields

$$(3.5) \quad \dot{t} \det \begin{pmatrix} \dot{t} & \dot{x}^T & \dot{y}^T \\ \hat{H}_t & \hat{H}_x & \hat{H}_y \end{pmatrix} = \det \begin{pmatrix} H_x^r & -H_x^i \\ -H_x^i & -H_x^r \end{pmatrix} = \det(\hat{H}_x \hat{H}_y).$$

Since

$$D_{x,y}\hat{H} = (\hat{H}_x, \hat{H}_y) = \begin{pmatrix} H_x^r & -H_x^i \\ -H_x^i & -H_x^r \end{pmatrix},$$

it follows that if λ is an eigenvector of $D_{x,y}\hat{H}$ having corresponding eigenvector $(\begin{smallmatrix} u \\ v \end{smallmatrix})$, then so is $-\lambda$ an eigenvalue having corresponding eigenvector $(\begin{smallmatrix} v \\ -u \end{smallmatrix})$. Hence the eigenvalues of $D_{x,y}\hat{H}$ occur in symmetric pairs about 0 and

$$(3.6) \quad (-1)^N \det D_{x,y}\hat{H} \geq 0 \text{ holds everywhere.}$$

In particular,

$$(3.7) \quad \det D_{x,y}\hat{H} \text{ never changes sign.}$$

Hence, by (3.5) we have

$$(3.8) \quad \text{If } \hat{c} = \{(t(s), x(s), y(s)) | s \in [s_0, s_1]\} \subset \hat{H}^{-1}(0),$$

then $\dot{t}(s)$ changes sign exactly when

$$\det \begin{pmatrix} \dot{t}(s) & \dot{x}(s)^T & \dot{y}(s)^T \\ \hat{H}_t(t(s), x(s), y(s)) & \hat{H}_x(t(s), x(s), y(s)) & \hat{H}_y(t(s), x(s), y(s)) \end{pmatrix}$$

changes sign.

From (3.8) and the result of Crandall and Rabinowitz [5] we have:

$$(3.9) \quad \text{If } \hat{c} \text{ has a turning point in } t \text{ at } s^* \in (s_0, s_1), \text{ then } (t(s^*), x(s^*), y(s^*)) \text{ is a bifurcation point of } \hat{H}^{-1}(0).$$

Similarly, $(t(s^*), x(s^*), y(s^*))$ is a bifurcation point of $H^C(0)$ and $(t(s^*), x(s^*) + iy(s^*))$ is a bifurcation point of $H^{-1}(0) \subset \mathbb{R}^1 \times \mathbb{C}^N$.

From (3.8) we also have conversely,

$$(3.10) \quad \text{If } \hat{H}^{-1}(0) \text{ has no bifurcation, then } \hat{c}(a) \text{ is monotone.}$$

4. Complexification

Let $H: \mathbb{R}^1 \times \mathbb{R}^N \rightarrow \mathbb{R}^N$ be as in (2.3). The complex extension $H^C: \mathbb{R}^1 \times \mathbb{C}^N \rightarrow \mathbb{C}^N$ of H is defined by

$$H^C(t, z) := H(t, \operatorname{Re}^{-1}(z)) \text{ where for } z = x+iy, \operatorname{Re}(x+iy) = x.$$

Since H is real analytic in x for all t , we have

$$H^C(t, \bar{z}) = \overline{H^C(t, z)} \text{ for all } (t, z) \in \mathbb{R}^1 \times \mathbb{C}^N.$$

Hence,

$$(4.1) \quad H^C(t, x+iy) = 0 \text{ if and only if } H^C(t, x-iy) = 0.$$

Since $H(0, a) = E(a) = 0$, also

$$\hat{H}(0, a, 0) = \begin{pmatrix} H^r(0, a, 0) \\ -H^i(0, a, 0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ by (1.3) and}$$

$$\hat{D}\hat{H}(0, a, 0) = \begin{pmatrix} H_x^r(0, a, 0) & H_y^r(0, a, 0) \\ -H_t^i(0, a, 0) & -H_y^i(0, a, 0) \end{pmatrix} = \begin{pmatrix} H_x(0, a) & 0 \\ 0 & -H_x(0, a) \end{pmatrix}.$$

Since $H_x(0, a)$ has rank N , $\hat{D}\hat{H}(0, a, 0)$ has rank $2N$. Thus, $(0, a, 0) \in \mathbb{R}^1 \times \mathbb{R}^{2N}$ is a regular zero point of \hat{H} . Since $H(t, x) = 0$ implies $\hat{H}(t, x, 0) = 0$, the curve

$$\begin{aligned} \hat{c}(a) &= \{(t, x, 0) \in \mathbb{R}^1 \times \mathbb{R}^{2N} | \hat{H}(t, x, 0) = 0\} \\ &= \{(t, x) \in \mathbb{R}^1 \times \mathbb{R}^N | H(t, x) = 0\} \times \{0\} \\ &= c(a) \times \{0\} \subset \mathbb{R}^1 \times \mathbb{R}^{2N}. \end{aligned}$$

From (3.9) we see that if $c(a)$ has a turning point (t^*, x^*) then $\hat{H}^{-1}(0)$ has a bifurcation point at $(t^*, x^*, 0)$.

Let

$$\hat{A}(s) = \begin{pmatrix} \dot{t}(s) & \dot{x}(s)^T & \dot{y}(s)^T \\ H_t^r(t(s), x(s), y(s)) & H_x^r(t(s), x(s), y(s)) & H_y^r(t(s), x(s), y(s)) \\ -H_t^i(t(s), x(s), y(s)) & -H_x^i(t(s), x(s), y(s)) & -H_y^i(t(s), x(s), y(s)) \end{pmatrix}$$

denote the augmented Jacobian corresponding to \hat{H} . Then on $\hat{c}(a)$ we have

$$\hat{A}(s) = \begin{pmatrix} \dot{t} & \dot{x}^T & 0^T \\ H_t^r & H_x^r & -H_x^i \\ -H_t^i & -H_x^i & -H_x^r \end{pmatrix} = \begin{pmatrix} \dot{t} & \dot{x}^T & 0^T \\ H_t & H_x & 0 \\ -H_t^i & 0 & -H_x \end{pmatrix}$$

Since $\text{rank} \begin{pmatrix} \dot{t} & \dot{x}^T \\ H_t & H_x \end{pmatrix} = N + 1$ on $c(a)$, also $\text{rank } H_x \geq N - 1$ on $c(a)$. Hence,

$\text{rank } \hat{A}(s) \geq 2N$ on $\hat{c}(a)$. Thus,

(4.2) If $(t(s^*), x(s^*))$ is a turning point of $c(a)$, then $(t(s^*), x(s^*), 0)$ is a simple bifurcation point of $\hat{H}^{-1}(0)$.

It is also not difficult to show

(4.3) If $(t(s^*), x(s^*), y(s^*))$ is a turning point of $\hat{c}(a)$, then $\pm(0, \dot{x}(s^*), \dot{y}(s^*))$ and $\pm(0, -\dot{y}(s^*), \dot{x}(s^*))$ are the unit tangents to $\hat{H}^{-1}(0)$ at the bifurcation point $(t(s^*), x(s^*), y(s^*))$.

In view of (3.5), it is always possible to choose $(\dot{t}(s), \dot{x}(s), \dot{y}(s))$ so that $\dot{t}(s) \geq 0$. Hence

(4.4) It is possible to extract from $\hat{H}^{-1}(0)$ a piecewise smooth curve $\tilde{c}(a) \subset \hat{H}^{-1}(0)$ such that

- (i) $(0, a, 0) \in \tilde{c}(a)$
- (ii) $\tilde{c}(a)$ is monotone in t .

However, (4.4) merely guarantees that

$$\det \begin{pmatrix} H_x^r & -H_x^i \\ -H_x^i & -H_x^r \end{pmatrix}$$

doesn't change sign on $\tilde{c}(a)$ and not that $\det H_x^r$ doesn't change sign on $\tilde{c}(a)$. We give an example below which illustrates this point (see (6.7)).

From (4.1) we have $\hat{H}(t, x, y) = 0$ if and only if $\hat{H}(t, x, -y) = 0$. Hence $\hat{H}^{-1}(0)$ is symmetric about $\mathbb{R}^1 \times \mathbb{R}^N \times \{0\}^N$. Since we are mainly interested in obtaining real critical points for which $\tilde{c}(a)$ should preserve the Morse index, it is only crucial to test whether the Morse index can be preserved on one of the branches emanating from a bifurcation point $(t(s^*), x(s^*), 0)$ for which $y(s^* - \epsilon) \neq 0$ on some interval $\epsilon \in (0, \epsilon_0)$, $\epsilon_0 > 0$.

5. Convergence

In view of (4.4), a piecewise smooth monotone path $\tilde{c}(a) \subset \hat{H}^{-1}(0)$ exists. Since H was assumed to be real analytic for each t , so is \hat{H} . Hence by (2.11),

(5.1) The set of singular points of $\tilde{c}(a)$ has no finite accumulation point.

However, to ensure that $\tilde{c}(a)$ reaches the level $t = 1$, it is necessary to make some further requirement. An example of such a requirement is a

(5.2) Bounding condition:

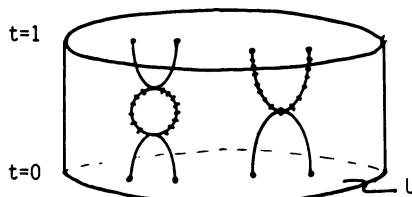
Let $U \subset \mathbb{R}^{2N}$ be an open bounded set such that

- (i) $(a, 0) \in U$
- (ii) $\hat{H}^{-1}(0) \cap ([0, 1] \times \partial U) = \emptyset$.

Then there exists a piecewise smooth monotone curve $\tilde{c}(a) \subset \hat{H}^{-1}(0)$ such that $\tilde{c}(a) \cap (\{1\} \times U) \neq \emptyset$.

(5.3) Conversely, for every $(x, y) \in U$ such that $\hat{H}(1, x, y) = 0$, there exists a piecewise smooth monotone path $\tilde{c}(a) \subset \hat{H}^{-1}(0)$ such that $(1, x, y) \in \tilde{c}(a)$ and $(0, a, 0) \in \tilde{c}(a)$.

Figure (5.4) illustrates the bounding condition.



— corresponds to a real curve: $y = 0$

↔ corresponds to a complex curve: $y \neq 0$.

For polynomial mappings in which the degrees of the components are properly matched, a bounding condition may be verified [4].

For simply concluding convergence, it often isn't necessary to take great care about matching the degrees. We illustrate this in the case of polynomial maps.

Suppose $F_j: \mathbb{R}^N \rightarrow \mathbb{R}^1$ is a polynomial map for each $j = 1, \dots, N$. Suppose $E_j: \mathbb{R}^N \rightarrow \mathbb{R}^1$ is chosen to be a polynomial map with

$$(5.5) \quad \deg E_j \leq \deg F_j, \quad j = 1, \dots, N,$$

then by Bezout's theorem, for $t \neq 0$, the number of solutions to

$$H(t, x+iy) = (1-t)E(x+iy) + tF(x+iy) = 0$$

is constant. (In particular, it is the product of the degrees of the F_j .) Thus,

(5.6) The only hyperplane normal to the t -axis to which $H^{-1}(0)$ can be asymptotic is $t = 0$.

Thus if $\hat{c}(a)$ isn't monotone increasing, then it is asymptotic to $t = 0$ from above. However, then there exists a point at which $t = 0$ and $\hat{c}(a)$ has a singular point. By (5.1) there can only be finitely many singular points in $\hat{c}(a)$. Thus by (5.6) $\hat{c}(a)$ eventually reaches $t = 1$ if (5.5) holds.

We illustrate this in the examples of the following section.

6. Examples

Let $f(x) = -\frac{1}{4}x^4 + \frac{1}{2}x^2$. Then

$$F(x) = f'(x) = -x(x^2 - 1).$$

Thus the critical points of f are:

$$x = 0 \quad (\text{a local minimum})$$

$$x = \pm 1 \quad (\text{global maxima}).$$

Let $e(x) = \frac{1}{2}(x - \frac{16}{11})^2$. Then

$$E(x) = e'(x) = x - \frac{16}{11}.$$

Thus the critical point of e is

$$x = \frac{16}{11} \quad (\text{a global minimum}).$$

Let $H(t, x) = (1 - t)E(x) + tF(x)$

$$(6.1) \quad \begin{aligned} &= (1 - t)(x - \frac{16}{11}) - tx(x^2 - 1) \\ &= (x - \frac{16}{11}) - t(x^3 - \frac{16}{11}). \end{aligned}$$

Thus $H(t, x) = 0$ yields

$$(6.2) \quad t(x) = (x - \frac{16}{11}) / (x^3 - \frac{16}{11}) \quad \text{for } x \neq \sqrt[3]{\frac{16}{11}}.$$

The turning points are readily seen to be

- (6.3) at $x = 2$: $(t(x), x) = (\frac{1}{12}, 2)$,
at $x = (1 + 3\sqrt{5})/11 := \alpha_+ : (t(\alpha_+), \alpha_+) = (.6788248, .7007458)$
at $x = (1 - 3\sqrt{5})/11 := \alpha_- : (t(\alpha_-), \alpha_-) = (1.237842, -.5189276)$.

In addition, t is asymptotic to 0 from above as $x \rightarrow \pm\infty$ and

$$\begin{aligned} t \uparrow \infty \text{ as } x \uparrow \sqrt[3]{\frac{16}{11}} \\ t \downarrow -\infty \text{ as } x \downarrow -\sqrt[3]{\frac{16}{11}}. \end{aligned}$$

The complexification of $H(t, x)$ is given by

$$H(t, x + iy) = (1 - t)(x + iy - \frac{16}{11}) - t[(x + iy)^3 - (x + iy)].$$

Hence $H(t, x + iy) = 0$ leads to the pair of equations

- (6.4) (i) $x - \frac{16}{11} - t(x^3 - 3xy^2 - \frac{16}{11}) = 0$
(ii) $y(1 - t(3x^2 - y^2)) = 0.$

When $y = 0$, the real curve in (6.2) results. For $y \neq 0$, (6.4)(ii) yields

$$(6.5) \quad y^2 = 3x^2 - \frac{1}{t}.$$

Thus $y^2 > 0$ if and only if $t < 0$ or $t > \frac{1}{3x^2}$. Replacing y^2 from (6.5) in (6.4)(i) yields

$$t(x) = (x + \frac{8}{11})/(4x^3 + \frac{8}{11}) \text{ for } x \neq -\sqrt[3]{\frac{2}{11}} \text{ and}$$

$$y^2(x) = (-x^3 + \frac{24}{11}x^2 - \frac{8}{11})/(x + \frac{8}{11}) \text{ for } x \neq -\frac{8}{11}.$$

Thus the complex solution set given by (6.4) is

$$\left\{ \left(\frac{x + \frac{8}{11}}{4x^3 + \frac{8}{11}}, x, \pm \sqrt{\frac{-x^3 + \frac{24}{11}x^2 - \frac{8}{11}}{x + \frac{8}{11}}} \right) \right\} \text{ for } \alpha_+ \leq x \leq 2 \text{ or} \\ -\sqrt[3]{\frac{2}{11}} < x \leq \alpha_- \text{ or } -\frac{8}{11} < x < -\sqrt[3]{\frac{2}{11}}.$$

The complex solution curve is portrayed in Figure (6.6). From (6.3) we have that its bifurcation points are:

$$(t, x, y) = (\frac{1}{12}, 2, 0), (t(\alpha_+), \alpha_+, 0), (t(\alpha_-), \alpha_-, 0).$$

Its asymptotes are:

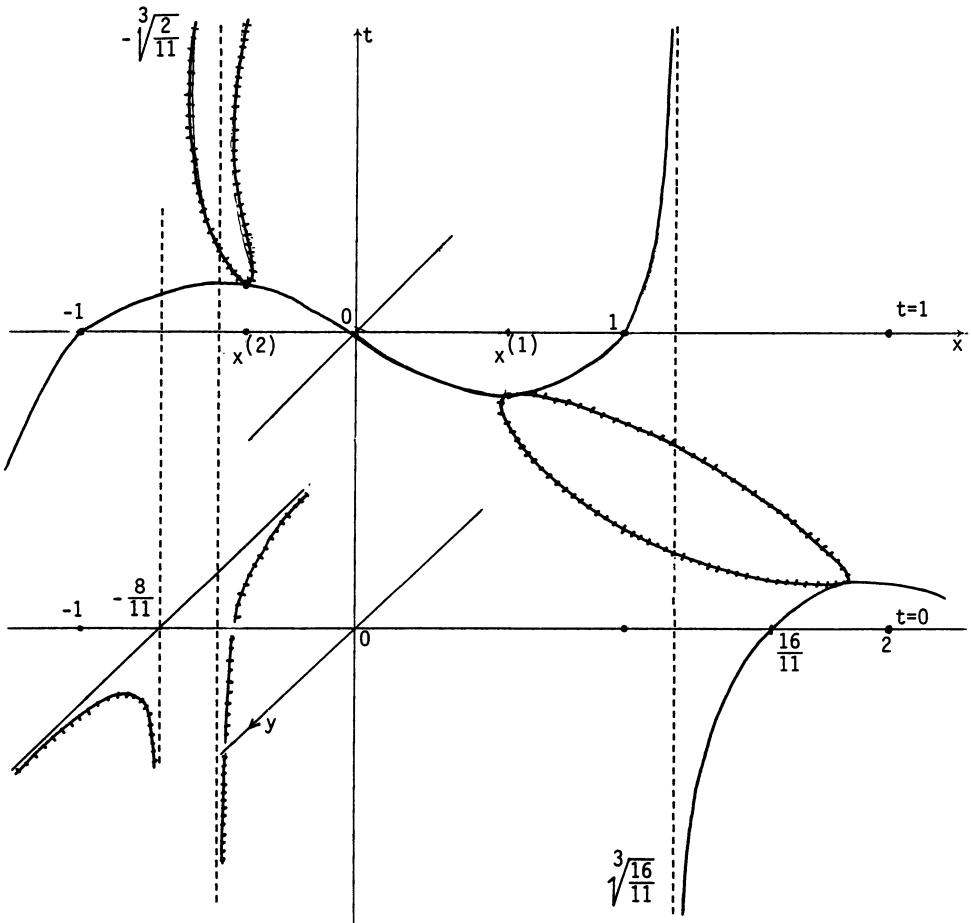
$$(t, -\sqrt[3]{\frac{2}{11}}, \pm\sqrt{3}\sqrt[3]{\frac{2}{11}}) \quad (\text{vertical})$$

$$(0, -\frac{8}{11}, \pm y) \quad (\text{horizontal}).$$

(6.6) Figure

----- = complex curve
 ——— = real curve.

In this example the complexification connects a minimum of e to a minimum of f .



(6.7) Example 2. This example is a two-dimensional extension of Example 1. Let

$$f(x_1, x_2) = -\frac{1}{2} \left[\frac{1}{2}x_1^4 - x_1^2 + \frac{1}{2}x_2^4 - x_2^2 \right], \quad e(x_1, x_2) = \frac{1}{2} \left[(x_1 - \frac{16}{11})^2 + (x_2 - 1)^2 \right].$$

$$\text{Then } \nabla f(x_1, x_2) = \begin{pmatrix} x_1(x_1^2 - 1) \\ x_2(x_2^2 - 1) \end{pmatrix} := F(x_1, x_2) \text{ and } \nabla e(x_1, x_2) = \begin{pmatrix} x_1 - \frac{16}{11} \\ x_2 - 1 \end{pmatrix} := E(x_1, x_2).$$

Hence f has the critical points

$$(x_1, x_2) = (0, 0) \quad (\text{a local minimum})$$

$$\begin{aligned}(x_1, x_2) &= (0, \pm 1), (\pm 1, 0) && \text{(saddle points)} \\ (x_1, x_2) &= (\pm 1, \pm 1) && \text{(global maxima).}\end{aligned}$$

\mathbf{e} has the single critical point

$$(x_1, x_2) = \left(\frac{16}{11}, 1\right) \quad \text{(the global minimum).}$$

Let $H(t, x_1, x_2) = (1-t)E(x_1, x_2) + tF(x_1, x_2)$. Then $H(t, x_1, x_2) = 0$ yields as in Example 1,

$$\begin{aligned}(6.8) \quad (i) \quad t &= (x_1 - \frac{16}{11})/(x_1^3 - \frac{16}{11}) \quad \text{and} \\ (ii) \quad \text{either} \quad (a) \quad x_2 &\equiv 1, \quad \text{or} \\ (b) \quad t(x_2^2 + x_2 + 1) &= 1.\end{aligned}$$

If (6.8)(ii)(a) holds, i.e., $x_2 \equiv 1$, the solution set is basically the same as in (6.2), except that $x_2 \equiv 1$

$$(6.9) \quad \left\{ \left(x_1 - \frac{16}{11} \right) / \left(x_1^3 - \frac{16}{11} \right), x_1, 1 \mid x_1 \neq \sqrt[3]{\frac{16}{11}} \right\}.$$

For the complex extension $H(t, x_1 + iy_1, x_2 + iy_2)$ of $H(t, x_1, x_2)$, the connected component $tK \subset H^{-1}(0)$ which contains the set in (6.9), is precisely the same as the solution set in Example 1, except that $x_2 \equiv 1$ and $y_2 \equiv 0$.

The real part of the complex extension is

$$H^r(t, x_1, x_2, y_1, y_2) = \begin{cases} \left(x_1 - \frac{16}{11} \right) - t(x_1^3 - 3x_1y_1^2 - \frac{16}{11}) \\ \left(x_2 - 1 \right) - t(x_2^3 - 3x_2y_2^2 - 1) \end{cases}.$$

Thus

$$(6.10) \quad H_x^r = \begin{pmatrix} 1 - t(3x_1^2 - 3y_1^2) & 0 \\ 0 & 1 - t(3x_2^2 - 3y_2^2) \end{pmatrix}.$$

Since on the complex part of K ,

$$\begin{aligned}y_1^2 &= 3x_1^2 - \frac{1}{t}, \quad x_2 \equiv 1, \quad y_2 \equiv 0, \\ (6.11) \quad H_x^r &= \begin{pmatrix} 6x_1^2 t - 2 & 0 \\ 0 & 1 - 3t \end{pmatrix}.\end{aligned}$$

Thus in traversing any piecewise smooth monotone curve $\tilde{c}(-\frac{16}{11}, 1) \subset \hat{H}^{-1}(0)$ we see from (6.10),

- (6.12) (a) both eigenvalues of H_x^r are positive at $t = 0$,
 (b) one eigenvalue of H_x^r changes sign at $t = \frac{1}{12}$,

from (6.11),

- (c) for $\frac{1}{3} < t < t(\alpha_+)$ (on K) both eigenvalues of H_x^r are negative,

(d) at $t = 1$, at least one eigenvalue of H_x^r is negative.

In particular,

(6.13) No piecewise smooth monotone curve $\tilde{c}(-\frac{16}{11}, 1) \subset H^{-1}(0)$ preserves the Morse index of the critical points it connects.

The remainder of the solution set $H^{-1}(0)$ is also somewhat interesting (at least for $t > 0$). It is schematically portrayed in Figure (6.15). It consists of a single connected component which satisfies the following equations:

$$(6.14) \quad (i) \quad t = (x_1 - \frac{16}{11})/(x_1^3 - 3x_1y_1^2 - \frac{16}{11})$$

$$(ii) \quad t = (x_2 - 1)/(x_2^3 - 3x_2y_2^2 - 1), \quad x_2 \neq 1$$

$$(iii) \quad y_1[1 - t(3x_1^2 - y_1^2)] = 0$$

$$(iv) \quad y_2[1 - t(3x_2^2 - y_2^2)] = 0.$$

The solution set to (6.14) for $t > 0$ has the following bifurcation points and asymptotes.

$$\text{Bifurcations: } (t, x_1, x_2, y_1, y_2) = (\frac{1}{12}, 2, \frac{-1 \pm \sqrt{45}}{2}, 0, 0)$$

$$(t(\alpha_+), \alpha_+, \frac{1}{2}(-1 \pm \sqrt{\frac{4}{t(\alpha_+) - 3}}), 0, 0)$$

$$(t(\alpha_-), \alpha_-, \frac{1}{2}(-1 \pm \sqrt{\frac{4}{t(\alpha_-) - 3}}), 0, 0)$$

$$(\frac{4}{3}, \beta, -\frac{1}{2}, 0, 0)$$

where $\beta = 1.0474549$ is the unique real solution to the cubic equation obtained from

$$(x_1 - \frac{16}{11})/(x_1^3 - \frac{16}{11}) = \frac{4}{3}.$$

$(\frac{4}{3}, \gamma, -\frac{1}{2}, \pm \sqrt{3\gamma^2 - \frac{3}{4}}, 0)$ where $\gamma = -.52372747$ is the unique real solution to the cubic equation obtained from

$$\frac{4}{3} = (x_1 - \frac{16}{11})/[x_1^3 - 3x_1(3x_1^2 - \frac{3}{4}) - \frac{16}{11}].$$

Note that the last two bifurcation points are in the complex domain.

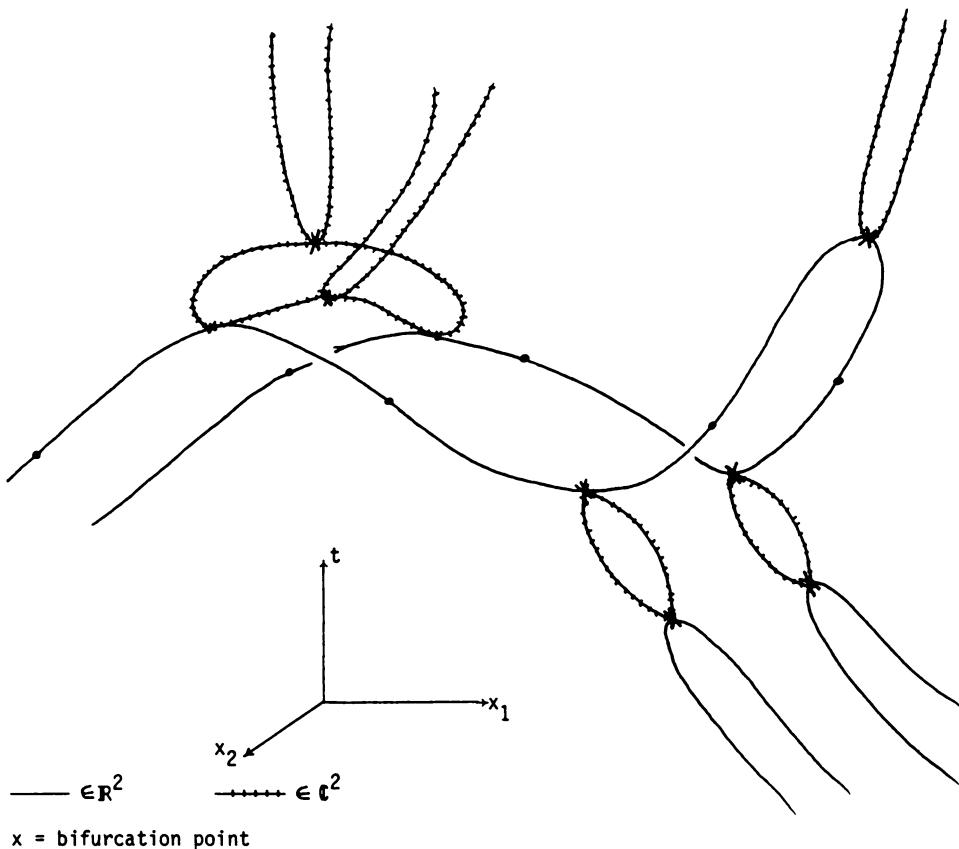
Asymptotes: $(0, x_1, -\frac{1}{2} \pm x_1, 0, 0)$ as $x_1 \rightarrow \pm\infty$ (horizontal)

$(0, \frac{16}{11}, x_2, 0, 0)$ as $x_2 \rightarrow \pm\infty$ (horizontal)

$(t, \sqrt[3]{\frac{16}{11}}, -\frac{1}{2}, 0, \frac{\pm\sqrt{3}}{2})$ as $t \rightarrow \infty$ (vertical)

$(t, -\sqrt[3]{\frac{16}{11}}, -\frac{1}{2}, \pm\sqrt{3}\sqrt[3]{\frac{2}{11}}, \frac{\pm\sqrt{3}}{2})$ as $t \rightarrow \infty$ (vertical).

(6.15) Figure



7. Numerical Aspects

The numerical tracing of $\tilde{c}(a)$ can be accomplished by a typical pseudo-arc length algorithm. The branch switching may be accomplished by using the orthogonal directions given in (4.3) when a point with $t < 0$ has been encountered. It is also possible to trace $\tilde{c}(a)$ via a derivative-free update continuation algorithm which effects branching off via perturbations. For discussions of this approach, see [6], [7]. Both of the above-mentioned approaches have been used on several examples of the sort which were used as illustrations in Section 6.

For the application at hand, it is important to test whether the eigenvalue structure of $H_x^r(0, a, 0)$ and $H_x^r(t(s^* + \epsilon), x(s^* + \epsilon), 0)$ are the same when s^* is a point at which $H_x^{-1}(0)$ has a bifurcation and $y(s^* - \epsilon) \neq 0$ for $\epsilon \in [0, \epsilon_0]$, $\epsilon_0 > 0$.

In the case of testing for local minima, this can be effected rather cheaply by using a selective iteration process, such as those discussed by Scheurle [8][9].

Let us suppose that $(t(s^*), x(s^*), 0)$ is a bifurcation point of $H^{-1}(0)$ with $y(s^* - \epsilon) \neq 0$ for $(t(s^* - \epsilon), x(s^* - \epsilon), y(s^* - \epsilon)) \in H^{-1}(0)$ for $\epsilon \in [0, \epsilon_0]$. Suppose further that

$$(7.1) \quad (t_1, x_1, 0) \approx (t(s^* + \epsilon_1), x(s^* + \epsilon_1), 0)$$

is a point which has been numerically obtained as an approximation to a point of $H^{-1}(0)$ with $t_1 = t(s^*) + h_1$, $h_1 > 0$. There are, of course, in general two possible points of the type (7.1), and possibly both points will need to be considered.

Now we may perform the iterative process

$$(7.2) \quad x_{n+1} = x_n - H^r(t^* + h_1, x_n)/M \|H_x^r(t^* + h_1, x_1)\|;$$

$n = 1, 2, \dots$, where M is chosen so that $M > 1$.

It is now easily seen that (7.2) will be repelling if $H_x^r(t(s^* + \epsilon_1), x(s^* + \epsilon_1))$ has a negative eigenvalue and converging otherwise.

References

- [1] Alexander, J.C. and Yorke, J.A.: The homotopy continuation method: Numerically implementable topological procedures, Trans.Amer.Math.Soc. 242 (1978), 271-284.
- [2] Allgower, E.L. and Georg, K.: Simplicial and continuation methods for approximating fixed points and solutions to systems of equations, SIAM Review 22 (1980), 28-85.
- [3] Chow, S.N., Mallet-Paret, J. and Yorke, J.A.: Homotopy methods that are constructive with probability one, Math.Comp. 32 (1978), 887-899.
- [4] Chow, S.N., Mallet-Paret, J. and Yorke, J.A.: A homotopy method for locating all zeros of a system of polynomials. In: Functional-Differential Equations and Approximation of Fixed Points, H.-O. Peitgen and H.-O. Walther (eds.) Springer Lecture Notes in Mathematics, 730 (1979), 77-88.
- [5] Crandall, M.G. and Rabinowitz, P.H.: Bifurcation from simple eigenvalues, J. Func.Anal. 8 (1971), 321-340.
- [6] Georg, K.: On tracing an implicitly defined curve by quasi-Newton steps and calculating bifurcation by local perturbation, SIAM J.Sci.Stat.Comp. 2 (1981), 35-50.
- [7] Georg, K.: Zur numerischen Realisierung von Kontinuitätsmethoden mit Prädiktor-Korrektor oder simplizialen Verfahren, Habilitationschrift, Univ. Bonn, 1982.
- [8] Scheurle, J.: Selective iteration and applications, J.Math.Anal.Appl. 59 (1977), 596-616.
- [9] Scheurle, J.: Ein selektives Projektions-Iterationsverfahren und Anwendungen auf Verzweigungsprobleme, Numer.Math. 29 (1977), 11-35.
- [10] Zangwill, W.I.: Determining all minima of certain functions, preprint Univ. of Chicago, August, 1981.

APPROXIMATION OF HOPF BIFURCATION
FOR SEMILINEAR PARABOLIC EQUATIONS

by

Christine BERNARDI and Jacques RAPPAZ

Summary : For a system of semilinear parabolic equations which presents a Hopf bifurcation, we define a general discrete problem which retains the bifurcation property and state an error estimate between the exact and approximate periodic solutions.

I. Introduction

Let us consider a semilinear elliptic problem

$$(1) \quad F(\lambda, u) = 0$$

depending on a real parameter λ and assume some hypotheses, in order to ensure the existence of a simple Hopf bifurcation for this problem, so that a family of periodic solutions of the parabolic equation

$$(2) \quad \frac{du}{dt} + F(\lambda, u) = 0$$

appears at a solution of (1). The analysis of equation (2) made in section II allows us to parametrize these periodic solutions. In section III, we define a discrete problem which retains the bifurcation property and give error estimates between the exact and approximate solutions. Examples of approximation are given in section IV.

Let V and H be two real Hilbert spaces, so that V is contained in H with a continuous imbedding and is dense in H . We denote by (\cdot, \cdot) the scalar product of H as well as the duality pairing between V and its dual space V' . We also consider another Banach space W satisfying : $V \hookrightarrow W \hookrightarrow H$ such that the imbedding of V into W is compact.

We also need the duality pairing between the spaces $L^2(0, 2\pi; V)$ and $L^2(0, 2\pi; V')$, defined by

$$[f, v] = \frac{1}{2\pi} \int_0^{2\pi} (f(s), v(s)) ds .$$

Let \mathfrak{X} be the closure of the space of 2π -periodic functions of $\mathcal{D}([0,2\pi];V)$ in $L^2(0,2\pi;V) \cap H^{1/2}(0,2\pi;H)$; we still denote by $[\cdot, \cdot]$ the duality pairing between \mathfrak{X} and its dual space \mathfrak{X}' .

Henceforward, we assume that the function F is of the form

$$F(\lambda, v) = A v + G(\lambda, v)$$

where :

1° the operator A is an isomorphism from V onto V' and satisfies

$$(3) \quad \forall v \in V, \quad (A v, v) \geq \alpha \|v\|_V^2, \quad \alpha > 0;$$

2° G is a \mathcal{C}^p -mapping ($p \geq 4$) from $\mathbb{R} \times V$ into V' and from $\mathbb{R} \times \mathfrak{X}$ into \mathfrak{X}' , with the derivative $D^p G$ bounded on the bounded sets of $\mathbb{R} \times \mathfrak{X}$; moreover, the partial derivative $D_v G(\lambda, 0)$ can be extended into a continuous operator from V into V' .

Only for the sake of simplicity, we assume that

$$(4) \quad \forall \lambda \in \mathbb{R}, \quad G(\lambda, 0) = 0,$$

so that the trivial branch is a branch of solutions of (1) and (2).

We make the hypotheses which ensure the existence of a Hopf bifurcation at a point $(\lambda_0, 0)$ of $\mathbb{R} \times V$.

(H1) The spectrum of the operator $L = D_v F(\lambda_0, 0)$ for the injection I from V into V' contains two purely imaginary I -eigenvalues $\pm i\omega_0$, where ω_0 is a real positive number. The rest of the spectrum is disjoint from $i\omega_0 \mathbb{Z}$.

(H2) The eigenvalues $\pm i\omega_0$ are algebraically simple, i.e. the null space and the range of $L - i\omega_0 I$ in the natural complexified spaces V^C and V'^C satisfy

$$(5) \quad \dim N(L - i\omega_0 I) = \text{codim } R(L - i\omega_0 I) = 1,$$

where $N(L - i\omega_0 I)$ is spanned by a vector φ_0 of V^C which does not belong to $R(L - i\omega_0 I)$.

Let us denote by φ_0^* the vector of V^C such that

$$(6) \quad R(L - i\omega_0 I) = \{f \in V^C ; (f, \varphi_0^*) = 0\}$$

with $(\varphi_0^*, \varphi_0) = 1$.

(H3) The so-called Hopf condition can be written

$$(7) \quad \operatorname{Re}(D_{\lambda v}^2 F^0 \cdot \varphi_0, \varphi_0^*) \neq 0,$$

where $D_{\lambda v}^2 F^0 = D_{\lambda v}^2 F(\lambda_0, 0)$.

II. Analysis of the continuous problem.

First of all, we make the change of time scale : $s = \omega t$ so as to introduce ω as a new variable and look for 2π -periodic solutions. Our aim is to solve the equation

$$(8) \quad \mathfrak{F}(\lambda, \omega, v) = \omega \frac{dv}{ds} + F(\lambda, v) = 0$$

in a neighbourhood of $(\lambda_0, \omega_0, 0)$ in $\mathbb{R} \times \mathbb{R} \times \mathbf{x}$.

To study the linearized problem of (8), we define the operator \mathfrak{L} from \mathbf{x} into \mathbf{x}' by

$$(9) \quad \mathfrak{L}v = \omega_0 \frac{dv}{ds} + L v.$$

We set

$$(10) \quad \zeta_0 = \varphi_0 e^{-is}, \quad \zeta_0^* = \varphi_0^* e^{+is}$$

and introduce the spaces

$$(11) \quad \mathbf{x}'^1 = \{f \in \mathbf{x}' ; [f, \zeta_0^*] = 0\}, \quad \mathbf{x}^1 = \mathbf{x} \cap \mathbf{x}'^1.$$

The following result is proved in [1].

Proposition 1 : The operator \mathfrak{L} is an isomorphism from \mathbf{x}^1 onto \mathbf{x}'^1 .

Clearly, for any (λ, ω, v) solution of (8) and for any τ , $(\lambda, \omega, v(\cdot + \tau))$ is another solution of (8). So, we must anchor the solution in time ; noticing that

$$[\zeta_0^*, v(\cdot + \tau)] = e^{-i\tau} [\zeta_0^*, v]$$

we add to equation (8) the following

$$(12) \quad [\operatorname{Im} \zeta_0^*, v] = 0$$

(Let us remark that any element of \mathbf{x}' that does not annihilate ζ_0 and $\bar{\zeta}_0$ can be chosen instead of $\operatorname{Im} \zeta_0^*$, which is important for practical purposes).

Hence, we have to solve in a neighbourhood of $(0,0,0)$

$$(13) \quad H(\sigma, x, v) = \begin{bmatrix} [\operatorname{Im} \zeta_0^*, v] \\ \mathcal{F}(\lambda_0 + \sigma, \omega_0 + x, v) \end{bmatrix} = 0 ,$$

where H is defined from $\mathbb{R} \times \mathbb{R} \times \mathbf{x}$ into $\mathbb{R} \times \mathbf{x}'$. We notice that

$$H(0) = 0 \quad D H(0) : (\sigma, x, v) = \begin{bmatrix} [\operatorname{Im} \zeta_0^*, v] \\ \mathcal{F} v \end{bmatrix}$$

According to [5, appendix I], $D H(0)$ is a Fredholm operator with index 1; its null space and range are

$$N(D H(0)) = \mathbb{R} \times \mathbb{R} \times \operatorname{Span} \{\operatorname{Re} \zeta_0\}$$

$$R(D H(0)) = \mathbb{R} \times \{f \in \mathbf{x}' ; [f, \zeta_0^*] = [f, \bar{\zeta}_0^*] = 0\} .$$

Then, the Lyapunov-Schmidt reduction is an easy consequence of proposition 1 and the implicit function theorem. Let Q denote the orthogonal projector from \mathbf{x}' onto \mathbf{x}'^1 . There exists a unique \mathbf{C}^P -mapping : $(\varepsilon, \sigma, x) \mapsto w(\varepsilon, \sigma, x)$ from a neighbourhood of $(0,0,0)$ in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ into \mathbf{x}^1 such that

$$(14) \quad \begin{cases} Q \mathcal{F}(\lambda_0 + \sigma, \omega_0 + x, \varepsilon \operatorname{Re} \zeta_0 + w(\varepsilon, \sigma, x)) = 0 \\ w(0,0,0) = 0 . \end{cases}$$

(for more details, see [1]).

Now, we can write the bifurcation equation

$$(15) \quad f(\varepsilon, \sigma, x) = [\mathcal{F}(\lambda_0 + \sigma, \omega_0 + x, \varepsilon \operatorname{Re} \zeta_0 + w(\varepsilon, \sigma, x)), \zeta_0^*] = 0 .$$

Let us notice that : $f(0) = 0$, $D f(0) = 0$ and that all the second partial derivatives of f in 0 are equal to 0, except

$$D_{\varepsilon\sigma}^2 f(0) = D_{\sigma\varepsilon}^2 f(0) = \frac{1}{2} (D_{\lambda\nu}^2 F^0 \cdot \varphi_0, \varphi_0^*)$$

$$D_{\varepsilon\chi}^2 f(0) = D_{\chi\varepsilon}^2 f(0) = -\frac{1}{2} i.$$

Then, following [6], we look for the non-degenerate characteristic rays of equation (15), i.e. the rays $(\varepsilon_0, \sigma_0, \chi_0)$ in \mathbb{R}^3 such that :

$$1^\circ D^2 f(\varepsilon_0, \sigma_0, \chi_0) = 0;$$

$$2^\circ D^2 f(\varepsilon_0, \sigma_0, \chi_0)(\varepsilon, \sigma, \chi) = 0 \text{ implies that } (\varepsilon, \sigma, \chi) \text{ is parallel to } (\varepsilon_0, \sigma_0, \chi_0).$$

Clearly, by (H3), the only non-degenerate characteristic ray is given by : $\varepsilon_0 = 1, \sigma_0 = 0, \chi_0 = 0$. According to [6, theorem 4.2], there exists a unique \mathbb{P}^2 -mapping : $\alpha \mapsto (\varepsilon(\alpha), \sigma(\alpha), \chi(\alpha))$ from a neighbourhood of 0 in \mathbb{R} into $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ such that

$$(16) \quad \begin{cases} f(\alpha \varepsilon(\alpha), \alpha \sigma(\alpha), \alpha \chi(\alpha)) = 0 \\ \varepsilon(0) = 1, \sigma(0) = 0, \chi(0) = 0. \end{cases}$$

Hence, we conclude that equation (8) has a unique branch of solutions $(\lambda(\alpha), \omega(\alpha), v(\alpha))$ in a neighbourhood of $(\lambda_0, \omega_0, 0)$, with

$$(17) \quad \begin{cases} \lambda(\alpha) = \lambda_0 + \alpha \sigma(\alpha) = \lambda_0 + O(\alpha^2) \\ \omega(\alpha) = \omega_0 + \alpha \chi(\alpha) = \omega_0 + O(\alpha^2) \\ v(\alpha) = \alpha \varepsilon(\alpha) \operatorname{Re} \zeta_0 + w(\alpha \varepsilon(\alpha), \alpha \sigma(\alpha), \alpha \chi(\alpha)) = \alpha \operatorname{Re} \zeta_0 + O(\alpha^2). \end{cases}$$

The corresponding branch of periodic solutions of equation (2) is

$(\lambda(\alpha), u(\alpha))$, with

$$(18) \quad u(\alpha)(t) = v(\alpha)(\omega(\alpha)t).$$

Of course, under more drastic non-degeneracy assumptions, one can obtain a more precise parametrization of the branch.

III. Definition and analysis of the discrete problem.

Let us notice that the parabolic operator \mathcal{A} defined by

$$(19) \quad \mathcal{A} v = \omega_0 \frac{dv}{ds} + A v$$

is an isomorphism from \mathbf{x} onto \mathbf{x}' . If τ denotes the inverse operator of ζ , equation (8) is equivalent to

$$(20) \quad \zeta f(\lambda, \omega, v) = v + \zeta \{(\omega - \omega_0) \frac{dv}{ds} + G(\lambda, v)\} = 0.$$

Then, a very natural way to approximate (8) or (20) is the following. Let h be a strictly positive real parameter tending to 0; we consider a family $(V_h)_h$ of finite-dimensional subspaces V_h of V and a family $(\zeta_h)_h$ of operators $\zeta_h : \mathbf{x}' \rightarrow \mathbf{x} \cap L^2(0, 2\pi; V_h)$ which satisfy the approximation hypothesis

$$(AH) \quad \forall f \in \mathbf{x}', \quad \lim_{h \rightarrow 0} \|(\zeta - \zeta_h)f\|_{\mathbf{x}} = 0.$$

Our discrete problem consists in solving the equation

$$(21) \quad v + \zeta_h \{(\omega - \omega_0) \frac{dv}{ds} + G(\lambda, v)\} = 0$$

or equivalently

$$(22) \quad \zeta_h(\lambda, \omega, v) = \zeta v + \zeta \zeta_h \{(\omega - \omega_0) \frac{dv}{ds} + G(\lambda, v)\} = 0$$

in a neighbourhood of $(\lambda_0, \omega_0, 0)$ in $\mathbb{R} \times \mathbb{R} \times \mathbf{x}$.

Now, the discrete problem (22) can be studied in the same way as equation (8). If we set

$$(23) \quad H_h(\sigma, \chi, v) = \begin{bmatrix} [\operatorname{Im} \zeta_0^*, v] \\ \zeta_h(\lambda_0 + \sigma, \omega_0 + \chi, v) \end{bmatrix}$$

we notice that

$$H_h(0) = 0 \quad DH_h(0) = DH(0) - \begin{bmatrix} 0 \\ \zeta (\zeta - \zeta_h) D_v G^0 \end{bmatrix};$$

we also have for all ℓ , $0 \leq \ell \leq p$,

$$\forall (\sigma, \chi, v) \in \mathbb{R} \times \mathbb{R} \times \mathbf{x}, \quad \forall (\xi_1, \dots, \xi_\ell) \in (\mathbb{R} \times \mathbb{R} \times \mathbf{x})^\ell,$$

$$\lim_{h \rightarrow 0} \|(D^\ell H(\sigma, \chi, v) - D^\ell H_h(\sigma, \chi, v)) \cdot (\xi_1, \dots, \xi_\ell)\|_{\mathbb{R} \times \mathbf{x}'} = 0.$$

The Lyapunov-Schmidt reduction is then a consequence of the discrete implicit function theorem of [4, theorem 1] or [6, theorem 2.2].

Proposition 2 : There exist a compact neighbourhood B_0 of $(0,0,0)$ in $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ and two constants $h_0 > 0$ and $a_0 > 0$ such that, for $h \leq h_0$, there is a unique \mathbb{C}^p -mapping : $(\varepsilon, \sigma, x) \mapsto w_h(\varepsilon, \sigma, x)$ from B_0 into \mathbb{C}^1 which satisfies

$$(24) \quad \begin{cases} Q \tilde{f}_h(\lambda_0 + \sigma, \omega_0 + x, \varepsilon \operatorname{Re} \zeta_0 + w_h(\varepsilon, \sigma, x)) = 0 \\ \|w - w_h(\varepsilon, \sigma, x)\|_{\infty} \leq a_0. \end{cases}$$

Now, we can write the discrete bifurcation equation

$$(25) \quad f_h(\varepsilon, \sigma, x) = [\tilde{f}_h(\lambda_0 + \sigma, \omega_0 + x, \varepsilon \operatorname{Re} \zeta_0 + w_h(\varepsilon, \sigma, x), \zeta_0^*)] = 0.$$

We notice that f_h and the first partial derivatives of f_h in 0 are equal to 0, except

$$D_\varepsilon f_h(0) = -[\mathfrak{A}(\mathfrak{T} - \mathfrak{T}_h) D_V G^0 (\operatorname{Re} \zeta_0 + D_\varepsilon w_h(0)), \zeta_0^*].$$

Hence, to compute the discrete bifurcation point, we need once more the discrete implicit function theorem of [4].

Proposition 3 : There exist two constants $h_1 > 0$ and $a_1 > 0$ such that, for $h \leq h_1$, there is a unique pair (σ_h^0, x_h^0) in $\mathbb{R} \times \mathbb{R}$ which satisfies

$$(26) \quad \begin{cases} D_\varepsilon f_h(0, \sigma_h^0, x_h^0) = 0 \\ |\sigma_h^0| + |x_h^0| \leq a_1. \end{cases}$$

This result enables us to write the modified discrete bifurcation equation

$$(27) \quad \begin{cases} \tilde{f}_h(\varepsilon, \sigma, x) = f_h(\varepsilon, \sigma_h^0 + \sigma, x_h^0 + x) = \\ = [\tilde{f}_h(\lambda_0 + \sigma_h^0 + \sigma, \omega_0 + x_h^0 + x, \varepsilon \operatorname{Re} \zeta_0 + w_h(\varepsilon, \sigma_h^0 + \sigma, x_h^0 + x), \zeta_0^*)] = 0. \end{cases}$$

We notice that : $f_h(0) = 0$, $D f_h(0) = 0$ and still apply [6, theorem 4.2].

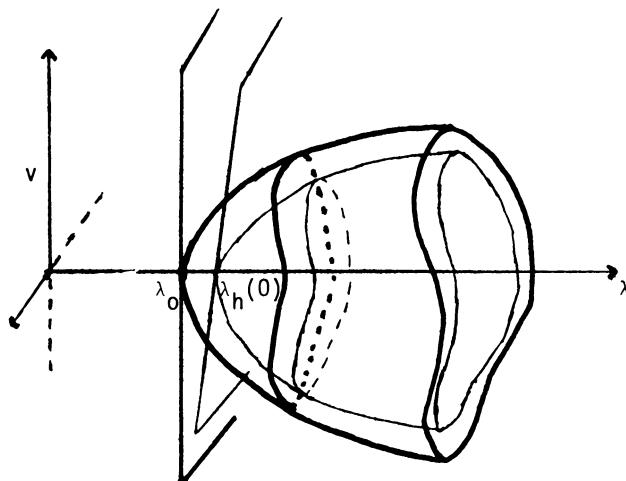
Proposition 4 : There exist three constants $\alpha^* > 0$, $h^* > 0$ and $a^* > 0$ such that, for $h \leq h^*$, there is a unique \mathbb{C}^{p-2} -mapping : $\alpha \mapsto (\varepsilon_h(\alpha), \sigma_h(\alpha), x_h(\alpha))$ from $[-\alpha^*, \alpha^*]$ into $\mathbb{R} \times \mathbb{R} \times \mathbb{R}$ which satisfies

$$(28) \quad \begin{cases} \tilde{f}_h(\alpha \varepsilon_h(\alpha), \alpha \sigma_h(\alpha), \alpha x_h(\alpha)) = 0 \\ |\varepsilon(\alpha) - \varepsilon_h(\alpha)| + |\sigma(\alpha) - \sigma_h(\alpha)| + |x(\alpha) - x_h(\alpha)| \leq \alpha^*. \end{cases}$$

Hence, we have proved that the discrete problem retains the bifurcation property. The previous results can be summarized in

Theorem 1 : For h sufficiently small, equation (21) has a unique branch of solutions $(\lambda_h(\alpha), \omega_h(\alpha), v_h(\alpha))$, $\alpha \in [-\alpha^*, \alpha^*]$, in a neighbourhood of $(\lambda_0, \omega_0, 0)$, with

$$(29) \quad \begin{cases} \lambda_h(\alpha) = \lambda_0 + \sigma_h^0 + \alpha \sigma_h(\alpha) \\ \omega_h(\alpha) = \omega_0 + x_h^0 + \alpha x_h(\alpha) \\ v_h(\alpha) = \alpha \varepsilon_h(\alpha) \operatorname{Re} \zeta_0 + w_h(\alpha \varepsilon_h(\alpha), \sigma_h^0 + \alpha \sigma_h(\alpha), x_h^0 + \alpha x_h(\alpha)). \end{cases}$$



Bifurcation diagram

— Continuous problem

— Discrete problem

Of course, the same discrete implicit function theorems yield error estimates (for technical details, see [1]).

Proposition 5 : There exists a constant C independent of h such that, for any k , $0 \leq k \leq p-3$, we have

$$(30) \quad \left\{ \begin{array}{l} \forall \alpha \in [-\alpha^*, \alpha^*] , \\ \left| \frac{d^k}{d\alpha^k} (\lambda(\alpha) - \lambda_h(\alpha)) \right| + \left| \frac{d^k}{d\alpha^k} (\omega(\alpha) - \omega_h(\alpha)) \right| + \\ \left\| \frac{d^k}{d\alpha^k} (v(\alpha) - v_h(\alpha)) \right\|_{\infty} \\ \leq C \sum_{\ell=0}^{k+1} \| (\tau - \tau_h)^{\frac{d^\ell}{d\alpha^\ell}} v(\alpha) \|_{\infty} \end{array} \right.$$

and, moreover,

$$(31) \quad \left\{ \begin{array}{l} |\lambda(0) - \lambda_h(0)| + |\omega(0) - \omega_h(0)| \\ \leq C \| [\Re(\tau - \tau_h) D_v G^0 \cdot (\operatorname{Re} \zeta_0 + w_h^1), \zeta_0^*] \| , \\ \text{with } \| w_h^1 \|_{\infty} \leq C \| (\tau - \tau_h) D_v G^0 \cdot \operatorname{Re} \zeta_0 \|_{\infty} . \end{array} \right.$$

Remarks : The error estimates we obtain depend only on the regularity of the periodic solutions of the continuous problem. We notice also that, by the approximation hypothesis (AH), for h sufficiently small, the discrete problem retains all the non-degeneracy properties of the continuous problem. Moreover, in most practical cases, inequality (31) yields an estimate of higher order than inequality (30) (an example is given in [1]).

IV. Examples of approximation.

In the sequel, we consider a bounded domain Ω of \mathbb{R}^m with Lipschitz-continuous boundary Γ and denote by $\|\cdot\|_{k,\Omega}$ the norm on the usual Sobolev spaces $H^k(\Omega)$ and $H^k(\Omega)^d$, where d is an integer ≥ 1 . We set : $V = H_0^1(\Omega)^d$ and $H = L^2(\Omega)^d$.

Let $a(\cdot, \cdot)$ be a symmetric continuous bilinear form on $H_0^1(\Omega)^d \times H_0^1(\Omega)^d$, which is coercive ; we define the operator A from $H_0^1(\Omega)^d$ into $H^{-1}(\Omega)^d$ by

$$(32) \quad \forall u \in H_0^1(\Omega)^d , \quad \forall v \in H_0^1(\Omega)^d , \quad (A u , v) = a(u, v) .$$

Example of Hopf bifurcation : For $\Omega =]0,1[$ and $d = 2$, we consider the classical problem

$$(33) \quad \begin{cases} \frac{\partial u_1}{\partial t} - \frac{\partial^2 u_1}{\partial x^2} - \lambda u_1 + \lambda u_2 + N(\lambda, u_1, u_2) = 0 & \text{in } (0,1) \\ \frac{\partial u_2}{\partial t} - \frac{\partial^2 u_2}{\partial x^2} - \lambda u_1 - \lambda u_2 + P(\lambda, u_1, u_2) = 0 & \text{in } (0,1) \\ u_1(0) = u_1(1) = 0 \quad u_2(0) = u_2(1) = 0 \end{cases},$$

where N and P are sufficiently regular mappings from $\mathbb{R} \times V$ into $H^{-1}(\Omega)$ which satisfy

$$(34) \quad \begin{cases} \forall \lambda \in \mathbb{R}, \quad \forall (u_1, u_2) \in V, \\ \|N(\lambda, u_1, u_2)\|_{-1, \Omega} + \|P(\lambda, u_1, u_2)\|_{-1, \Omega} \\ \leq C \{\|u_1\|_{1, \Omega}^2 + \|u_2\|_{1, \Omega}^2\} \end{cases}$$

One can check easily that hypotheses (H1) to (H3) are satisfied for $\lambda_0 = \pi^2$, so that a Hopf bifurcation takes place at the stationary solution $(\pi^2, 0)$.

Other examples from the real world can be found in [7] [8] for instance.

As Δ is the parabolic operator $\omega_0 \frac{d}{ds} + A$ from \mathbf{x} onto \mathbf{x}' , the inverse operator $\mathcal{T} = \Delta^{-1}$ is defined in the following way : for any f in \mathbf{x}' , $u = \mathcal{T}f$ is the only solution in \mathbf{x} of the equation :

$$(35) \quad \forall v \in V, \quad \omega_0 \left(\frac{du}{ds}, v \right) + a(u, v) = (f, v) \quad \text{a.e. in } (0, 2\pi).$$

Our aim is to define approximations of \mathcal{T} .

Let $(V_h)_h$ be a family of finite-dimensional subspaces V_h of V , and $(\pi_h)_h$ be a family of operators $\pi_h : V \rightarrow V_h$ such that, for an integer $r \geq 1$,

$$(36) \quad \begin{cases} \forall q \in \{1, \dots, r\}, \quad \forall v \in H^{q+1}(\Omega)^d, \\ \|v - \pi_h v\|_{i, \Omega} \leq C h^q \|v\|_{q+i, \Omega}, \quad i = 0 \text{ and } 1. \end{cases}$$

(Hypothesis (36) is satisfied if, for instance, the V_h are classical finite element spaces).

Then, the Galerkin approximation of u is the only solution u_h in $\mathbf{X} \cap L^2(0, 2\pi; V_h)$ of the equation

$$(37) \quad \forall v_h \in V_h, \quad \omega_0 \left(\frac{du_h}{ds}, v_h \right) + a(u_h, v_h) = (f, v_h) \quad \text{a.e. in } (0, 2\pi).$$

Let N_h be an integer ≥ 1 . The Galerkin-spectral approximation $\tilde{\mathbf{T}}_h f$ of u is defined by

$$(38) \quad \tilde{\mathbf{T}}_h f = \sum_{|n| \leq N_h} u^n e^{ins},$$

where the u^n , $|n| \leq N_h$, are solutions in V_h of

$$(39) \quad \forall v_h \in V_h, \quad i n \omega_0(u^n, v_h) + a(u^n, v_h) = (f^n, v_h),$$

and the f^n , $n \in \mathbb{Z}$, are the Fourier coefficients of f .

The following result is proved in [2].

Proposition 6 : If the solution $u = \mathbf{T} f$ of (35) is a periodic function in $\mathbf{X}(s) = L^2(0, 2\pi; H^{s+1}(\Omega)^d) \cap H^{(s+1)/2}(0, 2\pi; L^2(\Omega)^d)$, we have

$$(40) \quad \|(\mathbf{T} - \tilde{\mathbf{T}}_h)f\|_{\mathbf{X}} \leq C \{h^{\inf\{r, s\}} + N_h^{-s/2}\} \|u\|_{\mathbf{X}(s)}.$$

Also for an integer $N_h \geq 1$, we set : $k = \frac{2\pi}{N_h}$ and define the continuous functions φ^n , $0 \leq n \leq N_h - 1$, linear on each $[qk, (q+1)k]$, $0 \leq q \leq N_h - 1$, and such that

$$(41) \quad \begin{cases} \varphi^0(0) = \varphi^0(2\pi) = 1 & \varphi^0(qk) = 0, \quad 1 \leq q \leq N_h - 1 \\ \varphi^n(qk) = \delta_{nq} & 0 \leq q \leq N_h, \quad 1 \leq n \leq N_h - 1 \end{cases}$$

where δ_{nk} denotes the Kronecker's symbol.

The Galerkin-Euler approximation $\tilde{\mathbf{T}}_h f$ of u is defined by

$$(42) \quad \tilde{\mathbf{T}}_h f = \sum_{n=0}^{N_h-1} \tilde{u}^n \varphi^n,$$

where the \tilde{u}^n , $0 \leq n \leq N_h$, are solutions in V_h of

$$(43) \quad \begin{cases} \forall v_h \in V_h, \omega_0 \left(\frac{\tilde{u}^{n+1} - \tilde{u}^n}{k}, v_h \right) + a(\tilde{u}^{n+1}, v_h) = (\tilde{f}^n, v_h) \\ \tilde{u}^0 = \tilde{u}^N \end{cases}$$

and the \tilde{f}^n , $0 \leq n \leq N_h - 1$, are defined by : $\forall v \in V$, $(\tilde{f}^n, v) = \frac{1}{k} [f, \varphi^n v]$.
 (one can easily show that (43) has a unique solution $(\tilde{u}_0, \dots, \tilde{u}_{N_h-1})$ in $V_h^{N_h}$).

Using the same methods as in [3, theorem II.6], we prove the

Proposition 7 : If the solution $u = \mathcal{T}f$ of (35) is a periodic function in $\tilde{x}(s) = L^2(0, 2\pi; H^{s+1}(\Omega)^d) \cap H^{1/2}(0, 2\pi; H^s(\Omega)^d)$, $0 \leq s \leq r$, and in $\tilde{x} = H^1(0, 2\pi; H^1(\Omega)^d) \cap H^{3/2}(0, 2\pi; L^2(\Omega)^d)$, we have

$$(44) \quad \|(\mathcal{T} - \mathcal{T}_h)f\|_{\tilde{x}} \leq C\{h^s \|u\|_{\tilde{x}(s)} + N_h^{-1} \|u\|_{\tilde{x}}\}.$$

In both examples of approximation, one can prove by a classical density argument that hypothesis (AH) is satisfied if N_h tends to $+\infty$ as h tends to 0. Moreover, if the periodic solutions of the continuous problem are regular enough, one deduces from (40) and (44) error estimates between the exact and approximate periodic solutions in $O(h^s + N_h^{-s/2})$ and in $O(h^s + N_h^{-1})$ respectively.

REFERENCES.

- [1] C. BERNARDI. Approximation of Hopf bifurcation. Numer. Math. 39, 15-37 (1982).
- [2] C. BERNARDI. Numerical approximation of a periodic linear parabolic problem. SIAM J. Numer. Anal. 19, 1196-1207 (1982).
- [3] C. BERNARDI, G. RAUGEL. Approximation numérique de certaines équations paraboliques non linéaires. R.A.I.R.O. Anal. Numér. (to appear).

- [4] F. BREZZI, J. RAPPAZ, P.-A. RAVIART. Finite-dimensional approximation of nonlinear problems. Part I : branches of nonsingular solutions. *Numer. Math.* 36, 1-25 (1980).
- [5] J. DESCLOUX, J. RAPPAZ. On numerical approximation of solution branches of nonlinear equations. *Rapport Département de Mathématiques E.P.F.L.* (1981).
- [6] J. DESCLOUX, J. RAPPAZ. Approximation of solution branches of nonlinear equations, *R.A.I.R.O. Anal. Numér.* 16, 319-349 (1982).
- [7] R.F. HEINEMANN, A.B. POORE. Multiplicity, stability and oscillatory dynamics of the tubular reactor. *Chem. Eng. Science* 36, 1411-1419 (1981).
- [8] J.-P. KERNEVEZ. Enzyme Mathematics. Studies in Mathematics and its Applications 10, North-Holland (1980).

C. BERNARDI
Analyse Numérique
Université P. et M. Curie
4, place Jussieu
F - 75230 PARIS Cédex 05

J. RAPPAZ
Département de Mathématiques
Ecole Polytechnique Fédérale de Lausanne
Ecublens CH - 1015 LAUSANNE

DEFINING EQUATIONS FOR SINGULAR SOLUTIONS
AND NUMERICAL APPLICATIONS

Wolf-Jürgen Beyn

Singularity theory seems to play an important role not only in the theoretical but also in the numerical analysis of bifurcation problems. In this paper we establish a relation between the concept of a universal unfolding and direct methods for the numerical computation of singular points in bifurcation diagrams. In a direct method the unknown singular solution is computed as a regular solution of a so called defining equation. In particular, we discuss a defining equation for a multiple bifurcation point and demonstrate its application to a reaction diffusion system.

1. Introduction

The numerical computation of singular solutions in bifurcation problems has received special attention recently. We refer to [12] for a survey and comparison of numerical methods for turning points. There are by now also various approaches to more difficult singularities (see the papers of this volume), in particular to cusp points [16, 17] and bifurcation points [15, 13, 18, 19, 1, 9]. Basically, most of these methods consist in setting up a system of equations - we use the term *defining equations* - which has the unknown singularity as a regular solution. Newton's method for the defining equation would then converge locally and quadratically.

In this paper we present a list of defining equations and we want to demonstrate their relation to the concept of a universal unfolding in singularity theory. We consider a system of equations

$$(1) \quad T(z, c) = 0, \text{ where } T \in C^\infty(\mathbb{R}^M \times \mathbb{R}^p, \mathbb{R}^N), \quad M \geq N.$$

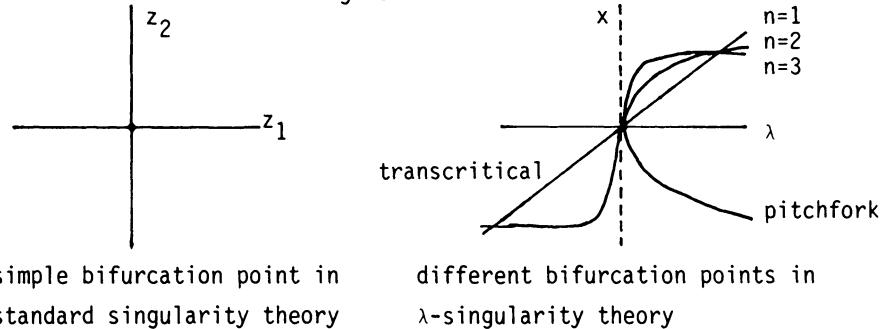
z and c are called the state and the control variable resp. We note that for some of the following definitions and results the operator T need only be defined locally and satisfy less smoothness assumptions. A solution (z_0, c_0) of equation (1) will be called *regular with respect to z* if the Jacobian of T w.r.t. z at (z_0, c_0) has maximum rank, i.e. $\text{rank } T_z(z_0, c_0) = N$. Otherwise it is called *singular with respect to z* .

Let (z_0, c_0) be a singular solution of (1) with respect to z . Then one of the fundamental results of singularity theory shows how to obtain in a qualitative way all possible solution sets of (1) in a neighbourhood of z_0 in the state space if c is kept fixed at a value close to c_0 . We refer to [6] for a brief account of some ideas and results of standard singularity theory as we use it. The harder parts of the proofs may be found e.g. in [5, 10]. It should be noted that standard singularity theory is distinct from classical catastrophe theory (which deals with gradient systems $T(z, c) = \nabla_z f(z, c) = 0$, $f \in C^\infty(\mathbb{R}^N \times \mathbb{R}^p, \mathbb{R})$, compare [5, 14]) and from the " λ -singularity theory" of Golubitsky and Schaeffer [8]. In the approach of [8] there are three types of variables: the state variable x , the bifurcation parameter λ and the control variable c . The (x, λ) -solution branches of a system

$$(2) \quad T(x, \lambda, c) = 0, \quad T \in C^\infty(\mathbb{R}^N \times \mathbb{R} \times \mathbb{R}^p, \mathbb{R}^N)$$

are then classified under the restriction that λ -slices are preserved. This turns out to be a refinement of the equivalence classes of singularities which are obtained in the standard theory by setting $z = (x, \lambda)$, $M = N + 1$. For example, the bifurcation points defined by $T_n(x, \lambda) = x(\lambda - x^n) = 0$ ($n \in \mathbb{N}$) are all different in λ -singularity theory, whereas they are equivalent to the simple bifurcation point $T(z_1, z_2) = z_1 z_2 = 0$ in the standard theory.

Figure 1



Our approach is to compute the bifurcation points of the types in fig. 1 by one and the same defining equation. However, if the particular behaviour of the emanating branches with respect to a bifurcation parameter λ is of interest then one should rather use the defining equations of Spence and Jepson (this volume) which are based on [8].

Our final remark concerns the question whether it is useful to compute more complex singularities at all. Usually, a dynamical system, which has (1) as

its steady state equation, only exhibits drastic changes near the most simple singularities of (1), namely folds (or turning points). As for the higher singularities we rather think of them to play the role of an "organizing center" [7] for the solution diagrams of the steady state equation (1). This aspect has been emphasized in [7 , 1 , 4].

I would like to thank Dipl. Math. J. Bigge for some of the numerical results and helpful discussions of the algebraic conditions for multiple bifurcation points.

2. Some fundamentals about singular solutions

For a singular solution (z_0, c_0) of (1) there are at least three important numbers to know:

the *index* $i = M - N$, the *corank* $n = N - \text{rank } T_z(z_0, c_0)$ and
the *codimension* $k \in \mathbb{N} \cup \{\infty\}$.

The codimension measures in some sense the complexity of the singularity and will be explained in detail later on (cf. [6 , V § 2]).

Let (z_0, c_0) be a regular solution of (1) w.r.t. z . Then the solution set of $T(z, c_0) = 0$ in a neighbourhood of z_0 is a smooth $i = M - N$ dimensional manifold. In general, this is no longer true near a singular solution with respect to z . However the solution set of (1) in a neighbourhood of (z_0, c_0) in the complete space $\mathbb{R}^M \times \mathbb{R}^P$ may be smooth again. For example, the equation $z^2 - c = 0$ has a fold w.r.t. z at $z = 0, c = 0$ (i.e. a singular solution of index 0) but defines a regular branch in $\mathbb{R} \times \mathbb{R}$. Similarly, $z_1^2 - z_2^2 - c = 0$ has a simple bifurcation point w.r.t. z at $(z_1, z_2) = 0, c = 0$ (i.e. a singular solution of index 1) but defines a regular surface in $\mathbb{R}^2 \times \mathbb{R}$. In what follows, the term "singular" will always mean "singular with respect to the state variable" and we will use the terms "point singularity" if index = 0 and "branch singularity" if index = 1.

The corank n of a singular solution (z_0, c_0) gives the number of equations to which the full system (1) can be reduced by the Liapunov-Schmidt method. The reduced equations are usually called the bifurcation equations. For numerical purposes it is important to note that this reduction may be performed without knowing the null space $N(T_z^0)$ or the range $R(T_z^0)$ (the upper index "0" always indicates the argument (z_0, c_0)).

We start with a decomposition into subspaces

$$(3) \quad \mathbb{R}^M = V \oplus W, \quad \mathbb{R}^N = X \oplus Y, \text{ where } \dim X = \dim V = N - n,$$

and we write $z = (v, w) \in \mathbb{R}^M$, $z_0 = (v_0, w_0)$. From (3) we obtain $\dim Y = n$, $\dim W = M - N + n = : m$. Let $P : \mathbb{R}^N \rightarrow X$ be the projector along Y . Our basic assumption is that for some open sets $\Omega \subset V$, $\Gamma \subset W \times \mathbb{R}^p$ the equation

$$(4) \quad PT(v, w, c) = 0, \quad v \in \Omega, \quad (w, c) \in \Gamma$$

defines a unique implicit function $v(w, c)$ such that

$$(5) \quad PT_V(v(w, c), w, c) : V \rightarrow X \text{ is nonsingular.}$$

For example, this is satisfied in some suitable neighbourhoods if

$(z_0, c_0) = (v_0, w_0, c_0)$ solves (4) and $PT_V^0 : V \rightarrow X$ is nonsingular. In this case $v(w_0, c_0) = v_0$ holds.

The mapping $S : \Omega \times \Gamma \rightarrow Y$, defined by

$$(6) \quad S(w, c) = (I - P)T(v(w, c), w, c)$$

will then be called a *Liapunov-Schmidt reduction of T w.r.t. (V, W, X, Y)* .

Theorem 1

Let $(z_0, c_0) = (v_0, w_0, c_0)$ be a solution of (4) where $T \in C^r(\mathbb{R}^M \times \mathbb{R}^p, \mathbb{R}^N)$, $1 \leq r \leq \infty$ and let $S(w, c)$ be a Liapunov-Schmidt reduction of T with respect to (V, W, X, Y) . Further, let $R : V \rightarrow X$ be linear and bijective. Then in a neighbourhood $U(z_0, c_0)$ a relation

$$(7) \quad \tau(z, c)T(\rho(z, c), c) = (R(v - v_0), S(w, c)), \quad z = (v, w)$$

holds where $\tau(z, c)$ is a C^{r-1} -family of regular $N \times N$ -matrices and $\rho \in C^r(U(z_0, c_0), U(z_0))$ satisfies $\rho(z_0, c_0) = z_0$, ρ_z^0 nonsingular. Moreover, if (z_0, c_0) is a singular solution of (1) w.r.t. z of corank n then $S(w_0, c_0) = 0$, $S_w(w_0, c_0) = 0$.

Remark: Formula (7) means that in a neighbourhood of a singular solution of corank n the operator T may be decomposed into a regular part R and a singular part S after a parameter dependent change of coordinates in \mathbb{R}^M and \mathbb{R}^N (cf. [8, Lemma 3.13]).

Proof: We drop the control variable c from the proof because it can simply be inserted at each step. By our assumptions the mapping $\sigma(z) = (v_0 + R^{-1}PT(z), w)$ satisfies $\sigma(z_0) = z_0$, $\sigma_z(z_0)$ nonsingular. Hence

$\sigma^{-1}(z) = : \rho(z) = (\rho_1(z), w) \in V \oplus W$ has the same property. By the definition of ρ we have

$$(8) \quad R(v-v_0) = PT(\rho(z)), \quad z = (v, w).$$

Setting $v = v_0$ in (8) we obtain $\rho_1(v_0, w) = v(w)$ from (4) and hence

$$(9) \quad S(w) = (I-P)T(\rho(v_0, w)).$$

Now let $\tau(z)$ be given by its representation $\begin{pmatrix} I & 0 \\ \tau_{21}(z) & I \end{pmatrix}$ with respect to $X \oplus Y$ where τ_{21} will be defined below.

Then (8) yields

$$\tau(z)T(\rho(z)) = (R(v-v_0), \tau_{21}(z)R(v-v_0) + (I-P)T(\rho(v, w))).$$

Expanding the last term to first order in v and using (9) we end up with the relation (7) if we set

$$\tau_{21}(z) = - \int_0^1 (I-P)T_z(\rho(v_0 + t(v-v_0), w))\rho_v(v_0 + t(v-v_0), w)R^{-1}dt.$$

Finally, $S(w_0, c_0) = 0, S_w(w_0, c_0) = 0$ are easy consequences of (7) if (z_0, c_0) is a singular solution of corank n . \square

The effect of theorem 1 is to concentrate the singular part of T into a low dimensional mapping $G(w) = S(w, c_0)$. In the following we need some basic notations from singularity theory [6, V]. Let $E_m^n = C^\infty(\mathbb{R}_0^m, \mathbb{R}^n)$ be the linear space of C^∞ -germs defined in a neighbourhood of $0 \in \mathbb{R}^m$ with values in \mathbb{R}^n . Similarly, $E_m^{n,n} = C^\infty(\mathbb{R}_0^m, \mathbb{R}^{n,n})$ contains the C^∞ -germs with values in the space $\mathbb{R}^{n,n}$ of $n \times n$ -matrices. Two germs $G_1, G_2 \in E_m^n$ are called *contact equivalent*, if a relation

$$(10) \quad G_1(w) = \tau(w)G_2(\rho(w))$$

holds for some $\tau \in E_m^{n,n}$, $\rho \in E_m^m$ with $\rho(0) = 0$ and $\rho_w(0), \tau(0)$ nonsingular. This essentially means that the solution sets of $G_1(w) = 0$ and $G_2(w) = 0$ near 0 are diffeomorphic. The result of theorem 1 may then be reformulated as the contact equivalence of the germs $T(z_0 + \cdot, c)$ and $R \otimes S(w_0 + \cdot, c)$.

A germ $F \in E_{m+1}^n$ is called an ℓ -parameter unfolding of $G \in E_m^n$ iff $G(w) = F(w, 0)$. It is called a *versal unfolding* of G if every j -parameter unfolding $H \in E_{m+j}^n$ of G satisfies $H(w, \beta) = \tau(w, \beta) F(\rho(w, \beta), \psi(\beta))$ for some $\tau \in E_{m+j}^{n,n}$, $\rho \in E_{m+j}^m$, $\psi \in E_j^1$ such that $\tau(w, 0) = I, \rho(w, 0) = w$ and $\psi(0) = 0$ (i.e. the germs $H(\cdot, \beta)$ and $F(\cdot, \psi(\beta))$ are contact equivalent). A versal unfolding of G with a minimum number of parameters is said to be *universal*.

For the determination of a universal unfolding of G one needs the so called *tangent space*

$$(11) \quad TG = \{\tau G + G_w \sigma : \tau \in E_m^{n,n}, \sigma \in E_m^m\} \subset E_m^n.$$

TG is a linear space over \mathbb{R} as well as a module over E_m^1 . The number $k = \text{codim } TG = \dim E_m^n / TG \in \mathbb{N} \cup \{\infty\}$ is the *codimension of the germ G at 0*. It is invariant under contact equivalence. Let us assume that G has finite codimension k and let $F(w, \alpha)$ be a k -parameter unfolding of G . Then the fundamental theorem on universal unfoldings [6, V § 3] states that F is universal if and only if the *transversality condition*

$$(12) \quad TG + \mathbb{R} \cdot \{F_{\alpha_1}(\cdot, 0), \dots, F_{\alpha_k}(\cdot, 0)\} = E_m^n$$

is satisfied. This condition means that the germs $F_{\alpha_j}(w, 0)$ ($j=1, \dots, k$) form a basis of E_m^n / TG .

It is worth noting that the codimension of a singularity of (2) in the theory of [8] is always greater than or equal to the standard codimension obtained by setting $z = (x, \lambda)$, $M = N + 1$.

3. Defining equations for singular solutions of low dimensional systems

We want to set up a system of equations which determines singular solutions (z_0, c_0) of (1) of a given corank n and a given codimension k . As we will see in section 5, this task can be reduced by theorem 1 to the determination of a singular solution (w_0, c_0) w.r.t. w of codimension k for a low dimensional system

$$(13) \quad G(w, c) = 0, \quad G \in C^\infty(\mathbb{R}^m \times \mathbb{R}^p, \mathbb{R}^n) \quad \text{with index } i = m - n \geq 0.$$

The prescribed codimension k suggests that we have to have at least $p \geq k$ parameters in (13) in order to find a singular solution of codimension k . In fact, we assume $p = k$ in (13). This can always be achieved by either fixing some control variables or inserting new ones (these should be of physical relevance for the underlying system).

A square system of equations

$$(D) \quad D_G(w, c) = 0, \quad \text{where } D_G \in C^\infty(\mathbb{R}^{m+k}, \mathbb{R}^{m+k})$$

will then be called a *defining equation* for G if it has the following property

- (P) P1: (w_0, c_0) is a regular solution of (D) \Leftrightarrow
 P2: (w_0, c_0) is a singular solution of $G(w, c) = 0$ w.r.t. w ,
 it is of corank n and $G(w_0 + w, c_0 + c)$ is a universal unfolding
 of $G(w_0 + w, c_0)$.

This is a rather strong requirement. For example, if we have computed a regular solution (w_0, c_0) of (D) then we are sure that the variation of c around c_0 will exhibit all possible solution pictures (up to diffeomorphisms) in the state space \mathbb{R}^m . The introduction of additional parameters in (13) will not create new phenomena. In the following table of defining equations the property (P) is satisfied in the strict sense only in the case of corank 1. For the corank 2 singularities some additional nondegeneracy conditions (A2), (A3) have to be added to P1 in order to make (P) correct.

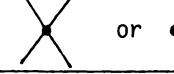
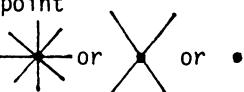
i	n	k	$D_G(w, c)$	name	representative germs $G(w, 0)$ near $w = 0$
0	1	k	(G, G_w, \dots, G_{wk}) where $G_{wk} = \frac{\partial^k G}{\partial w^k}$	fold ($k=1$), cusp ($k=2$) swallow-tail ($k=3$) butterfly ($k=4$)	w^{k+1}
0	2	4	(G, G_w) (A2)		$\begin{pmatrix} w_1^2 \\ w_2^2 \end{pmatrix}, \begin{pmatrix} w_1^2 \pm w_2^2 \\ w_1 w_2 \end{pmatrix}$
1	1	1	(G, G_w)	simple bifurcation point or isolia center (hermit) 	$w_1^2 \pm w_2^2$
1	1	2	$(G, G_w, \det G_{ww})$	cusp curve 	$w_1^2 - w_2^3$
1	2	5	(G, G_w) (A3)	multiple bifurcation point 	$\begin{pmatrix} w_1^2 \pm w_2^2 \\ w_2^2 \pm w_3^2 \end{pmatrix}, \begin{pmatrix} w_1^2 + w_2^2 - w_3^2 \\ -2w_1 w_3 + w_3^2 \end{pmatrix}$

table 1

(Am) ($m = 2, 3$) the homogeneous quadratic $q(w) = \frac{1}{2}G_{ww}(w_0, c_0)w^2$ satisfies $q(\tilde{w}) = 0$, $\tilde{w} \in \mathbb{C}^m$, $\tilde{w} \neq 0 \Rightarrow q_w(\tilde{w})$ has rank 2 over \mathbb{C} .

(A2) means that the conic sections $q_i(w) = 0$ ($i=1, 2$) are nowhere tangent in \mathbb{C}^2 except at 0, or what is the same, q_1 and q_2 have no common factor. Similarly, (A3) means that the quadratic surfaces $q_i(w) = 0$ ($i=1, 2$), $w \in \mathbb{C}^3$ are nowhere tangent except at 0. We note that the real version of (A3) is a well known assumption in the study of bifurcation at a double eigenvalue [11, 8, § 5] (the use of (A3) with \mathbb{R} instead of \mathbb{C} is sufficient in [8, § 5] because of the special way in which $\lambda = w_3$ enters into the problem there).

Table 1 shows that a defining equation does usually not determine a unique class but several classes of contact equivalent singularities. These were indicated by some simple representatives in the last column.

The proof of property (P) for all entries of table 1 would be too long, so here we restrict ourselves to the question for which n and m

$$(14) \quad D_G(w, c) = (G, G_w)(w, c) = 0, \quad G \in C^\infty(\mathbb{R}^m \times \mathbb{R}^k, \mathbb{R}^n)$$

is a defining equation. In order to make (14) a square system we need $k = n + nm - m$.

Theorem 2:

Let $n \leq m$, $k = k(n, m) = n + nm - m$ and let $G \in C^\infty(\mathbb{R}^m \times \mathbb{R}^k, \mathbb{R}^n)$. Then P2 implies P1. If P1 is assumed then (w_0, c_0) is a singular solution of (13) w.r.t. w which is of corank n and codimension $\geq k$. If the codimension is equal to k , then $G(w_0 + w, c_0 + c)$ is a universal unfolding of $G(w_0, c_0)$ and the pair (n, m) must be one of the following

$$(15) \quad \text{either } (n=1, m \in \mathbb{N}) \text{ or } (n=2, m \in \{2, 3\}).$$

Remark: We have the somewhat surprising result that (14) satisfies the property (P) only for some restricted values of n and m . This should be due to the occurrence of so called modal parameters (cf. [8, § 4,5]).

Proof: Without loss of generality we may assume $(w_0, c_0) = (0, 0)$. Let P2 be satisfied. Then obviously $G^0 = 0$ and $\text{rank } G_w^0 = 0$ hold, hence $D_G^0 = 0$. Moreover, we have

$$(16) \quad D_G^{1,0} = \begin{pmatrix} G_w^0 & G_c^0 \\ G_{ww}^0 & G_{wc}^0 \end{pmatrix} = \begin{pmatrix} 0 & G_c^0 \\ G_{ww}^0 & G_{wc}^0 \end{pmatrix} .$$

By the transversality condition (12) each $q \in E_m^n$ may be written as

$$(17) \quad q(w) = \tau(w)G(w,0) + G_w(w,0)\sigma(w) + G_c(w,0)\gamma$$

for some $\tau \in E_m^{n,n}$, $\sigma \in E_m^m$, $\gamma \in \mathbb{R}^k$. Evaluating q and q_w at $w = 0$ yields

$$(18) \quad \begin{pmatrix} q(0) \\ q_w(0) \end{pmatrix} = \begin{pmatrix} G_c^0\gamma \\ G_{ww}^0\sigma^0 + G_{wc}^0\gamma \end{pmatrix} = D_G^{1,0} \begin{pmatrix} \sigma^0 \\ \gamma \end{pmatrix} .$$

Since the left hand sides span \mathbb{R}^{n+nm} we obtain that $D_G^{1,0}$ is nonsingular.

Assume now that P1 is satisfied. If a relation (17) holds with $q = 0$ then we

find $\sigma^0 = 0$, $\gamma = 0$ from (18) and the regularity of $D_G^{1,0}$. Thus we have shown

that the functions $G_{cj}(\cdot, 0)$ ($j=1, \dots, k$) are linearly independent in

$E_m^n / TG(\cdot, 0)$. The codimension of $G(\cdot, 0)$ at 0 is therefore at least k and in case of equality $G(w, c)$ is a universal unfolding of $G(w, 0)$ since (12) is satisfied.

It remains to prove (15). Let $Q_2 \subset E_m^n$ be the linear subspace of homogeneous

quadratics. Each $q \in Q_2$ has a representation (17). From (18) we again find

$\sigma^0 = 0, \gamma = 0$ and differentiating (17) twice at $w = 0$ yields

$$(19) \quad q_{i,ww}^0 = \sum_{j=1}^n \tau_{ij}^0 G_{j,ww}^0 + G_{i,ww}^0 \sigma_w^0 + \sigma_w^{0T} G_{i,ww}^0 \quad (i=1, \dots, n).$$

We consider this as a linear system for the $n^2 + m^2$ unknowns τ^0 and σ_w^0 . Since the left hand sides of (19) span a linear space of dimension $\frac{1}{2}nm(m+1) = \dim Q_2$ we obtain $\frac{1}{2}nm(m+1) \leq n^2 + m^2$. But the homogeneous equation (19) also admits the nontrivial solution $\tau^0 = -2I$, $\sigma_w^0 = I$ so that in fact $\frac{1}{2}nm(m+1) \leq n^2 + m^2 - 1$. An elementary discussion of this inequality using $n \leq m$ then leaves us with the cases given in (15). \square

In the case $n=1, m \in \mathbb{N}$ the proof of property (P) is easily completed. Let us assume P1. Now we have $k = k(1, m) = 1$ and from (16) and the regularity of $D_G^{1,0}$ we find that the Hessian G_{ww}^0 is nonsingular. By the Morse lemma $G(w, 0)$ is then contact equivalent to a quadratic $\sum_{i=1}^m \epsilon_i w_i^2$, $\epsilon_i \in \{-1, 1\}$ and hence has codimension 1 (cf. [6, IV § 4]). The assertion P2 then follows from theorem 2.

In the case $n=2, m \in \{2, 3\}$ we assume P1 and (Am) and define $F(w) = \frac{1}{2} G_{ww}^0 w^2$, $w \in \mathbb{R}^m$. After some algebraic manipulations, which we omit, the property (Am) turns out to be equivalent to

(20) for each $q \in Q_2$ there exist matrices $A \in \mathbb{R}^{2,2}$, $B \in \mathbb{R}^{m,m}$ such that
 $q(w) = A F(w) + F_w(w) Bw, w \in \mathbb{R}^m$.

Since F agrees with $G(\cdot, 0)$ to second order we obtain from (20)

$$q(w) = A G(w, 0) + G_w(w, 0) Bw + O(\|w\|^3).$$

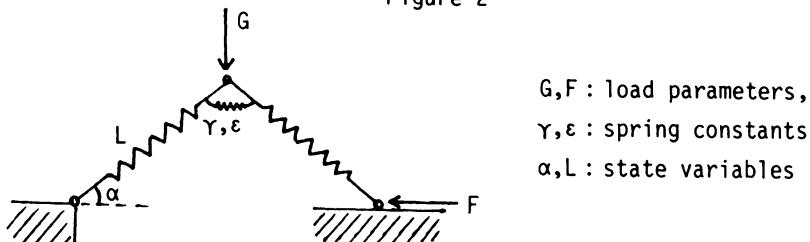
This result may be written in a more formal way as $M_{2,m}^n \subset TG(\cdot, 0) + M_m \cdot M_{2,m}^n$, where $M_{2,m}^n$ is the module generated by Q_2 in E_m^n and M_m is the maximal ideal in E_m^1 which is generated by the germs $w \rightarrow w_i$ ($i=1, \dots, m$). The lemma of Nakayama [6, IV § 2] then ensures $M_{2,m}^n \subset TG(\cdot, 0)$. Therefore, a basis of $E_m^n / M_{2,m}^n$, which consists of $nm + n$ linear germs, also spans $E_m^n / TG(\cdot, 0)$. By the regularity of $D_G^{0,0}$ we have $G_{ww}^0 \gamma = 0, \gamma \in \mathbb{R}^m \Rightarrow \gamma = 0$. This proves that the linear functions $p_i(w) = G_{ww}^0 e^i w$ ($i=1, \dots, m, e^i = i$ -th unit vector in \mathbb{R}^m) are linearly independent. Moreover $p_i(\cdot) = F_w(\cdot) e^i = G_w(\cdot, 0) e^i + (F_w(\cdot) - G_w(\cdot, 0)) e^i \in TG(\cdot, 0)$ $+ M_{2,m}^n \subset TG(\cdot, 0)$, $i=1, \dots, m$ so that finally $\dim E_m^n / TG(\cdot, 0) \leq nm + n - m = k$. Again theorem 2 completes the proof of P2.

Let us finally notice that the nontrivial solutions $\hat{w} \in \mathbb{R}^3$ of $G_{ww}(w_0, c_0) \hat{w}^2 = 0$ determine the bifurcation directions in the case $m = 3$. Due to (A3) there are either 0, 2 or 4 branches passing through w_0 .

4. Two applications

Our first example describes the buckling of a spring (fig. 2) and is taken from [14, §13.8] (with the exception of the parameter ϵ).

Figure 2



The total energy of the system is

$$U(\alpha, L; G, F, \gamma, \epsilon) = \frac{1}{2}\gamma(2\alpha)^2 + \frac{1}{3}\epsilon(2\alpha)^3 + (L-1)^2 + FL \cos \alpha + GL \sin \alpha.$$

From $U_\alpha = U_L = 0$ we can eliminate L and we find for the stationary states the scalar equation

$$(21) \quad T(\alpha, G, F, \gamma, \varepsilon) = 4\gamma\alpha^2 + 8\varepsilon\alpha^3 + G \cos \alpha - 2F \sin \alpha - GF \cos 2\alpha + (F^2 - \frac{1}{4}G)^2 \sin 2\alpha = 0.$$

This system has a butterfly point w.r.t. α at

$$(22) \quad \alpha = 0; \quad G = 0, \quad F = \frac{1}{4}, \quad \gamma = \frac{3}{32}, \quad \varepsilon = 0 \quad (\text{see [14]}),$$

where the four parameters provide a universal unfolding. Figure 3 shows a portion of the solutions in the (α, G, F) -space at fixed values $\gamma = 0.11$, $\varepsilon = 10^{-3}$. It also shows how we approached the butterfly point (22) numerically by solving the defining equations $(T, T_\alpha, \dots, T_{\alpha k}) = 0$ for increasing k . Starting with a regular solution $(\alpha = 1.8785, G = 0, F = 0.4)$ we computed a branch of regular solutions (r) by varying G . Then we switched to the defining equation of the fold $(T, T_\alpha)(\alpha, G) = 0$ and computed a branch of folds (f) by varying F . Proceeding in this way we obtained a branch of cusps (c) and swallow tails (s) which finally ends at the butterfly point (b).

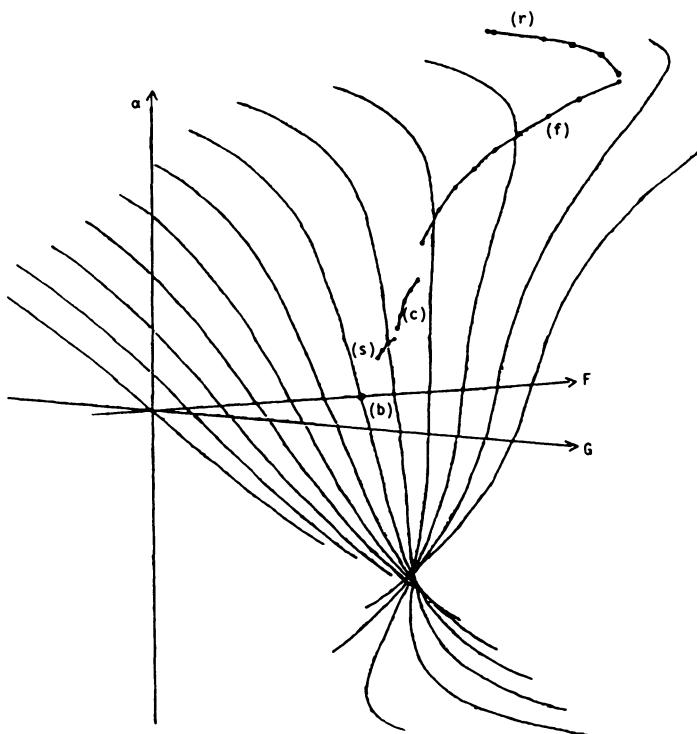


Figure 3

the jumps between the curves (r), (f), (c) and (s) are caused by the switching to the next defining equation

Our second example is an N-cell model with diffusion and reaction as discussed in detail in [2, 3]. The steady state equations are of the following form

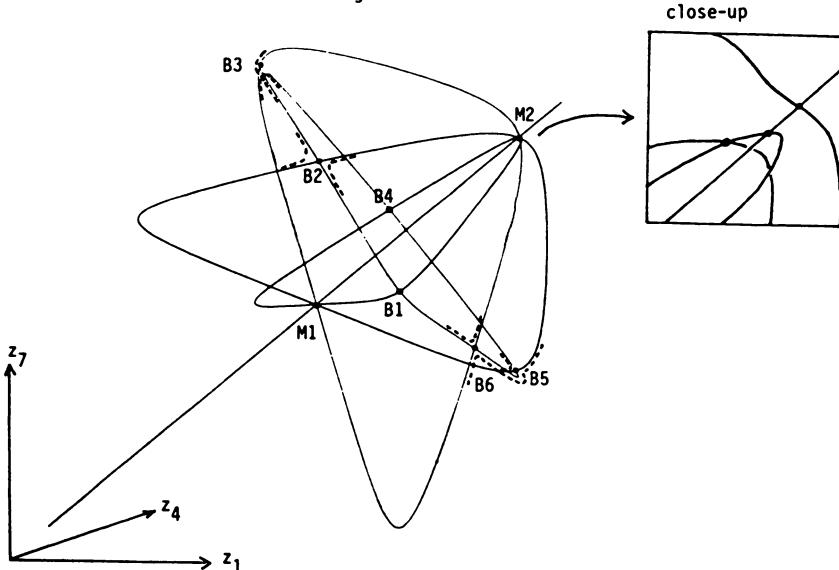
$$(23) \quad D_j(z_j - z_{j-1}) - D_{j+1}(z_{j+1} - z_j) + d_j z_j = h_j g(c_0 - z_j, \lambda, \mu), \quad j=1, \dots, N$$

$$z_0 = z_{N+1} = 0 \quad \text{where } z_j = c_0 - c_j$$

and c_j (resp. c_0) = substrate concentration in the j -th cell (resp. outer reservoir), D_j (d_j) = diffusion constant between the j -th cell and the $(j-1)$ -th cell (the outer reservoir), h_j = thickness of the j -th cell, $g(x, \lambda, \mu) = 10^\mu x(1+x+\lambda x^2)^{-1}$ = reaction rate of an inhibited Michaelis-Menten process. The cells with numbers 0 and $N+1$ should be interpreted as part of the reservoir. Since we are interested in the branches generated by varying λ we set $z = (z_1, \dots, z_N, \lambda)$, $c = (D_1, \dots, D_{N+1}, d_1, \dots, d_N, h_1, \dots, h_N, \mu)$ and write (23) in the form (1) with $M=N+1$.

Here we consider a special case of 7 cells where 3 cells are in contact with the outer reservoir ($D_j = 0.3$ ($j=2, \dots, 7$), $D_1 = D_8 = 1$, $d_4 = 2$, $d_j = 0$ otherwise, $h_4 = 2$, $h_j = 1$ otherwise, $c_0 = 4$, $\mu = 1.4$). A three-dimensional view of the numerical solution branches in the (z_1, z_4, z_7) -space is shown in fig. 4.

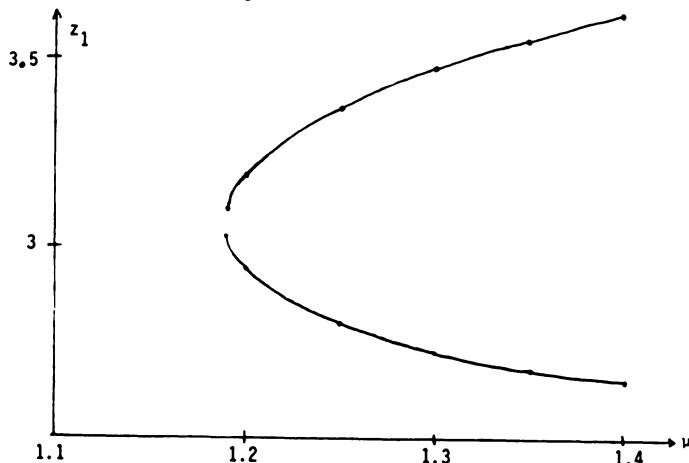
Figure 4



The numerical results suggest 6 simple bifurcation points (B1 - B6) and 2 multiple bifurcation points (M1, M2). However, a closer look at the

continuation procedure near these points reveals that only B1 and B4 are bifurcation points (which we have computed by the defining equations of section 3) whereas the others are actually separated as indicated by the dotted lines in fig. 4. Moreover, near M1 and M2 there are in fact 3 closely spaced simple bifurcation points as shown in the close up of fig. 4. These could only be determined by using very good initial guesses for the defining equation. A more detailed explanation of the branching structure in fig. 4 which uses the inherent symmetries in equation (23) is given in [4]. By table 1 we expect to find true multiple bifurcation points near M1, M2 if we let 5 parameters vary in the system (23). We used $c_1 = d_1$, $c_2 = d_4$ and the perturbations $c_3 z_4$, $c_4 z_7$, $c_5 z_1$ in the equations (23) with the numbers 1, 4, 7 respectively. Newton's method for the defining equation was then successful when started near M1, M2 and multiple bifurcation points with 4 bifurcation directions were detected (see section 5 for details of the reduction process). By varying the parameter μ we then found an upper and a lower branch of multiple bifurcation points (fig. 5). These seem to coalesce at a new singularity the type of which we do not know at present.

Figure 5



5. The reduction process

In order to compute a singular solution of the system (1) which is of codimension $k = p$ we apply the defining equations from table 1 to a Liapunov-Schmidt reduction $S(w, c)$ of T w.r.t. (V, W, X, Y) , i.e. we solve

$$(24) \quad D_S(u) = D_S(w, c) = 0, \quad u = (w, c) \in W \oplus \mathbb{R}^k (\cong \mathbb{R}^{m+k}).$$

If D_S has the property (P) and if the assumptions of theorem 1 hold then theorem 1 can be used to prove the following: (w_0, c_0) is a regular solution of (24) $\Leftrightarrow (z_0, c_0) = (v(w_0, c_0), w_0, c_0)$ is a singular solution of (1) of corank n and $T(z_0 + z, c_0 + c)$ is a universal unfolding of $T(z_0 + z, c_0)$.

We used a two stage process for the numerical solution of (24). This will only be briefly outlined here since less costly methods are available for special singularities [12, 9, 19]. In particular, in [9] singular solutions of corank 1 and of codimension 1 but of arbitrary index (these have the defining equation $(G, G_w) = 0$, compare section 3) are obtained in an efficient way by solving (1) together with a set of M+1-N scalar equations which characterize the points where rank $T_z = N - 1$.

Suppose D_S involves the derivatives S, S_w, \dots, S_{wr} ($r=1$ in most cases), then one Newton step for the system (24) needs $S(u)$ and the derivatives $S_{wj_u}(u)$ ($j=0, \dots, r$). We always used coordinate subspaces for V, W, X, Y and a few Newton steps for (4) in order to obtain a good approximation to $v(w, c) = v(u)$ and hence to $S(u)$ from (6). Differentiating (6) with respect to w and u shows that S_{wj_u} is of the form

$$(25) \quad S_{wj_u} = (I - P)(T_v v_{wj_u} + \kappa_j), \quad j = 0, \dots, r.$$

Here we have suppressed the arguments and denoted by κ_j all terms which involve only lower order derivatives of the implicit function $v(u)$. The expressions for κ_j get more and more complicated, for example $\kappa_0 = T_u$, $\kappa_1 = T_{vv}v_u + T_{vv}v_wv_u + T_{uv}v_w + T_{uw}$. The derivatives v_{wj_u} are computed from the linear systems

$$(26) \quad (PT_v) v_{wj_u} = -P\kappa_j \quad (j = 0, \dots, r)$$

which are obtained by differentiating $PT(v(u), u) = 0$. It is worth noting that the same matrix $PT_v(v(u), u)$ appears in all systems (26) so that one LU-factorization is sufficient. Moreover, our reduction process is essentially independent of the special form of the defining equation (24).

References

- [1] Beyn, W.-J.: Numerical analysis of singularities in a diffusion reaction model. To appear in Springer Lecture Notes, Proceedings of the EQUADIFF 82.
- [2] Bigge, J., Bohl, E.: On the steady states of finitely many chemical cells (to appear).
- [3] Bohl, E.: Discrete versus continuous models for dissipative systems (this volume).

- [4] Bohl, E., Beyn, W.-J.: Organizing centers for discrete reaction diffusion models (this volume).
- [5] Bröcker, Th.: Differentiable germs and catastrophes. London Math. Soc. Lecture Note Series 17, 1974.
- [6] Gibson, C.G.: Singular points of smooth mappings. Research notes in Mathematics 25, Pitman, 1979.
- [7] Golubitsky, M., Keyfitz, B.L.: A qualitative study of the steady state solutions for a continuous flow stirred tank chemical reactor. SIAM J. Math. Anal. 11, 316-339, 1980.
- [8] Golubitsky, M., Schaeffer, D.: A theory for imperfect bifurcation via singularity theory. Commun. Pure Appl. Math. 32, 21-98, 1979.
- [9] Griewank, A., Reddien, G.W.: Characterization and computation of generalized turning points. To appear in SIAM J. Numer. Anal.
- [10] Martinet, J.: Singularities of smooth functions and maps. London Math. Soc. Lecture Note Series 58, 1982.
- [11] McLeod, J.B., Sattinger, D.: Loss of stability and bifurcation at a double eigenvalue. J. Funct. Anal. 14, 62-84, 1973.
- [12] Melhem, R.G., Rheinboldt, W.C.: A comparison of methods for determining turning points of nonlinear equations. Computing 29, 201-226, 1982.
- [13] Moore, G.: The numerical treatment of non-trivial bifurcation points. Numer. Funct. Anal. Optimiz. 2, 441-472, 1980.
- [14] Poston, T., Stewart, I.: Catastrophe theory and its applications Pitman, London, 1978.
- [15] Seydel, R.: Numerical computation of branch points in nonlinear equations. Numer. Math. 33, 339-352, 1979.
- [16] Spence, A., Werner, B.: Nonsimple turning points and cusps. IMA J. of Numer. Anal. 2, 413-427, 1982.
- [17] Spence, A., Jepson, A.: The numerical computation of turning points of nonlinear equations. 169-183 in Treatment of Integral Equations by Numerical Methods (Ed.: C.T.H. Baker, G.F. Miller), Academic Press, 1982.
- [18] Weber, H.: On the numerical approximation of secondary bifurcation points. 407-425 in Springer Lecture Notes in Mathematics 878 (Ed.: E.L. Allgower et al.), 1981.
- [19] Werner, B., Spence, A.: The computation of symmetry-breaking bifurcation points. To appear in SIAM J. Numer. Anal.

Dr. W.-J. Beyn
 Fakultät für Mathematik
 der Universität Konstanz
 Postfach 5560
 D-7750 Konstanz

ORGANIZING CENTERS FOR DISCRETE REACTION DIFFUSION MODELS

W.-J.Beyn, E.Bohl

1. INTRODUCTION

This paper supplements the two papers [1,2] in this book. Our references are included in the list of references of [1,2]. We will refer to the j -th reference in [1], [2] by $[j]_1$, $[j]_2$, respectively.

Let us consider an assemblage of finitely many chemical cells as described in the introduction of [2]. More generally, we allow for more than just two cells to be connected to the outside reservoir via membranes. If h_j is the width of the j -th cell, if E_j is the diffusion constant between the j -th cell and the outside reservoir and if D_j is the diffusion coefficient between the $(j-1)$ -th cell and the j -th cell, then the corresponding system reads

$$(1a) \quad (E_1 + D_2)x_1 - D_2x_2 = h_1 f(x_1, \lambda),$$

$$(1b) \quad -D_j x_{j-1} + (E_j + D_j + D_{j+1})x_j - D_{j+1}x_{j+1} = h_j f(x_j, \lambda), \quad (j=2, \dots, N-1)$$

$$(1c) \quad -D_N x_{N-1} + (E_N + D_N)x_N = h_N f(x_N, \lambda).$$

The case $E_j = 0$ describes a cell which is not connected to the outside reservoir. The generation term f is qualitatively given by Fig.3 of [2]. There we have given examples in $(4)_2$ (we refer to the formula (j) in [2] as $(j)_2$). λ is again a control parameter. We consider our assemblage to be made up of end units (see Fig.1a, 1c) and middle units (see Fig.1b). For any unit, the number of cells which are not connected to the outside reservoir is arbitrary but at least one for end units and at least two for middle units. Hence the smallest assemblage constructed in this way consists of seven cells and is given in Fig.1.

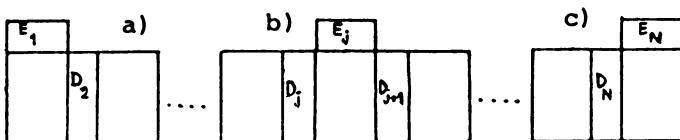


Fig.1

To understand the solution set of (1) we visualize all the parameters E_j , D_j , h_j , λ ($j=1, \dots, N$)

as control parameters and combine them in the control parameter vector $v \in \mathbb{R}^p$ where p is a sufficiently large natural number. Then an abbreviation for the system (1) is

$$(2) \quad F(x, v) = 0, \quad x \in \mathbb{R}^N, \quad v \in \mathbb{R}^p.$$

Here for any $v \in \mathbb{R}^p$ the function F maps \mathbb{R}^N into itself. We are trying to construct an organizing center (see [7], for this notion) in the solution set of (1) or (2) which determines the structure of this set in the neighborhood of the center. An organizing center is a singularity. Its universal unfolding structurally gives the complete picture of our solution set in the neighborhood of the center. In section 2 we first obtain the type of the singularity in a heuristic way. In section 3 we test our intuition numerically: Structures in the solution diagram of the universal unfolding of the singularity must have their counterpart in the solution set of (1) and vice versa. Hence we take some characteristic structures in one of the two diagrams and try to find a diffeomorphic picture in the other diagram by perturbation of suitable parameters. The results show remarkable agreement of the predictions and the answers. For all structures we have picked in one diagram we could find a counterpart in the other. We note that we only publish some of our tests in this paper. In fact, we tried many more situations and always found the expected answer.

2. A SINGULARITY

In the spirit of [2] we take our assemblage of cells apart into finitely many pieces of the form given in Fig.1. These are two end units and finitely many middle units as described in the introduction.

Let us first take a closer look at an end unit. In the case considered in section 2 of [2] the solution set of an end unit is given by Fig.5 in [2]. Let us concentrate on one of the hysteresis loops (see Fig.5 in [2]) which disappears during the transition described in [2]. We have seen in [2] that this dynamics can be understood by a perturbation of the solution set of the following simple algebraic equation

$$(3) \quad z^3 - \lambda = 0$$

(see (15)₂). Numerical studies on the system (1) for the middle unit show that also a middle unit produces hysteresis phenomena which behave like perturbations of the polynomial equation (3).

We now turn to the general case of our assemblage of cells. We assume that we can fix all control parameters except for λ such that the solution curve of any unit with respect to the parameter λ shows a hysteresis loop in the neighborhood of a common value λ_0 of the control parameter. All these loops are supposed to behave like a perturbation of (3) if some of the other control parameters in (1) undergo suitable perturbations. Then it is tempting to conjecture that the whole system (1) describing the complete assemblage of cells works locally like a perturbation of equations of the form (3) yielding the system

$$(4) \quad z_j^3 - \lambda = 0, \quad j = 1, \dots, K,$$

where K is the number of units which make up our complete assemblage. For each unit we simply put down an equation of the form (3). Eliminating λ from the system (4) we are left with the system

$$(5) \quad z_j^3 - z_{j+1}^3 = 0, \quad j = 1, \dots, K-1.$$

This describes a singularity at the point $(0, \dots, 0) \in \mathbb{R}^{K-1}$ which defines our organizing center mentioned in the introduction. We proceed now as described at the end of the introduction: The unfolding of (5) at the origin gives the solution pictures which we have to spot in the solution set of (1).

It is very difficult to obtain a universal unfolding for (5) in general. So we retreat in this paper to the two cases $K=2$ and $K=3$. Here we can give the universal unfolding of (5). We then know the perturbation pictures predicted by the singularity (5) and can try to adjust the parameters D_j, E_j, h_j ($j=1, \dots, N$) in (1) to find qualitatively the same pictures in the solution set of (1).

The case $K=2$ has already been considered in [2], [3b,4d]₂. Here our assemblage consists of two end units or an end unit and a middle unit or two middle units. In particular, the case of two end units has been studied [2], [3b,4d]₂ with the result that we could observe all predictions of the singularity

$$(6) \quad z_1^3 - z_2^3 = 0$$

(note $K=2$ in (5)) for the corresponding system which is in this case the system (1)₂. In particular the three pictures of Fig.6 in [2] are part of the universal unfolding of (6) which reads

$$(7) \quad z_1^3 - z_2^3 + \alpha_1 + \alpha_2 z_1 + \alpha_3 z_2 + \alpha_4 z_1 z_2 = 0,$$

This unfolding is discussed in detail in [1]₁.

3. THE CASE K=3

In this section we are concerned with the case $K=3$. We join the three parts of Fig.1 via membranes and arrive at a total of seven cells, three of which are connected to the outside reservoir and separated from each other by two cells with no connection to this reservoir. We assume the generation term f to be of the form (4a)₂ with

$$(8) \quad \lambda_1 = 10^{1.4}, \quad \lambda_2 = 4, \quad \lambda_3 = \lambda = \text{control parameter}.$$

The corresponding system (1) takes the form

$$(9a) \quad \begin{aligned} (E_1 + D_2)x_1 - D_2x_2 &= h_1 f(x_1, \lambda) \\ -D_2x_1 + (D_2 + D_3)x_2 - D_3x_3 &= h_2 f(x_2, \lambda) \end{aligned}$$

$$(9b) \quad \begin{aligned} -D_3x_2 + (D_3 + D_4)x_3 - D_4x_4 &= h_3 f(x_3, \lambda) \\ -D_4x_3 + (E_4 + D_4 + D_5)x_4 - D_5x_5 &= h_4 f(x_4, \lambda) \\ -D_5x_4 + (D_5 + D_6)x_5 - D_6x_6 &= h_5 f(x_5, \lambda) \end{aligned}$$

$$(9c) \quad \begin{aligned} -D_6x_5 + (D_6 + D_7)x_6 - D_7x_7 &= h_6 f(x_6, \lambda) \\ -D_7x_6 + (D_7 + E_7)x_7 &= h_7 f(x_7, \lambda) \end{aligned}$$

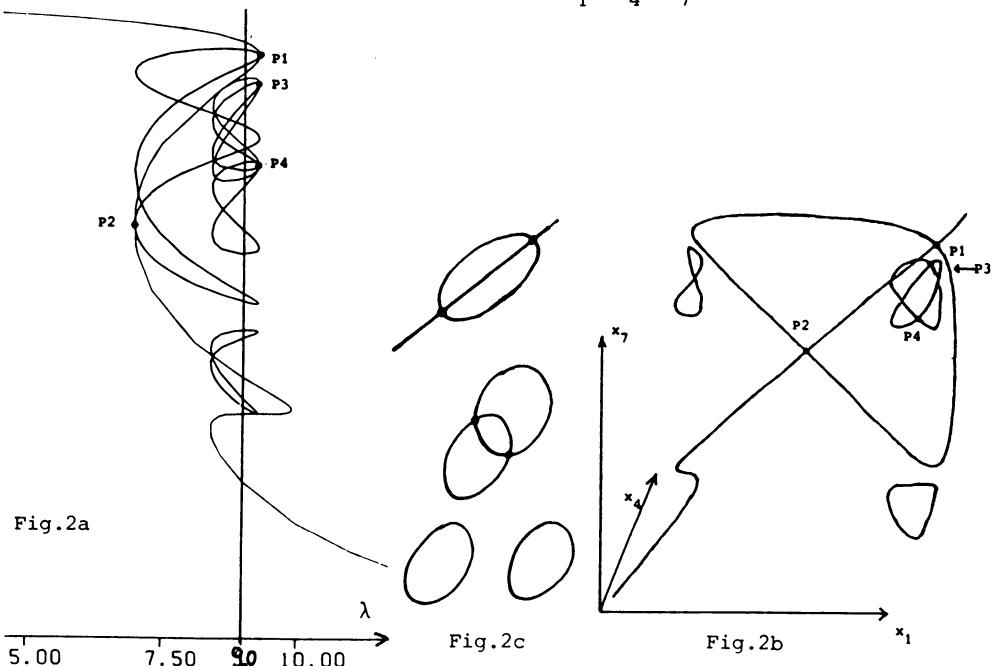
Here (9a) and (9c) govern the two end units and the three equations (9b) the middle unit. We now have to find diffusion constants E_j , D_j and cell lengths h_j such that all three parts (9a), (9b) and (9c) show hysteresis behaviour with respect to the control parameter λ .

This situation occurs if we put

$$(10) \quad E_j = 1 \quad (j=1, 4, 7), \quad D_i = .3, \quad h_i = 1 \quad (i=1, \dots, 7).$$

Then we can combine for the control parameter $\lambda=9$ out of three solutions for any of the subsystems (9a), (9b), (9c) a total of 27 solutions for the full system (9). We start a continuation procedure at each of these points and find

a net of branches as shown in Fig.2a. In the net four bifurcation points occur which we have marked by P1 to P4. The remaining intersections in Fig.2a do not correspond to bifurcation points of the system (9), they are caused by the choice of functionals which we made to draw the bifurcation diagram. In fact, the picture is simplified considerably if we plot the same branches in a (x_1, x_4, x_7) coordinate system as it is done in Fig.2b. This change of view may be compared with the elimination of the parameter λ from the singular system (4) (the vertical axis in Fig.2a is $(x_1 + 2x_4 + 3x_7)/6$).



A further simplification occurs if we draw a picture showing the topological type of the net without any reference to the values of the variables x_1, \dots, x_7, λ . The resulting diagram is given in Fig.2c. It consists of one open curve and five closed curves, one of which is connected to the open curve at two bifurcation points and two of which are tied together at two bifurcation points. The remaining two curves form isolas which are disconnected from each other and from the other branches.

Let us first try to retrieve this configuration from the universal unfolding of the singularity (5) with $K=3$. In this case we may rewrite (5) as

$$(11) \quad x^3 - y^3 = 0, \quad z^3 - y^3 = 0.$$

From the theory in [6]₂ we find that a universal unfolding of (11) needs 28 parameters and a particular one is given by

$$(12) \quad \begin{aligned} U_1(x,y,z,\alpha) &= x^3 - y^3 + \alpha_1 + \alpha_2 x + \alpha_3 y + \alpha_4 z + \alpha_5 xy + \alpha_6 y^2 + \alpha_7 yz + \alpha_8 z^2 \\ &\quad + \alpha_9 xz + \alpha_{10} xy^2 + \alpha_{11} xz^2 + \alpha_{12} xyz + \alpha_{13} y^2 z + \alpha_{14} yz^2 + \alpha_{15} xy^2 z + \alpha_{16} xyz^2 \\ U_2(x,y,z,\alpha) &= z^3 - y^3 + \alpha_{17} + \alpha_{18} x + \alpha_{19} y + \alpha_{20} z + \alpha_{21} x^2 + \alpha_{22} xy + \alpha_{23} yz \\ &\quad + \alpha_{24} xz + \alpha_{25} zx^2 + \alpha_{26} xyz + \alpha_{27} yx^2 + \alpha_{28} x^2 yz. \end{aligned}$$

Of course it is impossible to grasp all types of solution branches of the system $U_1=0, U_2=0$ when α varies in \mathbb{R}^{28} . Therefore we are compelled to drop many of the parameters from the unfolding (12). Here we are guided by special properties of the system (9) with the values of (10) which should be reflected by the unfolding. In particular, we keep in mind that the variables x, y, z correspond to the concentrations x_1, x_4, x_7 in those cells which are connected to the reservoir.

I. Symmetry

The complete system (9) at the values of (10) is invariant under the transformation $x_i \rightarrow x_{8-i}$ ($i=1, \dots, 7$). Hence we require $U_1(x,y,z,\alpha)=U_2(z,y,x,\alpha)$ which leaves us with a total of 12 parameters instead of 28 in (12).

II. Decoupling

$U_1=0$ models the coupling of the left end unit and the middle unit whereas $U_2=0$ does the same for the middle unit and the right end unit. It seems therefore reasonable to let U_1 be independent of z and U_2 be independent of x . This condition reduces the number of parameters in (12) to 10.

If we impose both conditions I and II on the universal unfolding (12) then we end up with the following 4-parameter unfolding

$$(13a) \quad F_1(x,y,\beta) = x^3 - y^3 + \beta_1 + \beta_2 x + \beta_3 y + \beta_4 xy = 0$$

$$(13b) \quad F_2(x,y,\beta) = z^3 - y^3 + \beta_1 + \beta_2 z + \beta_3 y + \beta_4 zy = 0.$$

We recognize that this system contains the unfolding (7) two times with a coupling through the variable y . Projecting the solution sets of (13) onto the (x,y) -plane and the (y,z) -plane will therefore yield the known solution curves of (7) (cf.[1]). On the other hand we can combine the solution set of (13) from the single curves (13a) and (13b). This combination follows some simple

rules which we have illustrated in an obvious way in Tab.1. These rules apply to any system of the type

$$(14) \quad F(x,y) = 0, \quad F(z,y) = 0$$

and could in fact be put into rigorous theorems.

(x,y) -branch of $F(x,y) = 0$	and (z,y) -branch of $F(z,y) = 0$	combines to (x,y,z) -branch of (14)

Tab.1

For example, if we consider the parameter set $\beta_1 = \varepsilon$, $\beta_2 = -1$, $\beta_3 = 1$, $\beta_4 = 0$ ($\varepsilon > 0$ small), then the cubic curves (13a), (13b) are of the following form



Fig.3

Combining the two pictures according to the rules of Tab.1 we get a structure which is diffeomorphic to the one given in Fig.2c. Hence we have found the counterpart of the structure of Fig.2a in an unfolding of the organizing center (11).

An even more interesting solution net of the system (9) occurs for the following set of parameters

$$(15) \quad E_1 = E_7 = 1, \quad E_4 = 2, \quad D_i = .3 \quad (i=1, \dots, 7), \quad h_i = 1 \quad (i=1, \dots, 7, i \neq 4), \quad h_4 = 2.$$

This situation differs from (10) only in E_4 and h_4 . In the (x_1, x_4, x_7) -space we obtain a solution picture for (9) as shown in Fig.4.

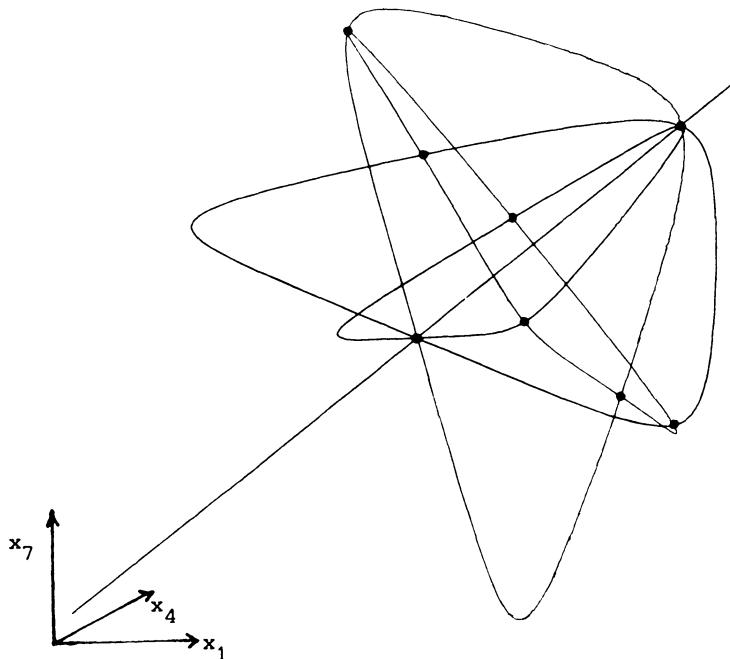
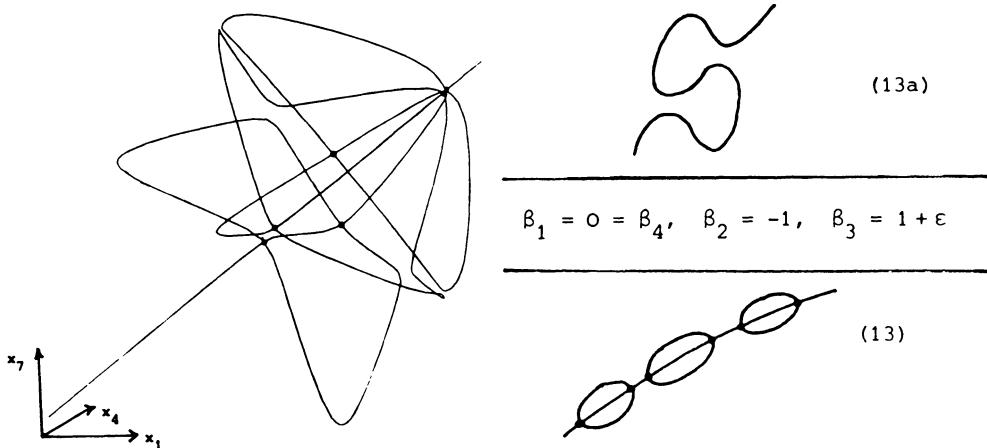
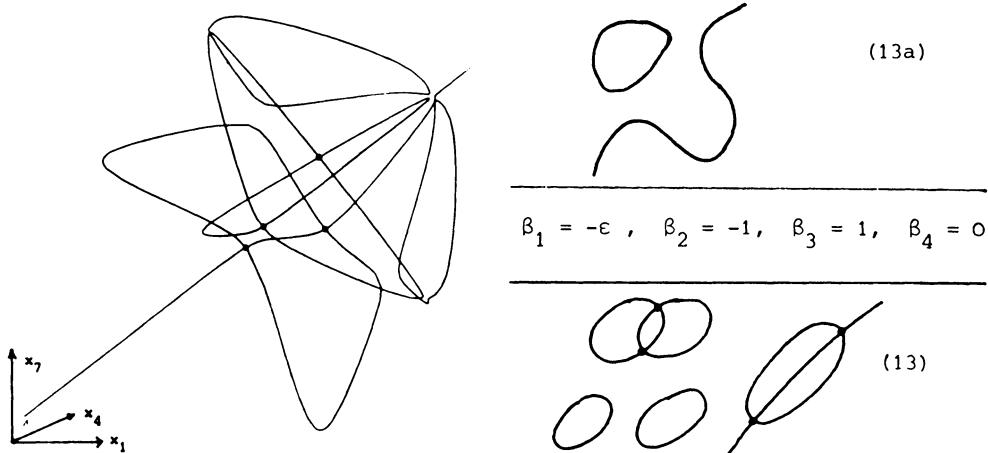


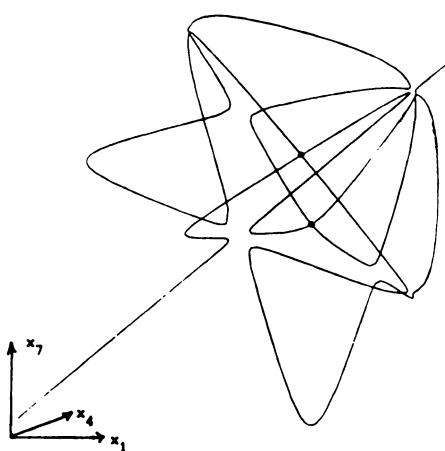
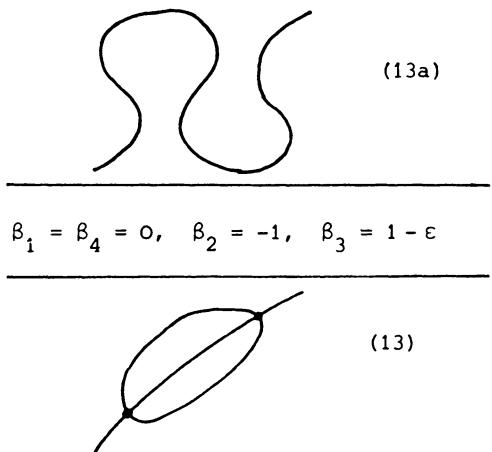
Fig.4

Let us postpone for the moment the explanation of this picture from the universal unfolding (12).

Instead we look at some typical solution curves of (13a) which occur for parameter sets $(\beta_1, \beta_2, \beta_3, \beta_4)$ close to $(0, -1, 1, 0)$. Combining these curves by the rules of Tab.1 yields the solution nets in the small pictures below. We compare these pictures with three perturbations of the situation (15) which are obtained by varying E_4 and h_4 .

Fig.5 ($h_4=2.055, E_4=2.051$)Fig.6 ($h_4=2, E_4=2.02$)

The case of Fig.6 is of course topologically equivalent to that of Fig.2. As a result, the small pictures in the figures 5, 6, 7 are the forecasts of the singularity (11) to the corresponding figures given by our system (9).

Fig. 7 ($h_4 = 1.948$, $E_4 = 1.952$)

(13)

Let us return to the case (15). The special feature of it is that the system (9) has in addition to our symmetry I the following subsymmetry:

III. Subsymmetry.

In the subspace of symmetric vectors x (i.e. $x_i = x_{8-i}$, $i=1, \dots, 7$) the system (9) is invariant under the transformation $x_i \rightarrow x_{5-i}$ ($i=1, \dots, 4$). Therefore we require $U_1(x, y, x, \alpha) = -U_1(y, x, y, \alpha)$ in (12).

Imposing this condition on (13) yields the one parameter unfolding

$$(16) \quad \begin{aligned} x^3 - y^3 - \gamma(x-y) &= 0 \\ z^3 - y^3 - \gamma(z-y) &= 0. \end{aligned}$$

The solution net of this system for $\gamma > 0$ consists of one straight line, three ellipses and one circle. These are coupled by 6 simple and 2 multiple bifurcation points as indicated in Fig. 8a.

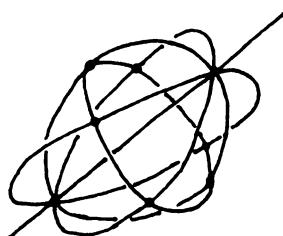


Fig. 8a

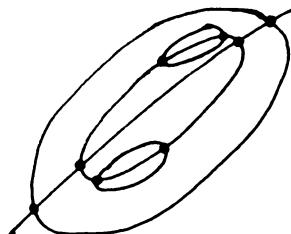


Fig. 8b

This seems to explain Fig.4. However, a close inspection of the branching structure of the system (9) with values (15) showed that in fact the topological type of Fig.8b obtains numerically. The deviations cannot be visualized in the scale of Fig.4, but we have indicated them in Fig.4 of [1].

The problem now is that Fig.8b is not possible in the partial unfolding (13). Hence we are forced to skip at least one of the simplifying assumptions I, II or III. The obvious candidate is the decoupling condition II which was only expected to hold approximately. If we impose the two symmetry conditions in the universal unfolding (12) we find the five parameter unfolding

$$\begin{aligned} x^3 - y^3 + \delta_1(x-y) + \delta_2(z-y) + (\delta_3 y + \delta_4 z + \delta_5 yz)(x-z) &= 0 \\ z^3 - y^3 + \delta_1(z-y) + \delta_2(x-y) + (\delta_3 y + \delta_4 x + \delta_5 yx)(z-x) &= 0. \end{aligned}$$

After some calculations it turns out that this system exhibits the net structure of Fig.8b for the parameter set

$$\delta_1 = -\gamma + \epsilon, \quad \delta_2 = -\epsilon \quad (0 < \epsilon \ll \gamma), \quad \delta_3 = \delta_4 = \delta_5 = 0.$$

Moreover, this situation is a perturbation of the system (16).

This ends the explanation of the situation (15) via the unfolding of the organizing center (11).

Acknowledgement: Our special thanks go to Dipl.Math. J. Bigge who did a great deal of the numerical computations for section 3. We also thank F. Nowaczeck for his help with the graphical representations.

REFERENCES

- [1] Beyn, W.-J., Defining equations for singular solutions and numerical applications. This book.
- [2] Bohl, E., Discrete versus continuous models for dissipative systems. This book.

Prof. Dr. E. Bohl, Dr. W.-J. Beyn
 Fakultät für Mathematik der Universität Konstanz
 Postfach 55 60
 D-7750 Konstanz

DISCRETE VERSUS CONTINUOUS MODELS FOR DISSIPATIVE SYSTEMS

Erich Bohl

1. INTRODUCTION

This paper is concerned with a nonlinear system of the form

$$(1a) \quad (D_0 + D)x_1 - Dx_2 = f(x_1, \lambda),$$

$$(1b) \quad -Dx_{j-1} + 2Dx_j - Dx_{j+1} = f(x_j, \lambda) \quad (j=2, \dots, N),$$

$$(1c) \quad -Dx_{N-1} + (D+D_1)x_N = f(x_N, \lambda).$$

We can regard (1) as a numerical model for the boundary value problem

$$(2) \quad -dx'' = f(x, \lambda), \quad x(0) = x(1) = 0,$$

if we put

$$(3) \quad D_0 = D = D_1 = h^{-2}d, \quad h = (N+1)^{-1}$$

in (1) where h is the stepwidth of an equidistant grid. (2) is the simplest diffusion-reaction equation, $d > 0$ is a diffusion constant and the right hand side f defines a creation term dependent on a vector λ of control parameters. Typical examples are given as follows:

$$(4) \quad f(x, \lambda) = \lambda_1(\lambda_2 - x)(1 + (\lambda_2 - x) + \lambda_3(\lambda_2 - x)^2)^{-1},$$

$$(5) \quad f(x, \lambda) = \lambda_1(\lambda_2 - x)\exp(-\lambda_3(1+x)^{-1}).$$

Here (4) describes a Michaelis-Menton process (with inhibition if $\lambda_3 > 0$).

(5) models an exothermic chemical reaction [4a,6]. From the physical interpretation the control parameters λ_i are always ≥ 0 .

(1) may also be interpreted as a mathematical model of an assemblage of N chemical cells (see Fig.1) in which a chemical reaction takes place given by the reaction term f . The cells communicate via membranes allowing for diffusive flow with diffusion constant D . In addition the two end cells are connected to an outside reservoir via a membrane characterized by the diffusion constants D_0, D_1 respectively (see Fig.1). These cell models occur quite frequently in the biological and chemical literature [5,6,7].

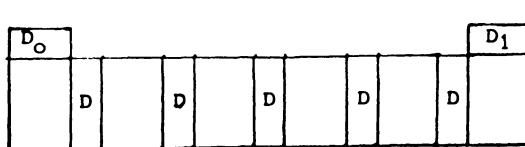


Fig.1

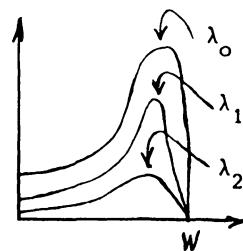


Fig.3

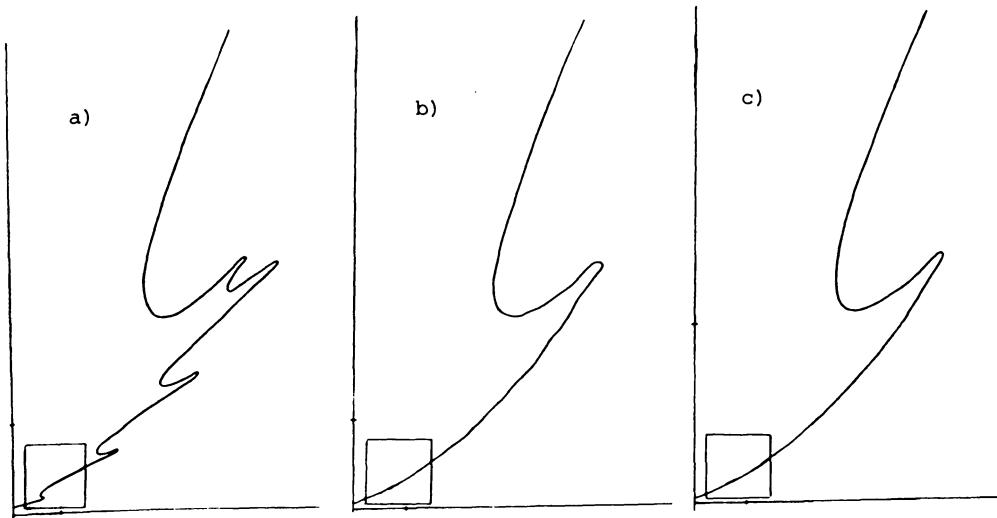


Fig.2

In this paper we are analyzing the solution set of (1). To this end we fix all control parameters except for one, which we again call λ , and try to understand the bifurcation diagram with respect to λ . It is well known that on the transition $h \rightarrow 0$ ($N \rightarrow \infty$) a fairly complicated picture for the system (1) develops into a much simpler diagram for (2). As an example we consider (2) with $d=1$ and

$$(6a) \quad f(x, \lambda) = (\lambda - x)h(x, \lambda),$$

$$(6b) \quad h(x, \lambda) = 10^3 (1 + 36(\lambda - x)) (1 + \lambda - x + 30(\lambda - x)^2 + 36(\lambda - x)^3)^{-1}.$$

Here (1), (3) yields Fig.2a for $h=.1$, Fig.2b for $h=.05$ and Fig.2c for $h=.025$.

The smoothing process on the transition $h \rightarrow 0$ is obvious from these pictures. Note that the discrete case yields more solutions which get lost on the transition to the continuous case. These solutions are frequently termed spurious solutions [8,10,11]. Indeed, they are spurious with respect to the continuous problem (2). However they are physically observed states with respect to the discrete cell system [6,7]. The occurrence of these solutions is not so much a matter of discretizing the differential equation (2), it is rather a matter of describing the physical process by a continuous or a discrete mathematical model.

In section 2 we cut our cell system in two pieces creating two separated parts of the system (1). We analyze the solution set of each of these parts and construct out of it in section 3 the solutions of the full system (1) (see also [4b] for this procedure). Finally in part 4 we apply our method to the example (2), (6) which we have discussed above. It will turn out that the solution diagram of Fig.2a is not complete. There are many more solutions to the corresponding system as we shall see in section 4.

Our study is mainly numerical work. For all our computations we assume the reaction term to be of the form (4) with λ_1 or λ_3 as continuation parameter. What we basically need is a generation term of the form given in Fig.3 which is monotone in the continuation parameter. Fig.3 assumes monotone decreasing behaviour with respect to λ .

2. SPLITTING THE CELL SYSTEM IN TWO PARTS

It is well known that the boundary value problem (2) is equivalent to the problem

$$(7) \quad -dx'' = f(x, \lambda), \quad x(0) = x'(.5) = 0.$$

Any solution \bar{y} of (7) produces a solution of (2) in the following way:

$$(8) \quad x(t) = \begin{cases} \bar{y}(t) & \text{for } t \in [0, .5], \\ \bar{y}(1-t) & \text{for } t \in [.5, 1]. \end{cases}$$

In the next section we want to use this process for the discrete problem (1). So we cut our assemblage of cells in two parts by turning off the diffusion

between two middle cells. This leaves us with one part of the system (1) which reads

$$(9a) \quad (D_0 + D)x_1 - Dx_2 = f(x_1, \lambda)$$

$$(9b) \quad -Dx_{j-1} + 2Dx_j - Dx_{j+1} = f(x_j, \lambda) \quad (j=2, \dots, M-1),$$

$$(9c) \quad -Dx_{M-1} + Dx_M = f(x_M, \lambda),$$

where $M < N$. Note that (9) may be regarded as a numerical model for the boundary value problem (7) if we define D_0 and D by (3). In this interpretation we are naturally interested in the solution set of (9) if D , D_0 and M are very big, which corresponds to the case of a very small step width h . This leads us to view our actual system (9) (with D_0 , D and M given) as a member of the family of systems which emerge if D varies from 0 to infinity while D_0 and M are fixed. If we understand this transition we have learned more about the system (9) at hand. Both ends ($D=0$, $D=\infty$) of the spectrum are easily understood:

1. $D=0$: Here all cells are separated and we only have to deal with

$$(10a) \quad D_0 x_1 = f(x_1, \lambda),$$

$$(10b) \quad 0 = f(x_j, \lambda) \quad (j=2, \dots, M).$$

We assume that we can solve (10a) for λ yielding the function $\lambda(x_1, D_0)$. Then (10) has the solution branch

$$(11) \quad (\lambda(x, D_0), x, w, w, \dots, w) \in \mathbb{R}^{M+1}, \quad x \in [0, w],$$

where w is the zero of f as indicated in Fig.3.

2. $D \rightarrow \infty$: We may assume that all solutions $(\lambda, \bar{x}) = (\lambda, \bar{x}_1, \dots, \bar{x}_M)$ of (9) satisfy the a priori bound (see Fig.3)

$$(12) \quad |\bar{x}_j| \leq w \quad (j=1, \dots, M).$$

Next we note that (9) is equivalent to the system (13), (9b), (9c) with

$$(13) \quad D_0 x_1 = \sum_{j=1}^M f(x_j, \lambda).$$

Dividing (9b), (9c) by D and observing (12) we arrive for $D \rightarrow \infty$ at $x_j = x_M$ for $j=1, \dots, M$. Therefore we learn from (13) that for fixed λ any sequence $\bar{x}(D_n)$ of solutions of (9) ($D_n \rightarrow \infty$) has a convergent subsequence which tends to a vector $(\lambda, \sigma, \dots, \sigma) \in \mathbb{R}^{M+1}$ where σ satisfies the equation

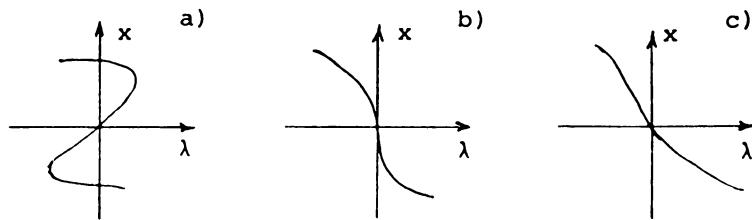


Fig.4

$$(14) \quad D_0 s = Mf(s, \lambda).$$

Conversely, if we are given a vector $(\bar{\lambda}, \bar{s}, \bar{s}, \dots, \bar{s}) \in \mathbb{R}^{M+1}$ where \bar{s} solves (14) then this vector satisfies (13) and leaves a small defect on (9b), (9c) if we divide these equations by D and consider D very big. Furthermore, the Jacobian at $(\bar{\lambda}, \bar{s}, \dots, \bar{s})$ is nonsingular and bounded independent of D if $D_0 \neq M D_s f(\bar{s}, \bar{\lambda})$ ($D_s f$ = partial derivative with respect to the first variable). Then only technical assumptions are needed to apply a local theorem [9(12.6.2)] of Newton's Method saying that in the neighborhood of our vector there is a solution of the system (9) which can be constructed as the limit of Newton's sequence. Hence, the solution branch of the scalar equation (14) defines an attractor of solutions of (9) if $D \gg 1$.

As a result our two extreme cases are characterized by the solution branches of the two scalar equations (10a) and (14). Under our qualitative assumptions on f (see Fig.3) it is reasonable to assume that the solution branch of these two scalar equations is qualitatively given by Fig.4a) or 4c). The transition $D \xrightarrow{D \rightarrow \infty}$ is given in the sequence of Fig.5 (see also [4c,d]): Here we took the more interesting case where (10a) and (14) yield a solution set qualitatively given by Fig.4a). Fig.5 is the result of numerical studies. These

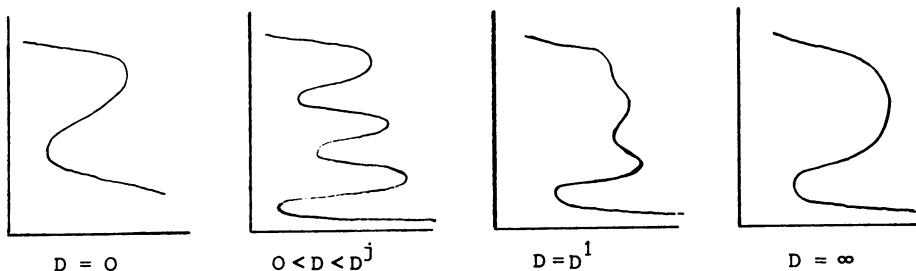


Fig.5

	1.Hyster.	2.Hyster.	3.Hyster.	4.Hyster.	$D=10^3$
λ	10.7847	50.2397	123.8050	207.1973	151.7758
x_1	3.5863	2.0363	1.4013	1.0769	3.9101
x_2	3.9844	3.8219	2.7250	2.1125	3.9130
x_3	3.9994	3.9933	3.8912	3.0841	3.9149
x_4	3.9999	3.9997	3.9958	3.9247	3.9159
μ	10.1026	38.5066	87.7250	53.8190	24.3805
y_1	3.1763	1.8563	1.3213	.8286	2.0351
y_2	3.9684	3.4137	2.5364	1.5111	2.0366
y_3	3.9988	3.9775	3.5584	2.0081	2.0376
y_4	3.9999	3.9991	3.9823	2.2740	2.0381
D^j	1.43	2.5	3.06	-	-
λ	11.8	38.3	69.5	-	-

Tab.1

reveal that for $D > 0$ the branch consists of a number of hysteresis loops (3 in Fig.5). To any of these loops (with a possible exception of one of them) there exists a critical diffusion parameter D^j at which this loop shrinks to a point \bar{z} and disappears if D exceeds D^j . Approaching D^j the two turning points forming the hysteresis converge to \bar{z} . The process is qualitatively

demonstrated by Fig.4. As an example we give in Tab.1 approximations to all turning points of (9) if f is given by (4) with $\lambda_1=10^{1.4}$, $\lambda_2=4$ and $\lambda=\lambda_3=\text{control parameter}$ and if $M=4, D_0=1$. The last two rows of Tab.1 show the critical values D^1 , D^2 , D^3 and values of λ where the corresponding hysteresis has almost disappeared. The last hysteresis in the 4th column does not disappear. It rather converges to a hysteresis whose turning points are approximately given in the last column of Tab.1. These values agree with the turning points of (14) ($D_0=1$, $M=4$) in accordance with our discussion above concerning the case $D \rightarrow \infty$ (see 2.). We have pointed out in [4d] that the sequence of Fig.4 which describes the fate of some of the hysteresis loops of our branch in the transition $D \rightarrow \infty$ is easily described by the simple algebraic equation

$$(15) \quad h(x, \lambda) + \alpha x = 0, \quad h(x, \lambda) = x^3 - \lambda$$

where α is a perturbation parameter. If α varies in the interval $[-1, 1]$, the series of pictures in Fig.4 is observed. This led us in [4d] to view (15) as a model to understand the behaviour of the assemblage of cells in the region of the critical parameter D^j . At this point singularity theory enters into the analysis of our bifurcation diagram. We go into this in [2] (see also [4d]).

3. JOINING THE TWO PARTS

In this section we are trying to find our way back to the system (1) putting together two systems of the form (9) discussed in section 2. This works along the lines in which we have constructed the solution (8) to the boundary value problem (2) out of a solution of (7) (see also [4b,c,d]). If $N=2M$ and if $(\lambda, \bar{x}_1, \dots, \bar{x}_M)$ solves (9) then the vector $(\lambda, \bar{x}_1, \dots, \bar{x}_M, \bar{x}_M, \dots, \bar{x}_1)$ obviously solves (1) for $D_o = D_1$. In general let $N=M_1+M_2$ and let $(\lambda, \bar{x}_1, \dots, \bar{x}_{M_1})$ be a solution of (9) for $N=M_1$. Similarly let $(\lambda, \bar{y}_1, \dots, \bar{y}_{M_2})$ be a solution of (9) with $N=M_2$ and $D_o = D_1$. Then the vector $(\lambda, \bar{x}_1, \dots, \bar{x}_{M_1}, \bar{y}_{M_2}, \dots, \bar{y}_1)$ leaves a defect on the system (1) whose maximum norm is given by

$$(16) \quad D|x_{M_1} - y_{M_2}|.$$

If this defect is small we can again apply the local Newton theorem [9] to construct a solution of (1). A glance at the figures of Tab.1 shows that the defect (16) may be small in the area of the vanishing hysteresis loops of the solution branch of the system of (9). Moreover, we have learned in section 2 that to any of these hysteresis loops there corresponds a critical value D^j at which the two turning points of this hysteresis loop converge. Hence if we move close to D^j and the control parameter λ between the two turning points of the corresponding hysteresis loop, then to λ there always exist three solutions on the branch which have almost identical M -th coordinates. Joining any two of these three solutions easily constructs a defect (16) which is arbitrarily small. This argument already shows (for $M_1=M_2$) that we can obtain unsymmetric and symmetric solutions to the system (1). Unsymmetric solutions have been observed in [5] for the first time (see also [6]).

Now the idea is to follow all these solutions via a continuation method on the full system (1) with respect to the control parameter λ . To get an idea of how the resulting solution diagram looks, we move the diffusion parameter D very close to but smaller than a critical parameter D^j . We construct a symmetric and an unsymmetric solution as described above. The result of a continuation method starting with these solutions is qualitatively given in Fig.6. The double-figure-eight in Fig.6b is the result of the continuation of an unsymmetric solution and the main branch is a result of the symmetric solution. Fig.6b typically results if $N=2M$ and if $D_o = D_1$. Figures 6a and 6c are pertur-

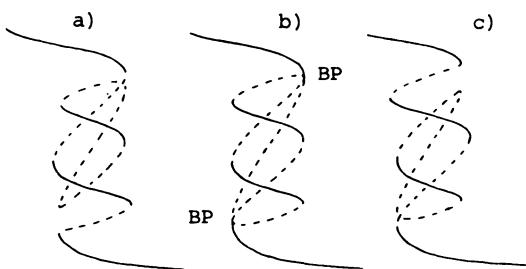


Fig.6: Only the intersections BP are bifurcation points

bations of Fig.6b which result e.g. if $D_0 \neq D_1$ (see also [2,3b]).

The solution diagram of (1) is now easily described: It is a symmetric branch with hysteresis loops accompanied by structures as shown in Fig.6 (see Fig.2 in [4b]). In the transition $D \rightarrow \infty$ all apart from pos -

sibly one of these hysteresis loops disappear and along with this process also the closed solution branches vanish so that eventually just one branch of symmetric solutions with at most one hysteresis loop is left.

Any one of the finite difference approximations of (2) of the form (1) are on this transition described above. Hence we must expect for moderate step width h that we are in a situation where structures as given in Fig.6 still appear. For h very small (but fixed) we have a system (1) where D is very big (see (3)). Hence we can expect that most of the hysteresis loops have disappeared. This suggests that the boundary value problem (2) with a right hand side qualitatively given by Fig.3 has a solution branch which shows only one hysteresis loop. The two corresponding turning points would mark ignition and extinction of the chemical process in a continuously distributed system. But our analysis also shows that in a distributed discrete system more such turning points exist, some of which even create unsymmetric solutions which are excluded in the continuous case.

4. ANALYZING AN EXAMPLE

In this section we briefly come back to the example (2), (6) with $d=1$. In the corresponding discrete equation (1) we have to put

$$(17) \quad D_0 = D_1 = D = (N+1)^2,$$

where N is the number of grid points in the interval $(0,1)$. In the introduction we already paid attention to the transition which had occurred between

the figures 2a ($N=9$), 2b ($N=19$) and 2c ($N=39$). Fig.2a shows a lot of extra hysteresis loops which are already smoothed out in Fig.2b. Let us concentrate on the first little hysteresis loop of the inset in Fig.2a in the area of $\lambda=7.8$ which does not show any more in Fig.2b. To analyze the system with the methods developed in the previous sections we note that the right hand side (6a) does not depend in a monotonic way on the control parameter λ . Therefore we fix $\lambda=7.8$ and consider the right hand side

$$(18a) \quad g(x, \mu) = (7.8-x)h_1(x, \mu),$$

$$(18b) \quad h_1(x, \mu) = 10^3(1+36(7.8-x))(1+7.8-x+\mu(7.8-x)^2+36(7.8-x)^3)^{-1}.$$

This leaves us with the analysis of the system

$$(19a) \quad ((N+1)^2 + D)x_1 - Dx_2 = g(x_1, \mu),$$

$$(19b) \quad -Dx_{j-1} + 2Dx_j - Dx_{j-1} = g(x_j, \mu), \quad (j=2, \dots, N)$$

$$(19c) \quad -Dx_{N-1} + (D + (N+1)^2)x_N = g(x_N, \mu)$$

for $\mu=30$ and $D=(N+1)^2$. Here we are in particular interested in the two cases $N=9$, $N=19$ which correspond to Fig.2a, Fig.2b respectively.

As in section 2, we now split (19) in two parts of the form (9):

$$(20a) \quad ((N+1)^2 + D)x_1 - Dx_2 = g(x_1, \mu),$$

$$(20b) \quad -Dx_{j-1} + 2Dx_j - Dx_{j-1} = g(x_j, \mu) \quad (j=2, \dots, M-1)$$

$$(20c) \quad -Dx_{M-1} + Dx_M = g(x_M, \mu),$$

$$(20d) \quad 2M = N+1 \quad (N=9 \text{ or } N=19).$$

			Indeed, a continuation process on (20) with respect
μ	31.8477	19.9353	to the first component x_1 shows for $N=9$, $D=100$ a
y_1	7.3476	6.1026	hysteresis loop in the area of $\mu=30$ whose turning
y_2	7.7768	7.7419	points are approximately given in Tab.2. For $N=19$
y_3	7.7981	7.7956	the corresponding branch shows no turning point for
y_4	7.7998	7.7996	$10 \leq \mu \leq 50$. Hence we can expect for (19) with $N=19$
y_5	7.7999	7.7999	a branch without turning points in the area of
Tab.2			$\mu=30$ in accordance with Fig.2b ($\lambda=7.8$). We now turn
Tab.2			to the case $N=9$. The turning points of Tab.2 prove
Tab.2			that the branch of (19) shows a hysteresis loop in

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
7.4826	7.7822	7.7985	7.7998	7.7999	7.7995	7.7945	7.7246	5.4365
7.4826	7.7822	7.7985	7.7998	7.7999	7.7997	7.7976	7.7694	7.1359
5.4365	7.7246	7.7945	7.7995	7.7999	7.7997	7.7976	7.7694	7.1359

Tab.3

the area of $\mu=30$. This coincides with Fig.2a for $\lambda=7.8$. However we can say more: The turning points of Tab.2 suggest that we can combine also unsymmetric solutions which create a branch separated from the main branch of the symmetric solutions given in Fig.2a. These two branches may be connected via bifurcation points or not (see Fig. 6 and [3b]). In Tab.3 we give some unsymmetric solutions for $\mu=30$ which we have calculated by the process described in section 3.

Finally let us pursue the hysteresis given in Tab.2 if D grows. The underlying system is (20) with $N=9$ (hence $M=5$). The results are given in Tab.4 which show the turning points as a function of D . Obviously, they tend to a point \bar{z} at the critical parameter $D \approx 197$ for $\lambda \approx 35.6$ (see the last column of Tab.4 and compare with our findings in section 3).

D	100	170	180	190	197
μ_1	31.8477	34.8347	34.8181	35.2905	35.6639
y_1	7.3476	7.2729	7.2505	7.2190	7.1547
y_2	7.7768	7.7635	7.7609	7.7579	7.7535
y_3	7.7981	7.7957	7.7953	7.7947	7.7941
y_4	7.7998	7.7994	7.7993	7.7992	7.7991
y_5	7.7999	7.7999	7.7999	7.7998	7.7998
μ_2	19.9353	33.1588	34.2265	35.1345	35.6626
x_1	6.1026	6.8329	6.9155	7.0090	7.1097
x_2	7.7419	7.7436	7.7450	7.7474	7.7512
x_3	7.7956	7.7937	7.7936	7.7936	7.7939
x_4	7.7996	7.7992	7.7991	7.7991	7.7991
x_5	7.7999	7.7998	7.7998	7.7998	7.7998

Tab.4

REFERENCES

- [1] Beyn, W.-J., Numerical analysis of singularities in a diffusion reaction model. To appear in the Conference Proceedings of the EQUADIFF 82.
- [2] Beyn, W.-J., Bohl, E., Organizing centers for discrete reaction diffusion models. This book.
- [3a] Bigge, J., Bohl, E., Deformations of the bifurcation diagram due to discretization. Submitted for publication.
- [3b] Bigge, J., Bohl, E., On the steady states of finitely many chemical cells. Submitted for publication.
- [4a] Bohl, E., Finite Modelle gewöhnlicher Randwertaufgaben. LAMM-51, Teubner Verlag, Stuttgart 1981 .
- [4b] Bohl, E., A numerical procedure to compute many solutions of diffusion-reaction systems. ISNM 62, 25-36, 1983.
- [4c] Bohl, E., A numerical study for a cell system. To appear in the Proceedings of the Conference "Mathematics in Biology and Medicine", Bari, 1983.
- [4d] Bohl, E., Verzweigungsbilder diskreter Transportmodelle. To appear in the Proceedings of the Conference on Numerical Methods for Differential Equations, Halle, 1983.
- [5] Bunow, B., Colton, C.K., Substrate inhibition kinetics in assemblages of cells. BioSystems 7, 160-171, 1975.
- [6] Kernevez, J.-P., Enzyme mathematics. Studies in mathematics and its applications Vol.10. Amsterdam, New York, Oxford, 1980.
- [7] Meinhardt, H., Models of biological pattern formation. Academic Press, New York, 1982.
- [8] Nussbaum, R.D., Peitgen, H.-O., Special and spurious solutions of $\dot{x}(t) = -\alpha f(x(t-1))$. Universität Bremen, Report Nr.91, 1983.
- [9] Ortega, J.M., Rheinboldt, W.C., Iterative solution of nonlinear equations in several variables. New York, San Francisco, London, 1970.
- [10] Peitgen, H.-O., Saupe, D., Schmitt, K., Nonlinear elliptic boundary value problems versus their finite difference approximations: numerically irrelevant solutions. J. reine angew. Math. 322, 74-117, 1981.
- [11] Peitgen, H.-O., Schmitt, K., Positive and spurious solutions of nonlinear eigenvalue problems. Universität Bremen, Report Nr.42, 1981.

Prof. Dr. E. Bohl
 Fakultät für Mathematik der Universität Konstanz
 Postfach 55 60
 D-7750 Konstanz

"Real" and "Ghost" Bifurcation Dynamics
in Difference Schemes for ODEs

F. BREZZI, S. USHIKI, H. FUJII

1. Introduction

The aim of this paper is to study bifurcations of dynamic behavior of solutions appearing in difference schemes for families of ODEs.

During the last decade, a great deal of attention has been paid to bifurcation problems and their numerical analysis. Especially for "stationary" bifurcation problems, a number of works have been published from various kinds of viewpoints; group-theoretic aspects [5][18] structural stability of singular points with respect to numerical perturbations [2][3][4][6], algorithmic standpoints [11] and so on.

Recently, much interest has been focused on more general bifurcation from both theoretical and numerical aspects. For instance, numerical study of systems - either of ODEs or PDEs, which may include Hopf singularities has been a subject of recent researches. In fact, a Hopf singularity yields a bifurcation of stationary solution to periodic orbits, and which is one of the simplest bifurcation phenomena next to simple stationary bifurcations. However, even for this simplest Hopf case, very little works have been done, at least to the authors' knowledge, about structure and dynamic behavior of numerical schemes, solutions of which are generally expected to "approximate" the "real" structure of continuous systems.

One aspect of numerical analysis of such problems is, certainly, to get branches of periodic solutions by, e.g., the Galerkin methods or etc. [12], where the problem is essentially regarded as a boundary-value problem in the time variable. Numerical methods which utilise the Poincaré map have been also proved successful [10][24].

Another aspect, which seems to be more fundamental from our viewpoints, is on the question which may be stated as follows. Suppose we are given a nonlinear evolution system, either a PDE, like the Navier-Stokes equation or a system of reaction-diffusion equations, or some ODE:

$$(1.1) \quad \frac{du}{dt} = f(u,s) \quad s \in \mathbb{R}.$$

Our interest is in the change of dynamic behavior of solutions of (1.1) when the "bifurcation parameter s " changes in \mathbb{R} . Emphasis should be made here that it is, in general, impossible or at least very difficult to know a priori of what kind and at what value of (u, s) there exist singularities of the system.

For a numerical analyst or an engineer who is interested in the behavior of solutions of (1.1) - either for an ODE or a PDE, he is obliged to "solve" with the aid of his computer an approximate system with discretized time in lieu of the continuous problem (1.1). However, he can have no a priori information what kind of bifurcations and where it may exist in his system - even for the simplest Hopf bifurcations. Hence, it is natural for him to pose a question "whether or not, and to what extent, does his discrete system reflect faithfully the dynamics of the continuous problem?"

To be more precise, let us suppose that the system (1.1) contains a Hopf singularity. Suppose we are given a discrete system (1.2); for instance, the simplest "explicit Euler" scheme for (1.1):

$$(1.2) \quad D_\tau u^n = f(u^n, s)$$

where

$$D_\tau u^n = (u^{n+1} - u^n)/\tau,$$

and $\tau > 0$ is the time step. Our basic question is what is the global dynamics of (1.2) for each s in \mathbb{R} . Can we detect the limit cycle bifurcating from the "approximate" Hopf singularity? Is the global dynamics of (1.2) the same, at least qualitatively, as the continuous one?

These are not at all trivial, since, as recent researches on iterations of mappings have extensively revealed (see, [9], [23] and etc.), a numerical scheme, which defines a mapping for each fixed time step, can have a variety of complex dynamics, and even bifurcation to chaotic dynamics.

The question should be singularity- and scheme-dependent. In other words, this leads us to the study of what schemes can and cannot reproduce faithfully the dynamics and the global structure of the original system, and which forms the core of our motivation in this paper.

For ODEs, an s -family of "time-periodic" orbits, called "limit cycles" appears when the bifurcation parameter s passes the Hopf

point. For the discrete system (1.2), how is the behavior of its exact solutions? Since the computation is made with a fixed $\tau > 0$, "limit cycles" in the sense of ODE can never appear.

On the other hand, the following is a classic result: see, for instance, R.J.Sacker [17] and etc.[15],[20]. The mapping $F^\tau : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$,

$$(1.3) \quad u^{n+1} = F^\tau(u^n, s)$$

where $\tau > 0$ is a fixed constant, and

$$(1.4) \quad F^\tau(u, s) \stackrel{\text{def}}{=} u + \tau f(u, s),$$

possesses an s -family of "invariant circles" whenever F^τ satisfies the Hopf condition in the mapping sense.

It is natural to expect that, as τ tends to zero, there exist invariant circles uniformly in τ , and that this τ -family of circles converges to the limit cycle of the original system. In other words, if one observes the discretized "cycle" on his graphic display, it looks like a real limit cycle notwithstanding the governing dynamics.

The known proof of the existence of invariant circles for mappings cannot be applied to our problem. The reason is that we are dealing with a τ -family of mappings, and that in the limit of $\tau \downarrow 0$, the proof appears to break down due to the degeneracy of eigenvalues. However, a careful analysis can show the uniform existence in τ of invariant circles and the convergence in the above sense. The detail including error estimates will appear in Fujii-Brezzi-Ushiki[6]. This result is summarized in Section 2.

One may believe at this point that everything goes well with the finite difference! We point out here that an important problem is hidden there. Namely, it should be recognized that the above result is a "local" one, and is valid only in a neighborhood of the origin $(u, s) = (0, 0)$ in $\mathbb{R}^n \times \mathbb{R}$ (and for sufficiently small τ).

The problem lies in the global structure of the system in the following two senses:

- [1] Suppose the value of s be fixed near the (discrete) Hopf point. How well realized is the dynamics, especially, the basin ("the attractive region") of the system?
- [2] As a global bifurcation problem in s , is the qualitative

dynamics of the system is "close" to that of the original problem ?

Here are places where "ghosts" can appear !

We shall show, taking a particular and simple system as an example, that:

however small the time mesh τ may be, a "ghost invariant circle" may appear, and the basin of the discrete system is limited to the inside of it. This ghost circle comes from $s=-\infty$ with τ .

"ghost chaos" comes from $s=+\infty$ with τ ! however small τ may be.

M.Yamaguti and his co-workers have reported in [21],[22] for the logistic equation that "chaos" in the sense of Li and Yorke [13] can appear if the time mesh τ is taken large. The second author has proved the existence of chaos in the central difference scheme of first order for the same equation, however small τ may be taken [19]. Their works can be considered to have revealed the mathematical structure of what have been thrown away into the garbage box as numerical instability by numerical analysts or engineers so far.

Our results may be regarded as a bifurcation-theoretic version of their "chaos caused by difference schemes". However, it should be noted that in our case, there is a region of the parameter s where the chaotic dynamics appears, however small τ may be chosen.

We shall discuss also in the last section about other difference schemes including the implicit Euler scheme, as well as asymmetric systems.

The following statement seems to be a broadly accepted belief: For a given system which is "structural stable" in an appropriate sense, employ a "stable" difference scheme with "sufficiently" small τ . Then, the dynamics and the structure of the system will be sufficiently well-reproduced. Viewed from our results, this belief is correct "in some sense". At the same time, they show clearly the "reality" and "dangers (or limitations)" of the word "sufficiently well".

A part of the above analyses is, of course, about a simple example. However, there is no a priori guarantee that the same phenomena can never appear in more general systems, and even in nonlinear partial differential equations in which numerical study of bifurcation phenomena is of crucial significance.

2. Hopf invariant circles

2.1 Main results

We consider in this section the family of mappings $F^\tau : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$, defined by (1.4), and study its Hopf invariant circles. Since (1.4) is the Euler approximation of the dynamical system in \mathbb{R}^n , (1.1), it is assumed in the following discussions that (1.1) has a Hopf point $(u^0, s^0) \in (0, 0) \in \mathbb{R}^n \times \mathbb{R}$. The precise form of this assumption will be stated later.

The main results of this section are summarized in the following propositions.

Proposition 2.1 There exist positive constants τ_0 , s_0 and $c_0 > 0$, and a smooth function $\delta = \delta(\tau)$, $\tau \in N_\tau^+ = [0, \tau_0]$, such that for each $\tau \in N_\tau^+$, F^τ has a Hopf bifurcation point $(0, \delta(\tau)) \in \mathbb{R}^n \times \mathbb{R}$ in the sense of mappings. Here, $\delta(\tau)$ satisfies that $|\delta(\tau)| \leq c_0\tau$. From the Hopf point $(0, \delta(\tau))$, there appears an s -branch of invariant circles, $s \in N_s^+ = [0, s_0]$, for each $\tau \in N_\tau^+$.

Proposition 2.2 This τ -family of branches of invariant circles converges to the branch of Hopf limit circles of (1.1) as $\tau \downarrow 0$, uniformly in $s \in N_s^+$.

Proposition 2.3 (Error estimate of circles) Let $r = r^0(\phi, s)$ and $r = r^\tau(\phi, s^\tau)$, $\phi \in T^1$, denote the radii of Hopf circles of, respectively, (1.1) and (1.3), where s^τ is the distance measured from $s = \delta(\tau)$, i.e. the Hopf point of F^τ . (T^1 : one-dimensional torus.) Then, we have the estimate :

$$\sup_{\phi \in T^1} |r^0(\phi, s) - r^\tau(\phi, s)| \leq \text{Const} \cdot \tau, \text{ for } s \in N_s^+ \text{ and } \tau \in N_\tau^+.$$

See, Fig. 2.1.

It is seen from Fig. 2.1 that the two surfaces made by the Hopf circles coincide together when $\tau \downarrow 0$, notwithstanding the governing dynamics which produces two circles, at least in a neighborhood of the Hopf point.

In the following, we shall give an outline of proofs. The details will be published in [6].

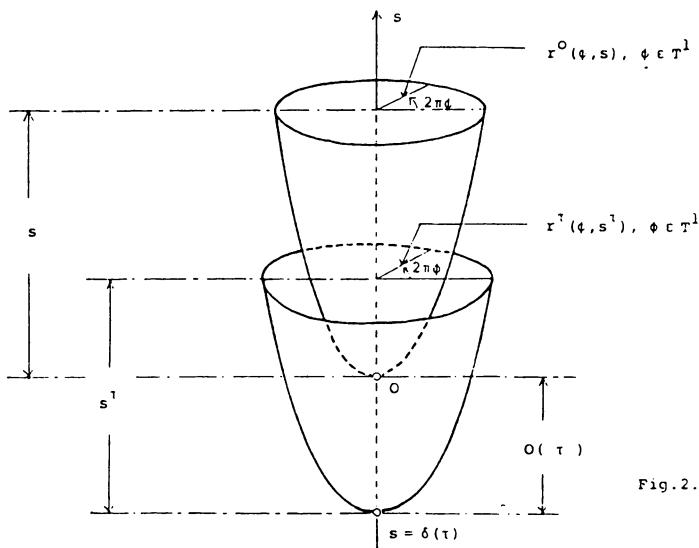


Fig. 2.1

2.2 Uniform existence of invariant circles

As stated before, the argument in this section is essentially classical, and in fact, makes use of the reduction to the Poincaré normal form as a first stage, and then of graph mapping techniques. See, e.g., [17], [14] or [9]. However, since we are working with a τ -family of mappings, we need to verify that at each stage all transformations and estimates are valid uniformly with respect to $\tau \in N_\tau^+$. The aim here is to show Prop. 2.1, but we give a slightly more details than may be necessary, to prepare basic tools for error estimates in the next subsection.

To simplify the discussion, we consider only the dynamical system on \mathbb{R}^2 :

$$(2.1) \quad \frac{du}{dt} = f(u, s), \quad (u, s) \in \mathbb{R}^2 \times \mathbb{R};$$

$f : \mathbb{R}^2 \times \mathbb{R} \rightarrow \mathbb{R}^2$ is a sufficiently smooth function, with $f(0, s) \equiv 0$ for all $s \in \mathbb{R}$.

The reduction from \mathbb{R}^n to \mathbb{R}^2 can be done with the aid of the centre manifold and its approximations. General results including such an extension are reported in [6].

The Hopf condition, which we assume throughout this section, is that there exists an $s^0 \in \mathbb{R}$ such that

$$(2.2) \quad \sigma(f_u(0,0)) = i\omega_0, \quad \bar{\sigma} = -i\omega_0 \quad (\sigma, \bar{\sigma} = \text{spectra}).$$

If $\sigma(s)$ and $\bar{\sigma}(s)$ are the eigenvalues of $f_u(0,s)$ near the origin of s , we also assume that

$$(2.3) \quad \frac{d}{ds} \operatorname{Re}\sigma|_{s=0} > 0.$$

Then, it is well-known that (1.2) has a Hopf bifurcation at $s=0$, and for $s > 0$ an invariant set $r=r^0(\phi, s)$ appears up to a suitable change of variables in polar coordinates. See, e.g., [14], [8].

We want to investigate the behavior of the simplest "explicit Euler" scheme for (2.1), which defines a τ -family of mappings (1.3). It will be convenient at least at the beginning to consider the maps F^τ for τ in some neighborhood $N_\tau = [-\tau_0, \tau_0]$ of the origin, although the interest is clearly on the case $\tau > 0$.

We identify \mathbb{R}^2 with the complex plane, and let

$$(2.4) \quad u = z\phi_0(s) + \bar{z}\bar{\phi}_0(s), \quad z \in \mathbb{C},$$

where $\phi_0(s)$ and $\bar{\phi}_0(s)$ are the eigenfunctions of $f_u(0,s)$ corresponding to, respectively, $\sigma(s)$ and $\bar{\sigma}(s)$. Then, (1.2) can be rewritten as

$$(2.5) \quad D_\tau z = \sigma(s)z + g(z, \bar{z}, s),$$

where

$$g(z, \bar{z}, s) = \langle f_1(z\phi_0(s) + \bar{z}\bar{\phi}_0(s)), \phi_0^*(s) \rangle$$

and

$$f_1(u, s) = f(u, s) - f_u(0, s);$$

$\phi_0^*(s)$ is the left eigenvector of $f_u(0, s)$ corresponding to $\sigma(s)$.

Eq. (2.5) defines a τ -family of mappings $z_s^{\tau, 0} : \mathbb{C} \times \mathbb{R} \rightarrow \mathbb{C}$,

$$(2.6) \quad z_s^{\tau, 0}(z) = \lambda^{\tau, 0}(s)z + \tau g(z, \bar{z}, s),$$

where

$$(2.7) \quad \lambda^{\tau, 0}(s) = 1 + \tau\sigma(s).$$

The next lemma shows the uniform existence of Hopf points in the

family $\{z_s^{\tau,0}\}_{\tau \in N_\tau^+}$.

Lemma 2.1 There exists a smooth function $\delta = \delta(\tau) : N_\tau \rightarrow \mathbb{R}$, such that

$$(2.8) \quad \delta(0) = 0 ; \quad |\delta(\tau)| \leq C|\tau|,$$

$$(2.9) \quad |\lambda^{\tau,0}(\delta(\tau))| = 1.$$

Proof Let

$$\begin{aligned} \gamma(\tau, s) &= (|\lambda^{\tau,0}(s)|^2 - 1)/\tau, \\ &= 2\operatorname{Re}\sigma(s) + \tau|\sigma(s)|^2. \end{aligned}$$

Then, since $\gamma(0,0) = 0$, and $(\partial\gamma/\partial s)(0,0) = 2\operatorname{Re}\sigma'|_{s=0} > 0$, by (2.2) and (2.3), there is a smooth function $\delta = \delta(\tau)$, $\tau \in N_\tau$, such that $\gamma(\tau, \delta(\tau)) = 0$ and $|\delta(\tau)| \leq C|\tau|$ by the implicit function theorem. In particular, (2.9) holds for $\tau \in N_\tau$.

Note : Here and in the following, the symbol N_τ will be a neighborhood of the origin for the variable τ , i.e., $N_\tau = [-\tau_0, \tau_0]$, where $\tau_0 > 0$ may be rechosen from time to time, which we shall not indicate explicitly. Also, we note that $N_\tau^+ = [0, \tau_0]$. The same is true for, e.g., N_s , N_s^+ and etc.

Let us introduce a δ -shift of origin into functions of s as :

$$\begin{aligned} (2.10) \quad \lambda^{\tau,\delta}(s) &= \lambda^{\tau,0}(s+\delta) [= 1 + \tau\sigma(s+\delta)], \\ g^\delta(z, \bar{z}, s) &= g(z, \bar{z}, s+\delta), \\ \text{and} \quad z_s^{\tau,\delta} &= \lambda^{\tau,\delta}(s)z + \tau g^\delta(z, \bar{z}, s). \end{aligned}$$

We use the symbols :

$$(2.11) \quad \lambda^\tau(s) = \lambda^{\tau,\delta(\tau)}(s) ; \quad g^\tau(z, \bar{z}, s) = g^{\delta(\tau)}(z, \bar{z}, s),$$

and so on. Then, we have a new τ -family of mappings :

$$\begin{aligned} (2.12) \quad z_s^\tau(z) &= z_s^{\tau,\delta(\tau)}(z) \\ &= \lambda^\tau(s)z + \tau g^\tau(z, \bar{z}, s). \end{aligned}$$

Thanks to this shift of origin, every member of the family $\{z_s^\tau(z)\}_{\tau \in N_\tau^+}$ has a common Hopf point at $(z, s) = (0, 0)$, in view of the relation (2.9), i.e., $|\lambda^\tau(0)| = |\lambda^{\tau,0}(\delta(\tau))| = 1$. It is easily seen from the construction that $z_s^\tau(z)$ is a smooth function of $\tau \in N_\tau$,

so long as $\sigma(s)$ and $g(\cdot, \cdot, s)$ are smooth enough in s .

Remark 2.1 In the classical proof, the condition $(d|\lambda^\tau|/ds)(0) > 0$ is required for a fixed τ . See, e.g., [17], [9]. However, in our case, since $(d|\lambda^\tau|/ds)(0) = \tau[\text{Re}\sigma'(0) + O(\tau)]$, the right hand side vanishes in the limit $\tau \downarrow 0$ (though it is positive whenever $\tau \in N_\tau^+$), showing a degeneracy in the eigenvalues.

The next stage is to reduce $z_s^\tau(z)$ to the Poincaré normal form.

We have the following

Lemma 2.2 (Uniform reduction to the Poincaré normal form)

Through the (τ, s) -dependent smooth change of variable :

$$\begin{aligned} z &= \zeta + x^\tau(\zeta, \bar{\zeta}, s) \\ &= \zeta + \sum_{2 \leq i+j \leq 4} x_{i,j}^\tau(s) \frac{\zeta^{i-j}}{i!j!} \quad (x_{j+1,j}^\tau(s) \equiv 0), \end{aligned}$$

which can be defined uniformly for (τ, s) near the origin, (2.12) is reduced to

$$(2.14) \quad x_s^\tau(\zeta) = \lambda^\tau(s)\zeta + \tau[\alpha^\tau(s)\zeta|\zeta|^2 + x_1^\tau(\zeta, \bar{\zeta}, s)],$$

where $\alpha^\tau(s)$ and $x_1^\tau(\zeta, \bar{\zeta}, s)$ are smooth functions of $\zeta, \bar{\zeta}, \tau$ and s , and satisfies

$$(2.15) \quad |x_1^\tau(\zeta, \bar{\zeta}, s)| \leq c|\zeta|^5, \quad \text{for } |\zeta| \in N_\zeta,$$

where $c > 0$ is independent of s and τ .

The argument is formally identical with the classical method of reduction. We have to check, however, that the transformation x^τ can be uniformly defined for $\tau \in N_\tau^+$. Let

$$(2.16) \quad g^\tau(z, \bar{z}, s) = g_{20}^\tau(s)z^2 + g_{11}^\tau(s)z\bar{z} + g_{02}^\tau(s)\bar{z}^2 + o(|z|^3).$$

Then, we find for $x_{i,j}^\tau(s)$ the formulas :

$$\begin{aligned} (2.17) \quad x_{2,0}^\tau(s) &= \frac{g_{20}^\tau(s)}{\sigma^\tau(s)[1+\tau\sigma^\tau(s)]} \\ x_{1,1}^\tau(s) &= \frac{g_{11}^\tau(s)}{\sigma^\tau(s)[1+\tau\sigma^\tau(s)]} \\ x_{0,2}^\tau(s) &= \frac{g_{02}^\tau(s)}{[2\sigma^\tau(s)-\sigma^\tau(s)+\tau[\bar{\sigma}^\tau(s)]]^2} \end{aligned}$$

and so on, where $\sigma^\tau(s) = \sigma^{\delta(\tau)}(s) = \sigma(s+\delta(\tau))$.

Since the denominators are bounded below as $s, \tau \rightarrow 0$, all of three functions are well-behaved for $(s, \tau) \in N_{s, \tau}$. The same conclusion holds for the other $x_{i,j}^\tau$'s. We note that $\alpha^\tau(s)$ is given by

$$(2.18) \quad \alpha^\tau(s) = x_{11}^\tau g_{20}^\tau + (x_{11}^\tau + \frac{x_{20}^\tau}{2}) g_{11}^\tau + \frac{1}{2} x_{02}^\tau g_{02}^\tau + \frac{1}{2} g_{21}^\tau,$$

where $x_{ij}^\tau = x_{i,j}^\tau(s)$ and $g_{ij}^\tau = g_{ij}^\tau(s)$. Hence, $\alpha^\tau(s)$ is also a smooth function in $N_{s, \tau}$.

Note that the same sequence of changes of variables for $\tau = 0$ transforms obviously (2.1) into its normal form :

$$(2.19) \quad \frac{\partial}{\partial t} \zeta = c^0(s) \zeta + \alpha^0(s) \zeta |\zeta|^2 + x_1^0(\zeta, \bar{\zeta}, s).$$

The third stage is to write $x_s^\tau(\zeta)$ in polar coordinates

$$(2.20) \quad \zeta = r e^{i2\pi\phi}, \quad x^\tau = R e^{i2\pi\phi} \quad (i = \sqrt{-1}).$$

We have that

$$(2.21) \quad R = r + \tau [s \theta_0^\tau(s) - a^\tau(s) r^2 + O(r^4)] r,$$

$$(2.22) \quad \phi = \phi + \tau [\theta_1^\tau(s) + b^\tau(s) r^2 + O(r^4)],$$

where we put

$$(2.23) \quad \begin{aligned} \theta_0^\tau(s) &= (|\lambda^\tau(s)| - 1)/\tau s, \\ \theta_1^\tau(s) &= \arg \lambda^\tau(s)/2\pi\tau, \\ a^\tau(s) &= -\operatorname{Re}[\alpha^\tau(s)/\lambda^\tau(s)] |\lambda^\tau(s)|, \end{aligned}$$

and $b^\tau(s) = \operatorname{Im}[\alpha^\tau(s)/\lambda^\tau(s)]/2\pi$.

We set now, for $s > 0$,

$$(2.24) \quad r^\tau = r_0^\tau(s)(1 + \xi s^{1/2}), \quad \text{and} \quad R^\tau = r_0^\tau(s)(1 + \Xi s^{1/2}),$$

where

$$(2.25) \quad r_0^\tau(s) = \{s \theta_0^\tau(s)/a^\tau(s)\}^{1/2},$$

and in the new variables, (2.21) and (2.22) become respectively

$$(2.26) \quad \Xi = \xi(1 - 2\tau s \theta_0^\tau(s)) + \tau s^{3/2} \Xi_1^\tau(\xi, \phi, s),$$

$$(2.27) \quad \phi = \phi + \tau [\theta_1^\tau(s) + s^{3/2} \phi_1^\tau(\xi, \phi, s)] \quad (\text{mod } 1).$$

We note that, since $\theta_0^\tau(s) = \text{Re}\sigma_0'(0) + O(|\tau|+|s|)$, $\theta_0^\tau(s) > 0$ for $(s, \tau) \in N_{s,\tau}^{++} = N_s^+ \times N_\tau^+$ from (2.3). Hence, the expression (2.25) has a meaning for $s \in N_s^+$ if $a^0(0) > 0$. We assume here that $a^0(0) > 0$, for convenience, (and consequently, $a^\tau(0) > 0$ for $\tau \in N_\tau^+$); if $a^0(0) < 0$, nothing changes, except that one has to choose $s < 0$ instead of $s > 0$ in what follows; the case $a^0(0) = 0$ corresponds to a higher degeneracy, and we are not going to consider this case here. (See, [9].)

In what follows, we use the smoothness that Ξ_1^τ and ϕ_1^τ are in $C^{0,1}$ with respect to (ξ, ϕ) uniformly in $\tau \in N_\tau$, $s \in N_s$, and which is the case if $f \in C^k$ ($k \geq 6$). (See., also Remark 2.2 below.)

We are going to look for an invariant manifold of the form $\xi = w^\tau(\phi) \in W$, where W is the complete metric space

$$(2.28) \quad W = \{w : T^1 \rightarrow \mathbb{R} ; |w(\phi)| \leq 1, |w(\phi) - w(\phi')| \leq |\phi - \phi'| \},$$

with

$$(2.29) \quad \text{dist}(w, w') = \|w - w'\| = \sup_{\phi \in T^1} |w(\phi) - w'(\phi)|.$$

Since for any $w \in W$,

$$(2.30) \quad \phi_w^\tau(\phi, s) = \phi + \tau[\theta_1^\tau(s) + s^{3/2}\phi_1^\tau(w, \phi, s)]$$

is a bi-lipschitz homeomorphism of T^1 onto itself, one can associate $\psi_w^\tau(\phi, s) = (\phi_w^\tau)^{-1} : T^1 \rightarrow T^1$, uniformly with respect to s and τ . Now, as in [17], [9], one can show that for $s \in N_s^+$,

$$(2.31) \quad \begin{aligned} \mathcal{F}^\tau(w)(\phi) &= (1 - 2\tau s \theta_0^\tau(s))w(\phi) + \tau s^{3/2}\Xi_1^\tau(w(\phi), \phi, s), \\ \phi &= \psi_w^\tau(\phi, s), \end{aligned}$$

defines a mapping $\mathcal{F}^\tau : W \rightarrow W$, uniformly for $\tau \in N_\tau^+$. We have that \mathcal{F}^τ is contractive, i.e.,

$$(2.32) \quad \|\mathcal{F}^\tau(w) - \mathcal{F}^\tau(w')\| \leq \epsilon^\tau \|w - w'\|,$$

with

$$(2.33) \quad 0 < \epsilon^\tau \leq 1 - C_0 \tau s,$$

where $C_0 > 0$ is independent of s and τ . See, [6] for details. We can now conclude the existence of a unique fixed point in W , for each $\tau \in N_\tau^+$, of $\mathcal{F}^\tau : \hat{w}^\tau = \mathcal{F}^\tau(\hat{w}^\tau)$. Hence, the corresponding radius:

$$(2.34) \quad r^\tau(\phi, s) = r_0(s)[1 + s^{1/2} \hat{w}^\tau(\phi, s)].$$

Remark 2.2 (Regularity of the fixed point \hat{w}^τ) Similar arguments (see, [9]) show that actually the fixed point \hat{w}^τ belongs to $W_\ell = \{w \in C^{\ell,1} \mid \|w\|_{\ell,1} \leq 1\}$ ($\ell \geq 0$), for each $\tau \in N_\tau^+$, if Ξ_1^τ and $\phi_1^\tau \in C^{\ell,1}$ with respect to (ξ, ϕ) uniformly in (s, τ) , which is the case if $f \in C^{\ell+6}$.

2.3 τ -limit of the circles and error estimates

We show Props. 2.2 and 2.3. We assume the fixed point $\hat{w}^\tau = \mathcal{F}^\tau(\hat{w}^\tau)$ belongs to W_1 in the following. (See, Remark 2.2.) For this, the following lemma provides the key estimate. Let $0 < \tau' < \tau$ (to fix the idea), and $\hat{w}^{\tau'}$ be the fixed point: $\hat{w}^{\tau'} = \mathcal{F}^{\tau'}(\hat{w}^\tau)$, for $\tau' \in N_\tau^+$.

Lemma 2.3 (Estimate of "truncation error")

$$(2.35) \quad \sup_{\phi \in T^1} \left| \frac{\mathcal{F}^\tau(\hat{w}^{\tau'}) - \hat{w}^\tau}{\tau} (\phi) \right| \leq C(\tau - \tau'), \quad \text{for all } s \in N_s^+,$$

where $C > 0$ is independent of s and τ .

Outline of the proof. We write $w \equiv \hat{w}^{\tau'}$, for simplicity. First,

$$\frac{\mathcal{F}^\tau(w) - w}{\tau} (\phi) = \frac{\mathcal{F}^\tau(w) - w}{\tau} (\phi) - \frac{\mathcal{F}^{\tau'}(w) - w}{\tau'} (\phi)$$

since the second term vanishes identically, and by definition,

$$(2.36) \quad \begin{aligned} &= [(1-2\tau\theta_0^\tau)w(\phi^\tau) - w(\phi) + \tau s^{3/2} \Xi_1^\tau(w(\phi^\tau), \phi^\tau, s)]/\tau \\ &\quad - [(1-2\tau'\theta_0^{\tau'})w(\phi^{\tau'}) - w(\phi) + \tau's^{3/2} \Xi_1^{\tau'}(w(\phi^{\tau'}), \phi^{\tau'}, s)]/\tau', \\ &= \left\{ \frac{w(\phi^\tau) - w(\phi)}{\tau} - \frac{w(\phi^{\tau'}) - w(\phi)}{\tau'} \right\} - 2\{\theta_0^\tau w(\phi^\tau) - \theta_0^{\tau'} w(\phi^{\tau'})\} \\ &\quad + \{s^{3/2} [\Xi_1^\tau(w(\phi^\tau), \phi^\tau, s) - \Xi_1^{\tau'}(w(\phi^{\tau'}), \phi^{\tau'}, s)]\}, \end{aligned}$$

where $\phi^\tau = \psi_w^\tau(\phi)$ and $\phi^{\tau'} = \psi_w^{\tau'}(\phi)$.

We note that for a given $w \in W_1$, ψ_w^τ is in $C^{1,1}$ with respect to $\tau \in N_\tau$, due to the smoothness $\theta_0^\tau(s) \in C^{1,1}$ in (s, τ) , and $\phi_1^\tau(\xi, \phi, s) \in C^{1,1}$ with respect to ξ, ϕ, τ and s . Consequently, we conclude that $w(\tau) = [w(\phi^\tau) - w(\phi)]/\tau$ is of $C^{0,1}$ with respect to τ , noting that $\psi_w^0(\phi) = \phi$. Hence, the first bracket $\{\cdot\}$ of the above expression is bounded by $C(\tau - \tau')$, for some positive constant independent of $s \in N_s^+$. To show (2.35), it is enough to bound the second and the third brackets in (2.36), by $C(\tau - \tau')$, which is an easy matter to show, for small $s \in N_s^+$.

Now, since

$$\begin{aligned} \|\hat{w}^\tau - \hat{w}^{\tau'}\| &\leq \|\mathcal{J}^\tau(\hat{w}^\tau) - \mathcal{J}^{\tau'}(\hat{w}^{\tau'})\| + \|\mathcal{J}^{\tau'}(\hat{w}^{\tau'}) - \hat{w}^{\tau'}\| \\ &\leq \epsilon^\tau \|\hat{w}^\tau - \hat{w}^{\tau'}\| + C_\tau(\tau - \tau') \end{aligned}$$

from (2.32) and (2.35), we have finally that

$$(2.38) \quad \|\hat{w}^\tau - \hat{w}^{\tau'}\| \leq \frac{C_0}{s} (\tau - \tau'), \quad \text{for } 0 < \tau' < \tau,$$

in view of (2.33). Hence, for the corresponding invariant sets r^τ and $r^{\tau'}$, we have the bounds :

Lemma 2.4

$$(2.39) \quad \|r^\tau(\cdot, s) - r^{\tau'}(\cdot, s)\| \leq C(\tau - \tau'), \quad \text{for all } s \in N_s^+.$$

Our strategy is to let $\tau' \downarrow 0$ in the above estimate to have Prop. 2.3. In fact, since we can show that $\{\mathbf{d}r^\tau/\mathbf{d}\phi\}_{\tau>0}$ also makes a Cauchy sequence in $\|\cdot\|$ norm, we have the existence of the limit function $r^0(\phi, s)$ which belongs to C^1 with respect to $\phi \in T^1$. It can be shown that the limit function satisfies

$$\begin{aligned} r^0(\phi, s) &= r_0^0(s)[1 + \hat{w}^0(\phi, s)s^{1/2}], \\ \text{and} \quad (2.40) \quad \frac{d\hat{w}^0}{d\phi} &= \frac{-2s\theta_0^0(s) + s^{3/2}\hat{z}_1^0(\hat{w}^0(\phi), \phi, s)}{\theta_1^0(s) + s^{3/2}\phi_1^0(\hat{w}^0(\phi), \phi, s)}. \end{aligned}$$

Hence, $r^0(\phi, s)$ is the invariant set produced by the Hopf limit circle of ODE (2.1).

Now taking the limit $\tau' \downarrow 0$, we have the desired estimate Prop. 2.3.

3. "Ghost" invariant circles - coming from the infinity even when small enough

In this and the subsequent sections, we examine a particular dynamical system as an example. The object here is to show rather surprising results that even in such a simple example, the behavior of the discretized dynamical system can be qualitatively quite different from the original system. Of course, since the discretized version defines an iteration of mappings, while the ODE defines a continuous dynamical system, this should be done under an appropriate interpretation.

What we shall show first is that in the discretized version, a "ghost" invariant circle may appear, and which may substantially deform the "domain of attraction" of the system. The domain of attraction consists of those points of the phase space which are attracted by an attractor.

The system we shall study is described as an ODE on the complex plane :

$$(3.1) \quad \dot{z} = z(i + s - z\bar{z}).$$

The system (3.1) is the simplest family of equations which has a Hopf singularity. In fact, (3.1) has a Hopf point at $s=0$, and for all $s > 0$, there is a branch of periodic orbits. The domain of attraction is infinitely large. See, Fig.3.1 for the global dynamics of (3.1).

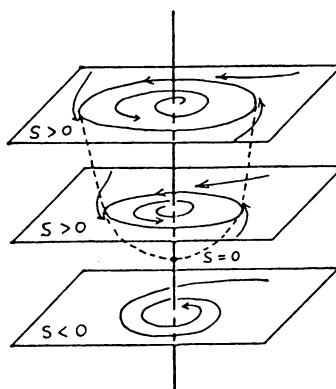


Fig.3.1

The Euler's finite difference scheme for (3.1) is :

$$(3.2) \quad z_{n+1} = z_n + \tau z_n (i + s - z_n \bar{z}_n).$$

The scheme (3.2) induces a one dimensional mapping in the radius $r = |z|$ of the polar coordinate:

$$(3.3) \quad r_{n+1}^2 = r_n^2 (\tau^2 + (1 + \tau(s - r_n^2))^2).$$

By applying the usual arguments for the dynamics of one dimensional mappings, the iteration (3.3) can have, in general, very complicated behavior.

Let us examine the fixed points of (3.3). These fixed points corresponds to invariant circles of the scheme (3.2). For $s > 0$, the continuous problem (3.1) has a limit cycle. Hence there should be an invariant circle for discretized version (3.2), at least for sufficiently small time step τ , which bifurcates from the origin as the parameter s passes a critical value, say s_τ , near the critical value $s = 0$ of the Hopf bifurcation for the ordinary differential equation. The solutions of equation

$$(3.4) \quad r^2 = r^2 (\tau^2 + (1 + \tau(s - r^2))^2)$$

are given by

$$(3.5) \quad R_-^2 = s + (1 - \sqrt{1 - \tau^2})/\tau$$

and

$$(3.6) \quad R_+^2 = s + (1 + \sqrt{1 - \tau^2})/\tau.$$

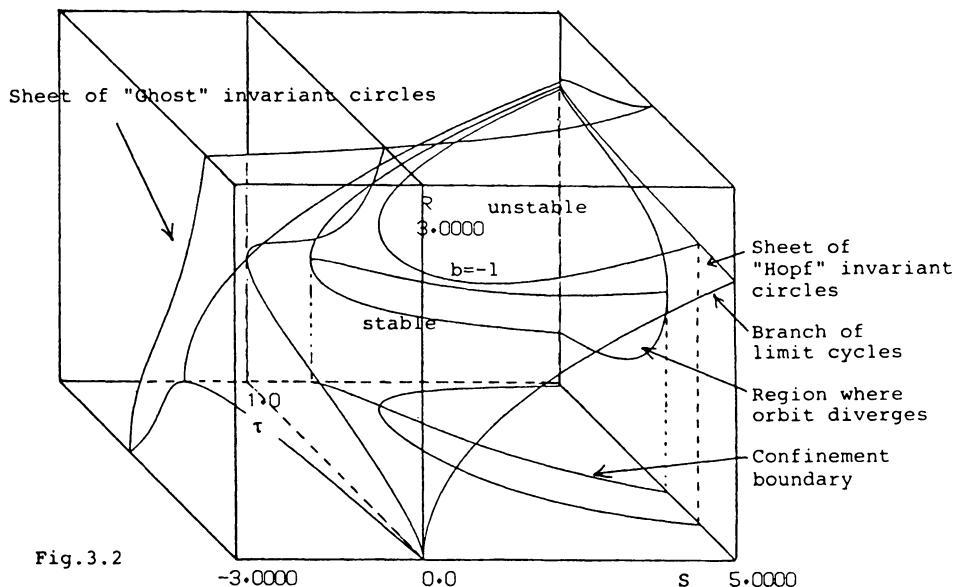
The fixed point R_- above corresponds to the bifurcated invariant circle for the original Hopf bifurcation. But the other fixed point R_+ of the iteration, which corresponds to an invariant circle for the discretized problem (3.2), does not represent any invariant circle for the ordinary differential equation (3.1). The radius of this "ghost" invariant circle tends to the infinity when τ tends to zero.

Outside this "ghost" invariant circle, all the orbits diverge to the infinity. The basin of attraction of the invariant circle

with radius R_- , which is attracting for small enough τ , is hence finite. There exists nothing that corresponds to the "ghost" invariant circle in the dynamical system defined by the ordinary differential equation (3.1).

These invariant circles exist when τ is between zero and the unity (we consider only the case of positive time step τ). More precisely, the branch R_- , which corresponds to the limit cycle bifurcating from the origin by the Hopf bifurcation, exists when τ is between zero and the unity and s is bigger than $-(1 - \sqrt{1 - \tau^2})/\tau$. And the branch R_+ of ghost invariant circles exists when τ is between zero and the unity and s is bigger than $-(1 + \sqrt{1 - \tau^2})/\tau$.

If the time mesh τ is larger than one, there exists no invariant circle for the discretized problem for any value of the parameter s . In this case, all points of the phase plane except the origin diverge to the infinity as we iterate the difference scheme (3.2). There exists no basin of attraction for (3.2).



The sheet of invariant circles is depicted in Fig.3.2. The sheet is two folded, one of which, the lower sheet, comes to coincide the Hopf branch for the continuous problem as τ tends to zero, and the other, the upper sheet, corresponds to the "ghost"

invariant circles. These two sheets are smoothly connected along a parabola contained in the plane $\tau=1$.

4. "Ghost" chaotic dynamics - coming from the infinity, even for small enough τ .

Now let us examine the stability of these invariant circles. By setting $q = r^2$, we have the iteration on the real half line $q \geq 0$:

$$(4.1) \quad q_{n+1} = q_n(\tau^2 + (1 + \tau(s - q_n))^2).$$

Differentiate this formula and evaluate it at the fixed points. We get

$$(4.2) \quad 3 \pm 2\tau\sqrt{1-\tau^2} (s + 1/\tau) - 2\tau^2.$$

Solving this with respect to s we get

$$(4.3) \quad s = \pm (b - 3 + 2\tau^2)/(2\tau\sqrt{1-\tau^2}) - 1/\tau,$$

where b denotes the value of the differential at the fixed point. This formula gives the contour curves of the eigenvalues at the fixed points in the (s, τ) plane for each value b . The curve (4.3) for $b = -1$ is the boundary of stability of the fixed points on the branches. Observe that for fixed τ between zero and the unity, the bifurcated branch R_- becomes unstable for sufficiently large s . In Fig.3.2 the curve (4.3) for $b = -1$ is drawn in the (s, τ) plane. This curve is also drawn in the same Figure as a curve in the lower sheet of the invariant circles.

Next let us consider the confinement for the discretized problem. Since orbits starting from a point outside the "ghost" invariant circle diverge to the infinity, we look for the values of parameters s and τ for which there exists a bounded region containing the origin such that if the initial point is taken in this region then the iterated image of this point never escape from the region.

The critical situation is described as follows. Suppose the situation where the invariant "ghost" fixed point is kept at a fixed

value and vary τ from zero to one, the value s being determined by (3.6). The graph of mapping (4.1) is monotone increasing for sufficiently small τ . For larger value of τ , there appears a hump between zero and the point corresponding to the Hopf branch. If the radius R_+ of the "ghost" invariant circle chosen is sufficiently large, then the hump grows tall enough to exceed the radius of the "ghost" circle. At the moment when the hump grows out the "ghost" circle, the confinement breaks down (see Fig.4.1 below).

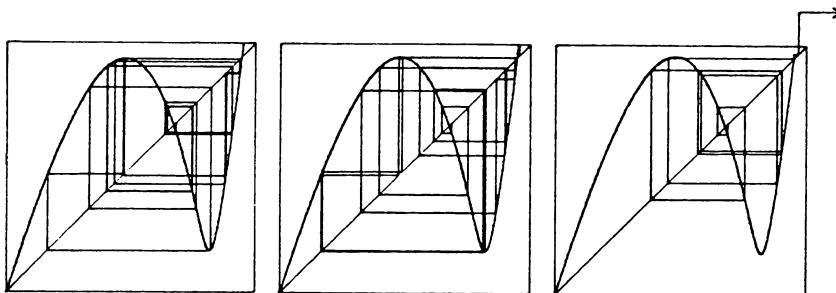


Fig.4.1

Let us compute more precisely. We look for the inverse image of the "ghost" circle R_+ of the mapping (4.1). Using the fact that R_+ and R_- given by (3.5) and (3.6) are the roots of the equation (3.4), the points in the inverse image of R_+ which are different from this fixed point, are the solutions of the quadratic equation in $q = r^2$:

$$(4.4) \quad q^2 - q(s + (1 - \sqrt{1 - \tau^2})/\tau) + 1/\tau^2 = 0.$$

The condition for this equation to have real positive roots is given by

$$(4.5) \quad s \geq (1 + \sqrt{1 - \tau^2})/\tau.$$

When (4.4) has double roots, the confinement becomes critical. At this moment, the radius of the circle mapped onto the "ghost" invariant circle is

$$(4.6) \quad r = \sqrt{q} = 1/\sqrt{\tau}.$$

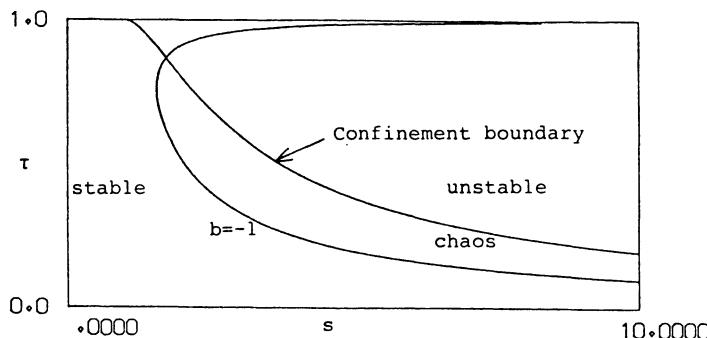


Fig.4.2

In Figure 4.2, two curves in the (s, τ) plane are drawn. One of which is the curve where the lower sheet of invariant circle becomes unstable by the eigenvalue of Jacobian getting smaller than -1. The other curve represents the points where the confinement condition breaks down. The region of initial points which are mapped by (4.1) into the region outside the "ghost" invariant circle is also plotted in Fig.3.2. The orbits starting from this region or passing it diverge to the infinity.

We shall prove the following proposition.

Proposition 4.1.

There exists a positive τ_0 such that for any positive τ smaller than τ_0 , the discretized mapping (3.2) is chaotic for some values of the parameter s , in the following sense: there exists a periodic point of period three for the iteration (4.1) so that (4.1) satisfies Li-Yorke's condition for the existence of chaotic scrambled set.

Proof.

We denote the mapping (4.1) by $f_\tau = f_\tau(q)$. Set $\tau_0 = 0.5$. Let τ be a value in the interval $(0, \tau_0)$. Take $s = (1 + \sqrt{1 - \tau^2})/\tau$. Then

$$(4.7) \quad Q_+ = R_+^2 = s + (1 + \sqrt{1 - \tau^2})/\tau = 2(1 + \sqrt{1 - \tau^2})/\tau$$

is the fixed point of f_τ . The interval $[0, Q_+]$ is mapped onto itself. The point $S=1/\tau$ is mapped into Q_+ . The mapping f_τ is monotone increasing on the interval $[S+1/\tau, Q_+]$. We see that $f_\tau(S+1/\tau) = S\tau^2 + \tau$ is smaller than S .

Let $I_1 = [S+1/\tau, Q_+]$ and $I_0 = f_\tau^{-1}(I_1) \cap I_1$. The image $f_\tau(I_1)$ includes the interval $I_2 = [1/\tau, S+1/\tau]$. And the image $f_\tau(I_2)$ includes I_0 . Hence there exists a periodic point of period three in I_0 , whose orbit visits $I_1 - I_0$ and I_2 .

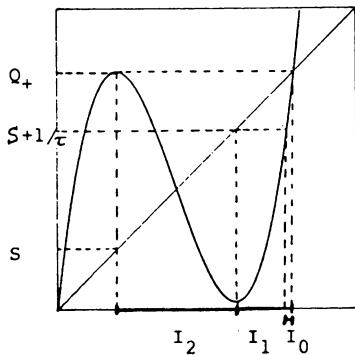


Fig.4.3

We have seen in the preceding that the discretized problem may have "ghost" invariant circles, which appear from the infinity, besides the Hopf invariant circle. The Hopf invariant circle produced in the neighborhood of the origin bifurcating from the stationary solution may become unstable if the time mesh τ is too large. Of course, if the time mesh τ is much larger, for example larger than one, then the behavior of the discretized numerical solution is completely different from the original problem. In fact, for this case, the discretized version has no invariant circle for any value of s whereas the continuous problem possesses an invariant limit circle for positive s .

But we proved that even for small τ , the bifurcated invariant circle becomes unstable if we execute the computation for large values of the bifurcation parameter s . If the taken time step mesh is very small, the range of the parameter values s is large. But if the time mesh is not very small, the parameter range in which the

numerical invariant circle exists and is stable becomes small.

The invariant circle breaks in two ways. It becomes unstable as described by (4.3), when the eigenvalue of the Jacobian at the fixed point becomes less than -1. Another way is that it disappears by corrison with the ghost solution when τ exceeds one.

When the invariant circle becomes unstable, there appear, in general, two invariant circles. To be more precise, there appears period two periodic orbit for the iteration (4.1) bifurcating as a secondary bifurcation from the Hopf branch; this periodic orbit corresponds to two circles in the plane of (3.2). The orbit starting from a point on one of these circles goes around the bifurcated invariant circles in a quasi-periodic manner visiting the two circles alternatively. These two circles compose an asymptotically stable invariant set.

If one follows a sequence of such numerical experiments, one observes a tertial bifurcation. The attractive invariant set of two circles becomes unstable and bifurcates into an attracting set consisting of four circles. A cascade of such "period doubling" series of bifurcations is expected. However, the iteration studied here is not uni-modal, the situation is not too simple. As is seen in the Figures the period doubling bifurcation does not, in general, continue to the infinite-th bifurcation.

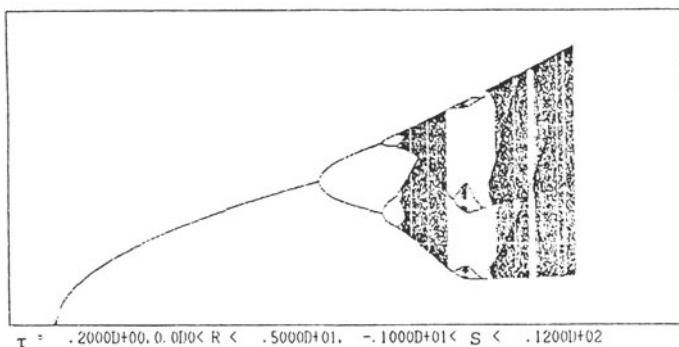


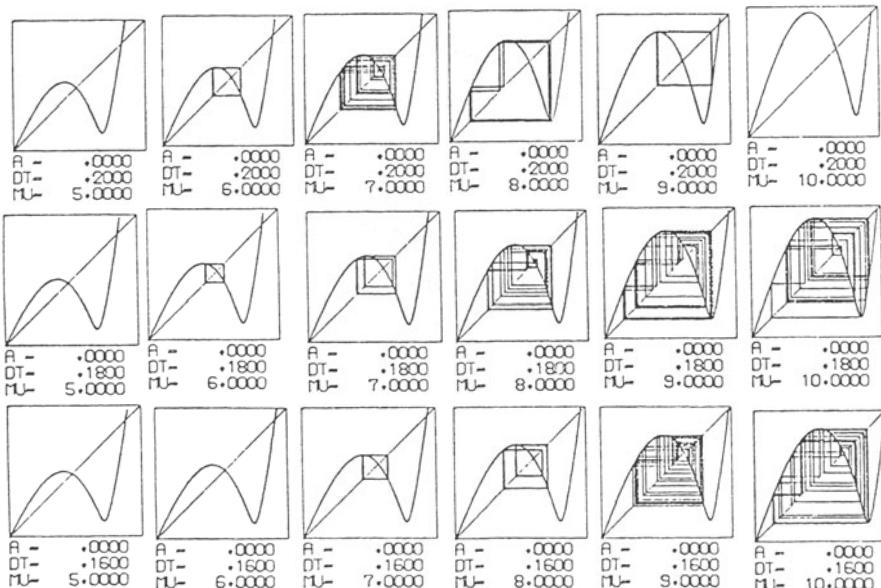
Fig.4.4

Figure 4.4 shows how the bifurcation proceeds as we vary the parameter s , for fixed time step T . In this Figure, the abscissa

represents the bifurcation parameter s and the ordinate represents the radius r . For each value of s , 200 points on an orbit are plotted. The cascade of period doubling bifurcations can be observed.

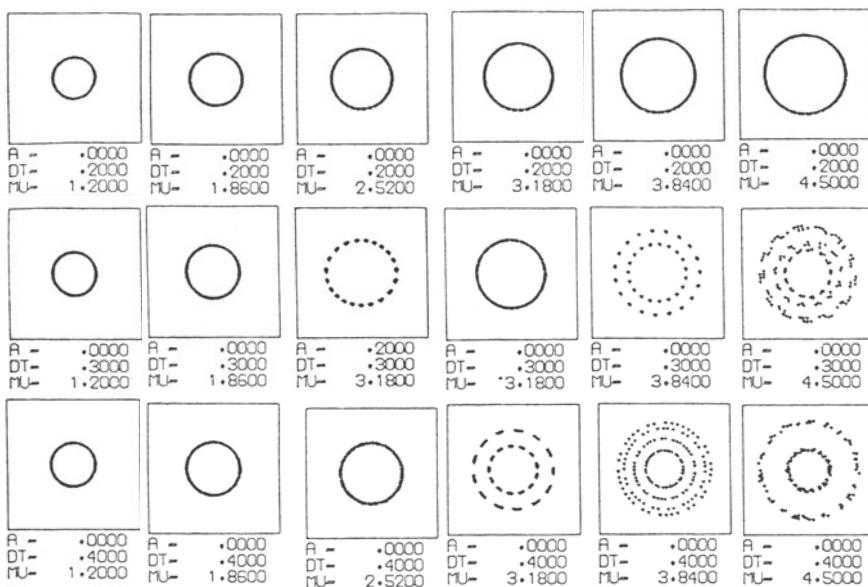
As can be imagined, such bifurcation seems to need at least two (or more) dimensional family to describe the global bifurcation diagram. The global bifurcation scheme does not have natural one dimensional structure.

Anyway, we proved in the proposition, that there are values of τ and s such that the dynamical system (4.1) has a region of confinement and that there is a chaos in it. Figure 4.5 is the plot of the graph of (4.1) for some values of τ and s . The orbit of the mapping is also plotted in the Figures. Figure 4.6 is the plot of some orbits in the phase space z of the mapping (3.1).



(Note : $MU \equiv s$, $DT \equiv \tau$)

Fig.4.5

Fig.4.6 (Note : MU \equiv s, DT \equiv τ)

5. Concluding Discussions

We give some comments. Firstly, since we studied the bifurcation of (3.2), and that this system is rotationaly symmetric the bifurcation is understood relatively well by examining the one dimensinal map (4.1). In general, however, the Hopf bifurcation to be studied has no rotational symmetry. As an example we show a numerical experiment for non-symmetric Hopf bifurcation. There, a skew term

$$\dot{x} = ax, \quad \dot{y} = -ay$$

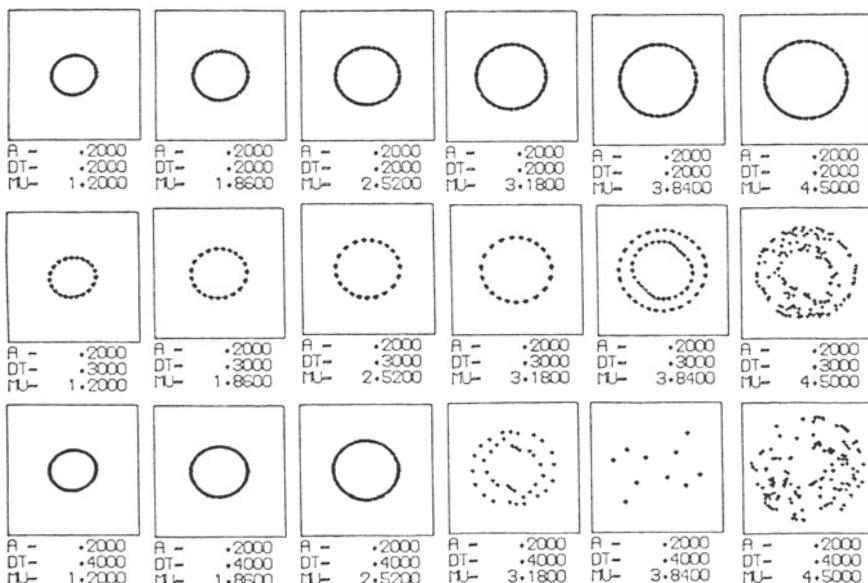
is added to (3.1). In this case, even the unstable inavariant circle can disappear. Figure 5.1. is the similar plot for non-zero skew parameter a.

Secondly, for the explicit Euler's scheme, the breakdown of the stable limit cycle into periodic or chaotic orbit is observed. What can happen, if we employ the implicit Euler's scheme, then ?

The implicit Euler's scheme does not determine the orbit uniquely. Now to fix the idea, suppose we choose the $n+1$ -th point of the orbit for the implicit scheme the nearest point from the n -th step. In this case chaotic phenomena is very difficult to occur. We observed that at least, period doubling from a fixed point can occur for one dimensional iterations.

Acknowledgements

This work has been done when the first author was visiting Kyoto University. He would like to express his thanks to Professor M. Yamaguti of Kyoto University for his hospitality and valuable discussions.



(Note : $MU \equiv s$, $DT \equiv \tau$)

Fig.5.1

REFERENCES

- [1] C.Bernardi, Approximation numériques d'une bifurcation de Hopf, C.R.Acad. Sc. Paris, t.292, série I, 103-106, 1981.
- [2] F.Brezzi, Approximations of non-linear problems, Lectures on the Numerical Solutions of Partial Differential Equations — Proc. of the Special Year in Numerical Analysis, Eds. I.Babuska, T.-P.Liu, J.Osborn, Univ. Maryland, 1981.
- [3] F.Brezzi-H.Fujii, Numerical imperfections and perturbations in the approximation of nonlinear problems, MAF ELAP IV, J.Whiteman ed., 431-452, Academic Press 1982.
- [4] F.Brezzi-J.Rappaz-P.A.Raviart, Finite dimensional approximations of nonlinear problems, II, Numer. Math. 37, 1981, 1-28; III, ibid. 38, 1981, 1-30.
- [5] H.Fujii, Numerical pattern formation and group theory, Computing Methods in Applied Science and Engineering, Eds. R.Glovinski, J.L.Lions, North Holland 1980.
- [6] H.Fujii-F.Brezzi-S.Ushiki, in preparation.
- [7] H.Fujii-M.Yamaguti, Structure of singularities and its numerical realization in nonlinear elasticity, J.Math.Kyoto Univ. 20 (1980).
- [8] B.D.Hassard-N.D.Kazarinoff-Y.-H.Wan, theory and applications of Hopf bifurcation, Cambridge Univ. Press 1981.
- [9] G.Iooss, Bifurcation of maps and applications, North-Holland 1979.
- [10] H.Kawakami, to appear in IEEE Trans. Circuits and Systems.
- [11] H.B.Keller, Constructive methods for bifurcation and nonlinear eigenvalue problems, Computing Methods in Applied Science and Engineering, Lec. Notes in Math. 704, 1979, Springer-Verlag.
- [12] W.F.Langford, Numerical solution of bifurcation problems for ordinary differential equations, Numer. Math. 28, 1977, 171-190.
- [13] T.Li-J.Yorke, Period three implies chaos, Amer Math. Monthly, 82, 1975, 985-992.
- [14] J.Marsden-M.F.McCraken, The Hopf bifurcation and its applications, Springer-Verlag 1975.
- [15] J.Neimark, On some cases of periodic motions depending on parameters, Dokl. Akad. Nauk. SSSR, 1959, 736-739.
- [16] M.Prüfer, Turbulence in multistep methods for initial value problems, preprint.
- [17] R.J.Sacker, On invariant surfaces and bifurcation of periodic solutions of ordinary differential equations, New York University, Report IMM-NYU 333, 1964; Comm. Pure Appl. Math. 18, 4 1965, 717-

732.

- [18] D.H.Sattinger, Group theoretic methods in bifurcation theory, Springer Lecture Notes in Math., n°762, 1979.
- [19] S.Ushiki, Central difference scheme and chaos, Physica 4D, 1982, 407-424.
- [20] Y.H.Wan, Bifurcation into invariant tori at points of resonance, Arch. Rational Mech. Anal. 68, 1978, 343-357.
- [21] M.Yamaguti-H.Matano, Euler's finite difference scheme and chaos, Proc. Jap. Acad. 55, 1979, 78-80.
- [22] M.Yamaguti-S.Ushiki, Chaos in numerical analysis of ordinary differential equations, Physica 3D, 3, 1981.
- [23] M.Yamaguti-S.Ushiki, Discréétisation et chaos, C.R.Acad.Sci. Paris, 290, 1980, 637-640.
- [24] N.Yamamoto, J.Math. Tokushima Univ, 16, 1982, 55-126.

F. Brezzi, Instituto di Analisi Numerica del CNR, Pavia, Italy

S. Ushiki, Department of Mathematics, Yoshida College, Kyoto University
Kyoto, Japan

H. Fujii, Institute of Computer Sciences, Kyoto Sangyo University,
Kyoto, Japan

THE USE OF EXTRAPOLATION FOR THE SOLUTION OF BIFURCATION PROBLEMS

R. Caluwaerts

The solution of a bifurcation problem is often obtained by replacing the original equation by an extended problem. An asymptotic expansion is derived for the discretization error of the approximate solution of the extended system.

1. Introduction

A family of non-linear equations, depending on a real parameter λ is given :

$$F(x, \lambda) = 0 \quad (1.1)$$

x belongs to some Banach space X and F is a non-linear operator from $X \times \mathbb{R}$ into a Banach space Y . One has a bifurcation problem when there are values of λ for which the Fréchet derivative with respect to x is singular. An important class of methods, for finding such critical values of λ , are the so called direct methods. In a direct method one usually replaces equation (1.1) by an extended system $H(y) = 0$ ($y = (x, \lambda, \phi, \dots)$) which has (x^*, λ^*, \dots) as a unique solution. For different types of problems different kinds of extended systems can be defined.

We will always assume that :

$$N(F'_x(x^*, \lambda^*)) = \text{span}(\phi); \phi \in X \quad (1.2)$$

$$N(F'^*_x(x^*, \lambda^*)) = \text{span}(\psi); \psi \in Y^* \quad (1.3)$$

and Range $(F'_x(x^*, \lambda^*))$ is closed.

Definition 1. (x^*, λ^*) is a turning point if and only if

$$\psi F'_{\lambda}(x^*, \lambda^*) \neq 0 \quad (3.4)$$

Definition 2. (x^*, λ^*) is a simple turning point if (1.4) is satisfied and if

$$C = \frac{1}{2} \psi F''_{xx}(x^*, \lambda^*) \neq 0 \quad (1.5)$$

G. Moore and A. Spence defined an extended system for simple turning points [7], for non-simple turning points this was done by D. Roose, R. Piessens and R. Caluwaerts [9,8,4]. Both methods can be proved to be regular.

Definition 3. $(\overset{\star}{x}, \overset{\star}{\lambda})$ is a bifurcation point if and only if it is not a turning point.

Definition 4. $(\overset{\star}{x}, \overset{\star}{\lambda})$ is a simple bifurcation point if

$$B^2 - 4AC > 0 \quad (1.6)$$

where $A = \frac{1}{2} \psi(F''_{\lambda\lambda}(x^*, \lambda^*)) + 2F''_{\lambda x}(x^*, \lambda^*)\omega + F''_{xx}(x^*, \lambda^*)\omega\omega$

$$B = \psi(F''_{x\lambda}(x^*, \lambda^*)\phi + F''_{xx}(x^*, \lambda^*)\phi\omega)$$

and ω is a solution of

$$F'_x(x^*, \lambda^*)\omega = -F'_{\lambda}(x^*, \lambda^*) \quad (1.7)$$

For simple bifurcation points an extended system was defined by G. Moore [6]. This system is also regular. However, some of the extended systems arising in bifurcation theory do not have an isolated solution. An example of this is the direct method used by R. Seydel [10] for the determination of bifurcation points.

$$\begin{aligned} F(x, \lambda) &= 0 \\ (1.8) \quad F'_x(x, \lambda)\phi &= 0 \\ k(\phi) - 1 &= 0 \end{aligned}$$

where k is a (not necessarily linear) function, from X into IR , with the property that

$$k'_x(\phi)\phi \neq 0 \quad (1.9)$$

It was proved by A. Spence and G. Moore [7] that system (1.8) is regular if it is used for a simple turning point. It can also be proved that it is singular for all other types of bifurcation.

In [3] it is proved that system (1.8) has a double solution $(\overset{\star}{x}, \overset{\star}{\lambda}, \phi)$ if

(x^*, λ^*) is a simple bifurcation point and if $C \neq 0$. The definition of a double solution will be given in section 2.

If the Banach space X is not finite dimensional, it will be necessary to use a discretization method for the approximate solution of $H(y) = 0$. This means we will replace this equation by a family of equations

$$H_h(y_h) = 0 \quad (1.10)$$

If $H_h(z) - H(z)$ has an asymptotic expansion in even powers of h , it is sometimes possible to improve the accuracy of the results by applying an extrapolation scheme to the results y_h for several values of h .

H. Stetter proved that the main condition for the existence of an asymptotic expansion, in even powers of h , for the global discretization error $y_h - y$, is that the solution y is isolated. In this article we will prove a generalization of that result.

2. The existence of asymptotic expansions

Let us consider an operator equation

$$F(y) = 0 \quad (2.1)$$

where F is a non-linear operator from a Banach space X into a Banach space Y . We define now a family of non-linear operators F_h and we replace the original problem by :

$$F_h(y_h) = 0 \quad (2.2)$$

where

$$F_h(z) = F(z) + \sum_{v=1}^n h^{2v} f_v(z) + O(h^{2n+1}); \quad z \in X \quad (2.3)$$

It was shown by H. Stetter [11] in 1965 (under more general conditions) that, if y is an isolated solution of (2.1), the approximate solution y_h has an asymptotic expansion in even powers of h . More precisely :

$$y_h = y + \sum_{v=1}^n h^{2v} y_v + O(h^{2n+1}) \quad (2.4)$$

We will give a generalization of that result. From now on we suppose that the Fréchet derivative $F'(y)$ is singular and that its nullspace is one-dimensional. We also assume that $F'(y)$ has a closed range

$$N(F^*(y)) = \text{span}(\phi); \quad \phi \in X; \quad \|\phi\| = 1 \quad (2.5)$$

$$N(F'^*(y)) = \text{span}(\psi); \quad \psi \in Y^*; \quad \|\psi\| = 1 \quad (2.6)$$

where Y^* denotes the space of all linear mappings from Y into IR . The solution of (2.1) is called "double" [5] if and only if

$$a_2 = \psi F''(y)\phi\phi \neq 0 \quad (2.7)$$

We are almost ready to state the main theorem.

The only problem is that an approximate equation for an equation with a double solution may have two solutions or no solutions at all! [5]

This implies that we will have to impose a condition on the kind of discretization used in the approximation in order to have real roots.

3. Proof of the main theorem

THEOREM 1.

If the non-linear operator F described in section 2 satisfies conditions (2.5), (2.6) and (2.7) and if the discretization F_h satisfies for $k \leq 2n+1$ and for all $z, z_1, \dots, z_k \in X$:

$$F_h^{(k)}(z)z_1 \dots z_k = F^{(k)}(z)z_1 \dots z_k + \sum_{v=1}^n h^{2v} f_{kv}(z)z_1 \dots z_k + O(h^{2n+1}) \quad (3.1)$$

where $f_{kv}(z)$ is a multilinear operator mapping X^k into Y and finally if

$$\psi f_{01}(y)/\psi F''(y)\phi\phi < 0 \quad (3.2)$$

then there exist y_1, \dots, y_{2n} such that (for h small enough)

$$F_h(y + \sum_{v=1}^{2n} h^v y_v) = O(h^{2n+1}) \quad (3.3)$$

$$\text{proof : Let } S = \sum_{v=1}^{2n} h^v y_v \text{ and } \tilde{y} = y + S \quad (3.4)$$

We will first construct equations for y_v ($v = 1, \dots, 2n$), afterwards we will show that those equations have (two) solutions which satisfy (3.3).

We expand $F_h(y+s)$ in a Taylor series around y :

$$F_h(y+s) = \sum_{k=0}^{2n} \frac{1}{k!} F_h^{(k)}(y)(s)^{(k)} + O(h^{2n+1}) \quad (3.5)$$

where $(T)^{(k)}$ denotes the k -tuple (T, \dots, T) . Using expansion (3.1) we find :

$$F_h(y+s) = \sum_{k=0}^{2n} \frac{1}{k!} [F^{(k)}(y)(s)^{(k)} + \sum_{v=1}^n h^{2v} f_{kv}(y)(s)^{(k)}] + O(h^{2n+1}) \quad (3.6)$$

If we use expression (3.4) for s and if we demand that $F_h(y+s) = O(h^{2n+1})$, (this means that the coefficient of h^s in (3.6) must be zero for $s \leq 2n$) then we find $2n$ equations. For $1 \leq s \leq 2n$:

$$\sum_{0 \leq k \leq 2n} \frac{1}{k!} F^{(k)}(y) y_{i_1} \dots y_{i_k} + \sum_{0 \leq k \leq 2n} \frac{1}{k!} f_{kv}(y) y_{i_1} \dots y_{i_k} = 0 \\ i_1 + \dots + i_k = s \quad i_1 + \dots + i_k + 2v = s \quad (3.7)$$

For $s = 1$ (3.7) reduces to $F'(y)y_1 = 0$ which has infinitely many solutions :

$$y_1 = \alpha_1 \phi; \alpha_1 \in \text{IR}.$$

For $s = 2$ we find, combining (3.7) and the expression for y_1 :

$$F'(y)y_2 = -\frac{1}{2} F''(y)\phi\phi \alpha_1^2 - f_{01}(y) \quad (3.8)$$

This equation is solvable if the right hand member belongs to Range ($F'(y)$).

Since $u \in \text{Range}(F'(y))$ if and only if $\psi u = 0$ we find :

$$\alpha_1 = \pm \left[\frac{-2\psi f_{01}(y)}{\psi F''(y)\phi\phi} \right]^{1/2} \quad (3.9)$$

The quantity under the square root is positive because of condition (3.2).

Let $y_{2,\min}$ be a solution of (3.8) the most general solution has the form :

$$y_2 = y_{2,\min} + \alpha_2 \phi; \alpha_2 \in \text{IR} \quad (3.10)$$

We now proceed by induction : suppose we found y_v for $v = 1, \dots, s-1$,

$y_v = y_{v,\min} + \alpha_v \phi$, where $\alpha_1, \dots, \alpha_{s-2}$ are known real numbers and α_{s-1} is still arbitrary. For $s \geq 3$ equation (3.7) can be transformed in an equation that can be solved for y_s .

$$F'(y)y_s = -F''(y)y_1 y_{s-1} - R_s \quad (3.11)$$

where R_s is independent of α_{s-1} and y_s .

Using the expressions for y_1 and y_{s-1} we find that equation (3.11) is solvable if and only if

$$\alpha_{s-1} = \frac{-\psi R_s - \alpha_1 \psi F''(y) \phi y_{s-1, \min}}{\alpha_1 \psi F''(y) \phi \phi} \quad (3.12)$$

$$\text{and again : } y_s = y_{s, \min} + \alpha_s \phi ; \alpha_s \in \text{IR} \quad (3.13)$$

Continuing this way we find expressions for y_1, \dots, y_{2n-1} and $y_{2n, \min}$. The constant α_{2n} remains arbitrary.

We find that $F_h(\tilde{y}) = O(h^{2n+1})$ which proves the theorem.

Theorem 1 proves that \tilde{y} is almost a solution. However, we need something more. We would like \tilde{y} to be as close as possible to y_h . Therefore we need some stability properties for the operators $\{F_h'(\tilde{y}) | h \leq h_0\}$.

Under certain conditions, it can be proved that $F_h'(\tilde{y})$ has an inverse, which is bounded. (e.g. Collective compactness of the family $\{F_h'(\tilde{y})\}$) [1].

The proof is very technical and is omitted here.

THEOREM 2.

Suppose there exist positive real numbers K, C and δ such that :

$$\|F_h(\tilde{y})\| \leq K h^{2n+1} \quad (3.14)$$

$$\|F_h'^{-1}(\tilde{y})\| \leq C h^{-1} \quad (3.15)$$

$$\text{and } \sup_{\|x-\tilde{y}\| \leq \delta} \|F_h^{(i)}(x)\| \leq M_i < \infty \quad \text{for } i = 1, 2 \quad (3.16)$$

Then there exists a positive real number R such that G_h is a contraction on B_R , where :

$$G_h(x) = x - F_h'^{-1}(x) F_h(x) \quad (3.17)$$

$$B_R = \{x \in X : \|x - \tilde{y}\| \leq Rh^2\} \quad (3.18)$$

Proof : First we will show that $\|G'_h\|$ is bounded by a constant smaller than 1.

Afterwards we will prove that $G_h(x) \in B_R$ if $x \in B_R$.

$$G'_h(x) = -F_h'^{-1}(x)F_h''(x)F_h'^{-1}(x)F_h(x) \quad (3.19)$$

For all $R > 0$ it is true that :

$$\|F_h(x)\| \leq K h^{2n+1} + M_1 R h^2 \leq 2M_1 R h^2 \quad (3.20)$$

The last inequality is valid if h is small enough.

Because of the continuity of $F_h'^{-1}(x)$ near \tilde{y} there exist a positive real number R_0 such that if $R \leq R_0$ and if $\|x - \tilde{y}\| \leq R h^2$:

$$\|F_h'^{-1}(x)\| \leq 2C h^{-1} \quad (3.21)$$

If $R \leq \min(R_0, (16 C^2 M_1 M_2)^{-1})$ it follows that

$$\|G'_h(x)\| \leq \|F_h'^{-1}(x)\|^2 \|F_h''(x)\| \|F_h(x)\| \leq \frac{1}{2} \quad (3.22)$$

This proves the first part of the theorem. Secondly we prove that

$G_h(x) \in B_R$ if $x \in B_R$. Using the triangle inequality we find

$$\|G_h(x) - \tilde{y}\| \leq \|G_h(x) - G_h(\tilde{y})\| + \|G_h(\tilde{y}) - \tilde{y}\| \leq \frac{1}{2} \|x - \tilde{y}\| + \|G_h(\tilde{y}) - \tilde{y}\|$$

Taking (3.14) and (3.15) into account it follows that $G_h(x) \in B_R$ if h is small enough.

Theorem 2 proves that G_h has a unique fixed point in B_R or equivalently : there is a unique solution, $y_h \in B_R$, of the approximate equation (2.2).

For this y_h (there is another one outside B_R !) we will show that

$\|y_h - \tilde{y}\| = O(h^{2n})$. Therefore we expand $F_h(y_h)$ in a Taylor series around \tilde{y} .

$$0 = F_h(y_h) = F_h(\tilde{y}) + F_h'(\tilde{y})(y_h - \tilde{y}) + \int_0^1 t F_h''(\tilde{y} + t(y_h - \tilde{y})) dt (y_h - \tilde{y}) \quad (2)$$

From condition (3.15) of theorem 2 it follows that

$$\|F_h'(\tilde{y})(y_h - \tilde{y})\| \geq \frac{h}{C} \|y_h - \tilde{y}\| \quad (3.24)$$

Substitution of (3.24) into (3.23) and taking the uniform bound for $\|F_h''\|$ into account :

$$\frac{1}{2} M_2 \|y_h - \tilde{y}\|^2 - \frac{h}{C} \|y_h - \tilde{y}\| + K h^{2n+1} \geq 0 \quad (3.25)$$

This inequality has two solutions for $\|y_h - \tilde{y}\|$, the first one is a lower bound for $\|y_h - \tilde{y}\|$ which cannot be satisfied by the solution y_h belonging to B_R . So $\|y_h - \tilde{y}\|$ must satisfy the other inequality :

$$\|y_h - \tilde{y}\| \leq \frac{h - [h^2 - 2 M_2 K C^2 h^{2n+1}]^{1/2}}{C M_2} \quad (3.26)$$

A simple calculation shows that (for h small enough) (3.26) implies :

$$\|y_h - \tilde{y}\| \leq 2 K C h^{2n} \quad (3.27)$$

Theorems 1 and 2 guarantee that the approximate solution can be written in the form :

$$y_h = y + \sum_{v=1}^{2n-1} h^v y_v + O(h^{2n}) \quad (3.28)$$

This is very interesting since it provides the possibility of using an extrapolation scheme on the approximate results y_h . Numerical examples will be given in section 4.

4. Numerical examples

The first example is a two point boundary value problem. [6]

$$\begin{aligned} \frac{d^2}{ds^2} x(s) - p(\lambda) \frac{d^2}{ds^2} X(s) + \pi^2 \lambda f(x(s) - p(\lambda)X(s)) &= 0 \\ x(0) = x(1) = 0 \end{aligned} \quad (4.1)$$

where $p(\lambda) = \lambda^4 e^{-\lambda/2}$; $f(z) = z^2 + z$ and $X(s) = s(1-s)e^s$.

It can be shown that $\lambda^* = 1$ is a simple bifurcation point, the corresponding solution equals $x^*(s) = s(1-s)e^{s-1/2}$.

If the extended system (1.8) is used and if we use the simplest finite difference method :

$$\frac{d^2}{ds^2} x(s) \approx \frac{x(s+h) - 2x(s) + x(s-h)}{h^2} \quad (4.2)$$

and if we replace the second boundary condition by

$$y_h(1) = 0.8 h^2 \quad (4.3)$$

then there are two real solutions (x_h, λ_h, ϕ_h) .

In table 1 we find the result of an application of a Richardson extrapolation scheme on the approximate values λ_h .

TABLE 1

$h = 1/15 : 0.9214$	
	1.001783
$h = 1/23 : 0.9414$	1.0002237
	1.000892
$h = 1/35 : 0.9670$	1.0000641
	1.000423
$h = 1/53 : 0.9783$	1.0000184
	1.000195
$h = 1/80 : 0.9857$	

The successive values of h were found by multiplying the number of nodes n each time by $3/2$. ($1/h = n-1$).

The second example is an integral equation

$$f(x) - \int_0^1 \frac{(f(y)-1)(f(y)-\mu) + x}{1+xy} dy = 1 - \ln(1+x) \quad (4.4)$$

The extended system was solved by the trapezoidal rule. There are two approximate solutions.

If we use the midpoint rule then the approximate equation has no real solutions. In table 2 we find the approximate values for μ_h . The underlined figures are exact.

TABLE 2 (μ_h)

$h = 1/2$: -	0.031	- 0.21938
$h = 1/4$: -	0.325	- 0.2024770
			- 0.20670
$h = 1/8$: -	0.165	- 0.2030173
			- 0.20393
$h = 1/16$: -	0.184	- 0.2031047
			- 0.20331
$h = 1/32$: -	0.194	

Both examples show that very accurate results can be obtained. The discretized system $H_h(x_h, \lambda_h, \phi_h) = 0$ was solved by Newton's method, starting from a solution $(x_{\bar{h}}, \lambda_{\bar{h}}, \phi_{\bar{h}})$ of the same problem for a larger value of h . The regularity of $H_h(x_h, \lambda_h, \phi_h)$ ensures quadratic convergence.

Acknowledgement

The author would like to thank Prof. Dr. R. Piessens for the reading of the manuscript.

References

1. K. Atkinson, A survey of numerical methods for the solution of Fredholm integral equations of the second kind. SIAM, Philadelphia, (1976).
2. R. Caluwaerts, An asymptotic expansion for the discretization error of a geometrically isolated solution of a non-linear equation. In preparation.
3. R. Caluwaerts, The use of a singular extended system for the numerical computation of bifurcation points. In preparation.
4. R. Caluwaerts, A direct method for the determination of non-simple turning points. In preparation.
5. H.B. Keller, Geometrically isolated non-isolated solutions and their approximation. SIAM J. Numer. Anal. v. 18, 822-838 (1982).

6. G. Moore, The numerical treatment of non-trivial bifurcation points. *Numer. Funct. Anal. and Optimiz.*, v. 2, 441-472 (1980).
7. G. Moore and A. Spence, The calculation of turning points of non-linear equations. *SIAM J. Numer. Anal.*, v. 17, 567-576, (1980).
8. D. Roose and R. Caluwaerts, Direct methods for the computation of non-simple turning points corresponding to a cusp. These proceedings.
9. D. Roose and R. Piessens, Numerical computation of nonsimple turning points and cusps. Report TW 60, Dept. of Computer Science, Universiteit Leuven (1983) (Submitted to *Numer. Math.*).
10. R. Seydel, Numerical computation of branch points in ordinary differential equations, *Numer. Math.*, v. 32, 51-68, (1979).
11. H. Stetter, Asymptotic expansions for the error of discretization algorithms for non-linear functional equations. *Numer. Math.*, v. 7, 18-31, (1965).

R. Caluwaerts
Catholic University of Leuven
Dept. of Computer Science
Celestijnlaan 200 A
B-3030 Heverlee (Belgium)

Techniques for Large Sparse Systems Arising from Continuation Methods

Tony F. Chan¹

Computer Science Dept., Yale University, New Haven, CT06520, USA.

Abstract: We survey numerical techniques for solving the nonlinear and linear systems arising from applying continuation methods to tracing solution manifolds of parameterized nonlinear systems of the form $G(u, \lambda) = 0$. We concentrate on large and sparse problems, e.g. discretizations of partial differential equations, for which this part of the computation dominates the overall cost. The basic issue is a tradeoff of the exploitation of the sparsity structure of the Jacobian G_u and the numerical treatment of its singularity. Among the techniques to be discussed are: Newton and quasi-Newton methods, low rank correction methods, implicit deflation techniques, Krylov subspace iterative methods and multi-grid methods.

1 Introduction

In this paper, we are concerned with the numerical solution of parameterized nonlinear systems of the form

$$G(u, \lambda) = 0, \quad (1)$$

where $u \in R^n$, $\lambda \in R^m$ and $G: R^n \times R^m \rightarrow R^n$. Such systems arise in many problems in scientific computing. In the modelling of nonlinear physical phenomena, u may correspond to a field variable and λ to a set of physical parameters. Another source of such parameterized systems is the class of homotopy continuation methods [32] for improving the global convergence of locally convergent methods (e.g. Newton's method) for solving nonlinear systems and fixed point problems. In these homotopy techniques, one transforms a nonlinear system $F(u) = 0$ by a homotopy, e.g. $G(u, \lambda) = (1-\lambda)(u-u_0) + \lambda F(u) = 0$, so that one starts from the known solution u_0 at $\lambda = 0$ and trace the solution curve of $G(u, \lambda)$ until $\lambda = 1$ to obtain the solution of $F(u) = 0$.

In general, the equation $G(u, \lambda) = 0$ defines a m -dimensional manifold in R^{n+m} . Very often, in addition to obtaining the solution u at a few selected parameter values, more physical insight can often be gained by knowing some general features of the solution manifold as a result of varying the parameters. A continuation procedure can generally be defined as a method for tracing parts of the solution manifold [2, 46, 63, 73]. The design of such a procedure would be straightforward if the solution manifold can be parameterized by the naturally occurring parameters λ . However, this cannot always be done because the solution manifold may contain singular points where the Jacobian G_u is singular and where this parameterization breaks down. Most continuation methods overcome this problem by using a different parameterization of the solution manifold implicitly defined by an additional set of equations

$$N(u, \lambda) = 0, \quad (2)$$

¹This work was supported in part by the Department of Energy under contract DE-AC02-81ER10996 and by the Army Research Office under grant DAAG-83-0177.

where $N: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^m$, so that a solution (u, λ) of (1) is always a *regular* solution of the coupled system (1) and (2), even at points where G_u is singular. Conventional methods can then be used to solve this coupled system, e.g. Newton's method, in a predictor-corrector fashion in which a nearby computed solution (u_0, λ_0) is used to generate an initial guess.

In this paper, we shall concentrate on large and sparse problems, for example, where $G(u, \lambda)$ may represent the discretization of nonlinear partial differential equations. For these problems, the solution of the coupled nonlinear system (1) and (2) constitutes the major computational cost of the continuation procedure. It is therefore important to exploit sparsity and structures in G (or G_u) so as to increase the computational efficiency. This can be achieved by any algorithm for solving (1) and (2) each step of which involves solving a subproblem involving G (or G_u) with λ fixed. However, this approach conflicts with the desire to avoid dealing with the possible singularity of G_u which was the reason for introducing the new parameterization in the first place. Therein lies the basic issue: how does one find a way to exploit the structures in G without running into numerical problems with the singularity of G_u ?

This basic conflict not only occurs in the basic continuation procedure, but also in many related algorithms. For example, many types of singular points of the solution manifold, such as turning points, bifurcation points and cusp points, can be characterized as regular solutions to coupled systems of the form of (1) and (2) [1, 8, 44, 51, 53, 55, 57, 60, 62, 67, 68] and consequently computational algorithms derived from this approach must deal with the same conflict. Many techniques to be discussed here have straightforward applications to these problems as well.

We note that it is possible to avoid dealing with such singularities if one stays away from singular points of the solution manifold. It then becomes possible to use G_u^{-1} explicitly in a computational algorithm and exploit the structures in G_u . Many large problems using continuation methods have been solved using this approach [19, 48, 50, 71] and we shall not elaborate on them in this paper. We feel that the regularization of the problem by introducing the N-equation presupposes the necessity of dealing with *possible* singularities of G_u and thus it is desirable to have computational procedures that automatically handle such singularities. Moreover, such a capability becomes indispensable if one is interested in computing the singular points themselves, such as locating turning points and bifurcation points and following folds in the solution manifold [64]. It is one of the themes in this paper to show that the extra cost in doing so is not too high for many numerical techniques suitable for solving large problems.

For ease of presentation, we shall restrict our discussions to the special case of $m = 1$ (i.e. one parameter). This case is also the most common because in practice one often alternately freezes all parameter values except one in tracing the solution manifold. We shall also only treat the case where the dimension of the null space of G_u is at most one. All of the techniques to be presented can be generalized to the higher dimensional cases in a straightforward manner.

2 Nonlinear Techniques

The coupled nonlinear system (1) and (2) can be considered as a single nonlinear system in the variable $z = (u, \lambda)$, namely:

$$F(z) = \begin{bmatrix} G(u, \lambda) \\ N(u, \lambda) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad (3)$$

Since we are seeking a regular solution of (3), a regular nonlinear iterative methods can be applied. Most of the methods to be presented here are of this nature.

An obvious approach is to use Newton's method or one of its many variants (chord, damped, discrete, truncated ...) [59], applied directly to the coupled system (3). At each iteration, a linear system involving the Jacobian:

$$M = \begin{bmatrix} G_u & G_\lambda \\ N_u & N_\lambda \end{bmatrix}$$

must be solved. If the parameterization N is chosen appropriately, the Jacobian M is nonsingular [46] and thus Newton's method has local quadratic convergence, *even when G_u is singular*. The usual drawback of lack of global convergence for Newton's method is not severe in continuation methods because the continuation step size can be controlled to insure local convergence. Newton's method, however, does require evaluation and storage of the Jacobians of G and N . For sparse problems, sparse estimation techniques can be used [20, 22, 61].

Georg [35] and Kearfott [45] have considered dense quasi-Newton methods [24] for solving (3). No Jacobian is needed and only evaluations of G and N are required. Superlinear rate of convergence is usually achievable. For problems where the Jacobians are not available or costly to evaluate, this represents an advantage. However, since the Jacobian plays a central role in bifurcation problems, it may be needed for other purposes anyway, for example, for branch switching [46]. For large and sparse problems, sparse update of the approximate Jacobian is needed and experience has shown that these do not perform as efficiently as Newton-like methods [38]. The successful application of these methods to large continuation methods remains to be proven.

A very interesting idea has been proposed in [37] and later used in [11]. The nonlinear system (3) is transformed into a least squares minimization problem for the functional $G^2 + N^2$, and a preconditioned nonlinear conjugate gradient method [31, 36] is used for finding the minimum. This technique is especially convenient in situations where a least squares method is already used in a finite element variational setting for solving the system $G = 0$ with fixed λ . However, since the use of least squares approach squares the condition number of G_u , it is extremely important for efficiency reasons to use a good preconditioning.

Lastly, nonlinear relaxation techniques can be attempted. For example, point nonlinear SOR methods [59] can be applied. Another possibility is to use block nonlinear SOR by relaxing u with the G equation and λ with the N equation alternately. By design, these methods exploit sparseness in G . However, straightforward application usually encounters convergence difficulty because the Jacobian M is often not positive definite or diagonally dominant. The point nonlinear SOR method, however, can be used as a smoother in a nonlinear multi-grid method (see Section 4).

3 Linear Techniques

Among the nonlinear methods discussed in the last section, the class of Newton-like methods is by far the most commonly used. It is the most general method and has fast local convergence. The need to evaluate Jacobians is compensated by the central role the Jacobians play in bifurcation problems. In the rest of the paper, we shall only deal with this class of methods.

In each iteration of a Newton-like method, a linear system of the form

$$M \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} A & b \\ c^T & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix}, \quad (4)$$

needs to be solved, where the n by n matrix A ($\equiv G_u$) is bordered by the vectors b and c to form a larger system of dimension $(n+1)$ by $(n+1)$. When A is large and sparse, one would like to exploit the structures in A when solving this system. However, M does not necessarily inherit desirable structures of A , such as bandedness, symmetry, positive definiteness and separability (for fast direct solvers). It is thus natural to consider algorithms for solving (4) that do exploit these structures in A . On the other hand, dealing with A directly necessarily leads to numerical problems with its possible singularity. These two competing goals constitute the fundamental issue that must be resolved by any practical algorithm. In this section, we shall discuss how some commonly used linear algorithms for large problems can be modified to handle this problem.

3.1 The Deflated Block-Elimination Algorithm

An algorithm that fully exploits structures in A is the following block-elimination algorithm (corresponding to block Gaussian Elimination on M):

Algorithm BE [46]

$$(1) \text{ Solve } A v = b, \quad (5)$$

$$A w = f. \quad (6)$$

$$(2) \text{ Compute } y = (g - c^T w) / (d - c^T v).$$

$$(3) \text{ Compute } x = w - y v.$$

Note that only a solver for A is needed. This solver could use any method that is appropriate for the particular problem, for example, sparse Gaussian Elimination, a fast direct solver, an iterative method or a multi-grid method. The last two cases will be discussed in more detail in later sections. In this section, we shall only consider methods based on Gaussian Elimination.

In general, the work consists mainly of one factorization of A and two backsolves with the LU factors of A . Moreover, for problems with many right hand sides (e.g. in chord-Newton methods), the factorization needs to be computed only once. However, since we use A^{-1} explicitly, we can expect problems when A is nearly singular [12]. Consider the following simple example ($n = 2$):

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & \epsilon & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 1+\epsilon \\ 1 \end{bmatrix}.$$

where $|\epsilon|$ is smaller than the machine epsilon of the computer, i.e. $1 + \epsilon = 1$ in floating point arithmetic. In exact arithmetic, $v = (-1/\epsilon, 1/\epsilon)^T$ and $w = (1 - 1/\epsilon, 1 + 1/\epsilon)^T$. In floating point arithmetic, $w = (-1/\epsilon, 1/\epsilon)^T$. Assuming that this is the only round-off error committed, Algorithm BE would give $x = (0, 0)^T$ which obviously has a large relative error.

In [12], deflation techniques are proposed for stabilizing Algorithm BE. Instead of

computing v and w directly from (5) and (6), numerically stable representations for them are computed by working in subspaces orthogonal to approximate null vectors ϕ and ψ of A . Based on these deflated decompositions of v and w , stable variants of Algorithm BE are derived, one version of which is:

Algorithm DBE: Deflated Block-Elimination Algorithm [12].

1. Compute an approximate normalized left singular vector ψ of A .
2. Compute $\phi = \delta A^{-1}\psi$, where $\delta = 1 / \|A^{-1}\psi\|$.
3. Compute $c_b = (\psi^T b)$ and $c_f = (\psi^T f)$.
4. Solve $Av_d = b - c_b \psi$ for v_d . (v is represented as: $v = v_d + (c_b/\delta)\phi$)
5. Solve $Aw_d = f - c_f \psi$ for w_d . (w is represented as: $w = w_d + (c_f/\delta)\phi$)
6. Compute $h_1 = g - c^T w_d$, $h_2 = d - c^T v_d$, $h_3 = h_1 c_b - h_2 c_f$, $h_4 = (c^T \phi) c_f - \delta h_1$, $D = (c^T \phi) c_b - \delta h_2$.
7. Compute $y = h_4 / D$ and $x = w_d + (h_3 \phi - h_4 v_d) / D$.

Note that only two solves with A is needed, exactly the same as in Algorithm BE. The major overhead for performing the deflation is the computation of ψ and one backsolve for ϕ . The vector ψ can be computed with only one or two backsolves. One possibility is to use a few steps of an inverse iteration [12, 13, 69]. Another algorithm that has been used in the literature is based on computing a LU-factorization of A with a small n -th pivot [16, 47, 49]. Although the usual pivoting strategies (partial and complete pivoting) [25] will exhibit such a LU-factorization for most nearly singular matrices, it is well known that there are counter-examples to this commonly assumed fallacy [16, 38, 74]. An algorithm that is guaranteed to produce such a factorization is given in [16]. This extra work in computing ϕ and ψ is compensated by the fact that they can be reused for several continuation steps and are also useful for switching branch at bifurcation points [46].

In [47, 49], a similar algorithm is independently proposed, but one that works only for the case where A is *exactly* singular. Errors occur if A is nearly but not exactly singular. On the other hand, Algorithm DBE can be proven to be numerically stable [12] *independent of the singularity of A*. Because of its robustness and low overhead, Algorithm DBE can be used at all continuation steps without necessitating a check on the singularity of A .

3.2 Sparse Matrix Methods

If A is sparse, then so is M . Therefore a sparse matrix solver [26, 28, 29, 34] can be used directly on M . However, even if A has a sparse LU factorization, M does not necessarily have a factorization that is just as sparse. This is because if A is nearly singular then some pivoting with the last row or column of M is needed for numerical stability when factoring M which may adversely affect the fill-ins. This is especially severe if a pivot with the last column or row occurs early in the elimination process, as the following simple example shows:

$$(a) \begin{bmatrix} x & & x \\ & x & 0 & x \\ & . & . & . \\ 0 & & . & . \\ x & x & \dots & x \end{bmatrix} \quad (b) \begin{bmatrix} x & x & \dots & x \\ x & x & & \\ . & . & . & 0 \\ . & 0 & . & \\ x & & & x \end{bmatrix}$$

Without pivoting, matrix (a) produces no fill-in, whereas pivoting with the $(n+1, n+1)$ th element gives matrix (b) which produces a complete fill-in.

Fortunately, there are situations where it can be shown that the last row or column of M does not have to be pivoted until towards the end of the elimination process. This is true, for example, if a pivoting strategy can be found to produce a *sparse* LU factorization of A with a small n -th pivot. Using the same pivoting sequence for factoring M , possibly with the last row and column of M appropriately scaled, we obtain at the last stage of the elimination process, a coefficient matrix of the form:

$$\begin{bmatrix} U & u & t \\ 0 & \epsilon & p \\ 0 & q & s \end{bmatrix}$$

where U is sparse and ϵ is the small pivot. Using complete or partial pivoting for the lower right hand 2 by 2 submatrix now will handle the singularity.

3.3 Banded Matrices

Discretizations of differential equations often give rise to banded rather than generally sparse matrices. If a banded LU factorization with a small n -th pivot can be found, then the method outlined in Section 3.2 can be used. This is possible for some two point boundary value problems where the parameter λ occurs in the boundary conditions [49].

For a general banded matrix, Rheinboldt [65] proposed the following method. The matrix M is splitted according to $M = S + R$, where S has the same form as M except the vectors b and c are both replaced by the k -th unit vector and R is a rank 2 matrix. The index k is chosen so that S is as well-conditioned as possible. Using the Sherman-Morrison-Woodbury formula [42], for every system in M , one can equivalently solve three systems in S . By taking advantage of the special form of S , it can be shown that a system in S can be reduced to one for A with its k -th row and column deleted, which preserves bandedness. Since it requires working with the (possibly complicated) storage structures of A , this algorithm is not as modular as Algorithm DBE. Moreover, generally one more backsolve is required.

3.4 Iterative methods

For many large and sparse problems, e.g. multi-dimensional PDEs, iterative methods may become competitive with direct methods, both in terms of storage and computational time. One of the most successful iterative methods is the class of Krylov subspace iterative methods [3, 18, 21, 27, 30, 40, 41, 43, 52, 68, 72, 75]. In addition to sparseness, the symmetry of the coefficient matrix often plays a critical role in both the efficiency and the convergence of these iterative methods. In general, efficient methods and rather complete theories exist for symmetric and positive definite problems, whereas the situation for indefinite and nonsymmetric problems are not as well-understood. We shall assume in this section only that A is symmetric.

Although M inherits the sparseness properties of A , M may be nonsymmetric while A is symmetric. Therefore the obvious approach of applying a nonsymmetric iterative method directly to (4) may fail to exploit the symmetry of A . In [17], some alternative algorithms are proposed. One approach that does exploit the symmetry in A is to use Algorithm BE. However, two linear systems of dimension n have to be solved for each system involving M . Moreover, deflation techniques may have to be used to handle the singularity of A . In principle, deflation techniques for conjugate gradient type methods can be obtained by applying the techniques developed in [13] to the tridiagonal factor produced by the underlying Lanczos process. This is currently under development. Another method that exploits symmetry of A is a low rank correction method. For example, if we split M as $M = S + uv^T$, where S has the same form as M except that the vector b is set to be equal to c , then the solution of (4) can easily be obtained via the Sherman-Morrison formula [42] by solving two systems with the symmetric and nonsingular matrix S . Mittelman [54] has even considered choosing the parametrization N in the continuation method so that $N_u = G_\lambda$ to produce a symmetric M . Finally, one can apply a symmetric positive definite method to the normal equations derived from the M -system. However, it is well-known that the convergence rate will suffer. In short, the alternatives are solving one nonsymmetric system or two symmetric systems or one symmetric positive definite ill-conditioned system.

Another issue is the choice of a good preconditioning, which is often essential for the successful application of Krylov subspace based iterative methods. Assume that a good preconditioning is available for the matrix A in the form of a symmetric matrix B such that $B^{-1} \approx A^{-1}$ and such that the matrix-vector product $B^{-1}x$ is easy to compute. The use of preconditioning in Algorithm BE is straightforward, because the preconditioning B^{-1} can be applied directly to the systems with A as coefficient matrix. Next, consider the matrices M and S . One way to obtain a preconditioning is to first express the exact inverse in terms of A^{-1} and then replace A^{-1} by B^{-1} . Thus, for example, we have

$$M^{-1} = \begin{bmatrix} A^{-1}(I - bu^T) & v \\ u^T & -y^{-1} \end{bmatrix} \quad (7)$$

where

$$y = c^T A^{-1} b - d, \quad u = A^{-1} c / y, \quad v = A^{-1} b / y. \quad (8)$$

Replacing A^{-1} by B^{-1} in (7) and (8), one obtains the following preconditioner for M :

$$P_1 = \begin{bmatrix} B^{-1}(I - b\tilde{u}^T) & \tilde{v} \\ \tilde{u}^T & -\tilde{y}^{-1} \end{bmatrix} \quad (9)$$

where the "hatted" quantities are defined by analogy to (8), but with A^{-1} replaced by B^{-1} . In addition to P_1 , we can use the following simpler preconditioning:

$$P_2 = \begin{bmatrix} B^{-1} & 0 \\ 0 & 1 \end{bmatrix} \quad (10)$$

In [17], numerical experiments were carried out to compare some of the above techniques. We applied them to the model nonlinear elliptic problem $G(u,\lambda) = \Delta u + \lambda e^u = 0$ with zero Dirichlet boundary condition on a unit square. This problem has a simple turning point. For the preconditioning, we use $B = \Delta$. We briefly summarize the results here. It was found that the use of a good preconditioner is extremely important. In particular the methods that do not use preconditioning are slow and sensitive to nonsymmetry near the turning point whereas symmetry is not as important for preconditioned systems. If a good preconditioning is available, it seems best to work directly with the nonsymmetric M than with the symmetric systems. In fact, the method $P_2 M$ gives the best results in execution time. As expected, the normal equations approach is not competitive.

Lastly, we point out that a Newton-Krylov subspace method can be implemented by directional differencing techniques without computing or storing the Jacobian matrix [14, 33, 58] and can be used in conjunction with inexact Newton algorithms [23].

4 Multi-Grid Methods

If G is a discretization of a differential or integral operator, then one may consider using a multi-grid (MG) method [10] for solving (3). For this, we need a hierarchy of nested grids on which the discretizations of the operators G , N and their Jacobians are defined. In addition, we need a smoother on each grid, for example, a point relaxation method or a conjugate gradient method. For the MG method to work, the operators and smoothers on the grids must be appropriately chosen to work together in a concerted manner. Although this is rather standard procedure for a large class of differential systems, very little of the theory and literature on MG is on solving coupled systems. It must be noted that the operator N may have very different smoothing and approximation properties on the hierarchy of grids than G and thus it is not obvious that MG can be made to work as efficiently on the coupled system (linear or nonlinear) directly as on G itself.

There are at least two ways in which multi-grid methods can be applied: (1) solve the linear systems that arise in Newton's method, or (2) solve the coupled nonlinear system directly. Consider case (1) first. The most straightforward approach is to use MG as the black box solver for A in Algorithm BE. However, the singularity of A again causes problems. It was first reported in [15] that MG diverges when A is nearly singular. This divergence is not caused by round-off errors but by the corrections from a coarse grid on which A is nearly singular. As a result of this singularity, the magnitude of the component of the correction in the null vector ϕ direction could be completely wrong. This could happen even if A on the finest grid is reasonably nonsingular. Fortunately, deflation techniques together with Algorithm DBE can be used to overcome this problem [6, 7]. The basic idea is to compute the deflated decompositions of the vectors v and w in Algorithm DBE by MG methods. Similar in spirit to algorithms proposed in [15], approximate null vectors ψ and ϕ are computed on all grids and the iterates on all grids except the finest are purged of these components after smoothing but before a transfer to another grid. The ϕ component on the finest grid is accumulated by a projection process. Again the overhead is low and the algorithm can be used without a check on the singularity of A . Another natural idea for handling the singularity of A is to add a small diagonal shift to M to make A nonsingular [5]. However, besides losing quadratic convergence in the Newton process, the shift has to be chosen carefully and thus is not as easy to implement robustly as the deflation techniques.

In addition to being used on A in conjunction with Algorithm BE, with modifications

standard MG techniques can be applied to M directly. Such an idea was proposed by Mittelmann and Weber [56]. On all the grids except the coarsest, smoothing is done only to x using the $Ax + yb = f$ equation with a fixed value for y . On the coarsest grid, a direct solver is used to solve the M -system with pivoting and y is updated there.

Now consider case (2). It is known that a version of MG, called the Full Approximation Scheme [10] (FAS), can be applied directly to a nonlinear system without first applying a linearization. Hackbusch [39] proposed a technique similar to Mittelmann and Weber's [56] except that a FAS is used on $G(u, \lambda)$ to smooth u on all the grids except the coarsest and λ is updated only on the coarsest grid. Stuben and Trottenberg [70], based on an idea of Brandt, proposed a slightly different algorithm in which both u and λ are updated on all grids. After a FAS smoothing step on a particular grid, u is scaled such that the constraint equation N is satisfied with the current λ , after which λ is updated by "averaging" the G equations on that grid. Recently, Bolstad and Keller [9] have combined the FAS with the r -extrapolation technique [10] for solving similar problems.

The above techniques of applying MG methods to continuation algorithms compute the solution (u, λ) on the finest grid as a solution of the coupled nonlinear system (3). This presupposes that a fine grid solution is needed on all points on a solution branch. Very often it suffices to compute the *qualitative* behaviour of the solution manifold, perhaps on a coarser grid, and a fine grid solution at a few selected points. If the main features of the solution manifold can be captured by a coarse grid, then a direct method based on Gaussian Elimination can be used on it without incurring a large computational effort. At a point where high accuracy is desired, a MG algorithm can be used to refine the solution. This idea is implemented in the MG-Continuation program PLTMGC [6]. This package can handle a general class of self-adjoint mildly nonlinear elliptic problems with a parameter dependence on a general two dimensional domain, and can compute target values in λ and $\|u\|$ with an adaptive stepping algorithm, detect and locate simple turning points and bifurcation points and switch branch at simple bifurcation points. It is based on an earlier package PLTMG [4] and uses Rayleigh-Ritz Galerkin techniques on piecewise linear triangular elements with adaptive mesh refinements. For refining the coarse grid solution using MG, MG deflation techniques [7] and Algorithm DBE are applied to ensure numerical stability.

References

- [1] J.P. Abbott. An Efficient Algorithm for the Determination of Certain Bifurcation Points. *Journal of Computational and Applied Mathematics* 4 :19 - 27, 1978.
- [2] E. Allgower and K. Georg. Simplicial and Continuation Methods for Approximating Fixed Points and Solutions to Systems of Equations. *SIAM Review* 22(1):28 - 85, 1980.
- [3] O. Axelsson. Conjugate gradient type methods for unsymmetric and inconsistent systems of linear equations. *Lin. Alg. Appl.* 29:1-16, 1980.
- [4] R.E. Bank. *PLTMG User's Guide, June, 1981 version.* Technical Report, Dept. of Mathematics, University of Calif. at San Diego, 1982.
- [5] Randolph E. Bank. Analysis of a Multilevel Inverse Iteration Procedure for Eigenvalue Problems. *SIAM J. Numer. Anal.* 19:886-898, 1982.
- [6] R.E. Bank and T.F. Chan. *PLTMGC: A Multigrid-Continuation Program for Parameterized Nonlinear Elliptic Systems.* Technical Report 261, Dept. of Computer Science, Yale University, 1983.
- [7] R.E. Bank and T. F. Chan. *Multi-Grid Deflation.* 1983. In preparation.
- [8] W. Beyn. Defining Equations for Singular Solutions and Numerical Applications. In T. Kupper, H. Mittelmann and H. Weber, Editors, *Numerical Methods for Bifurcation Problems*, Birkhauser Verlag, Basel, 1984 .
- [9] J. Bolstad and H.B. Keller. *A Multi-Grid Continuation Method for Elliptic Problems with Turning Points.* 1983. Paper presented at the Siam Fall Meeting, Norfolk, Virginia.
- [10] A. Brandt. Multi-level Adaptive Solution to Boundary Value Problems. *Math. Comp.* 31:333-390, 1977.
- [11] M.O. Bristeau, R. Glowinski, J. Periaux, G. Poirier. Non unique solutions of the transonic equation by arc length continuation techniques and finite element least squares methods. In *Proceedings of 5th international conference on finite elements and flow problems, Austin, Texas, , Jan. 23-26, 1984 .*
- [12] T.F. Chan. Deflation Techniques and Block-Elimination Algorithms for Solving Bordered Singular Systems. *Siam J. Sci. Stat. Comp.* 5 (1) March 1984.
- [13] T.F. Chan. *Deflated Decomposition of Solutions of Nearly Singular Systems.* Technical Report 225, Computer Science Department, Yale Univ., 1982. To appear in *Siam J. Numer. Anal.*, 1984.
- [14] T.F. Chan and K. Jackson. *Nonlinearly Preconditioned Krylov Subspace Methods for Discrete Newton Algorithms.* Technical Report 259, Dept. of Computer Science, Yale Univ., 1983. To appear in *Siam J. Sci. Stat. Comp.*, 1984.
- [15] T.F. Chan and H.B. Keller. Arclength Continuation and Multi-Grid Techniques for Nonlinear Eigenvalue Problems. *SIAM J. Sci. Stat. Comp.* 3(2):173-194, June 1982.
- [16] T.F. Chan. On the Existence and Computation of LU-factorizations with Small Pivots. *Math. Comp.* 42 (166) April 1984.
- [17] T.F. Chan and Y. Saad. *Iterative Methods for Solving Bordered Systems with Applications to Continuation Methods.* Technical Report 235, Computer Science Dept., Yale University, 1982. To appear in *Siam J. Sci. Stat. Comp.*, 1984.
- [18] R. Chandra. *Conjugate gradient methods for partial differential equations.* Ph.D. Thesis, Dept. of Computer Science, Yale Univ., 1978.

- [19] B. Chen and P. Saffman. Numerical Evidence for the Existence of New Types of Gravity Waves of Permanent Form on Deep Water. *Studies in Applied Math.* 62:1-21, 1980.
- [20] T.F. Coleman and J.J. More. Estimation of Sparse Jacobian Matrices and Graph Coloring Problems. *Siam J. Numer. Anal.* 20:187-209, 1983.
- [21] P. Concus, G.H. Golub and D.P. O'leary. A Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations . In J.R. Bunch and D.J. Rose, Editor, *Proceedings of the Symposium on Sparse Matrix Computations*, Academic Press, New York, 1975, pp. 309-332.
- [22] A.R. Curtis, M.J.D. Powell and J.K. Reid. On the Estimation of Sparse Jacobian Matrices. *J. Inst. Maths. Applies.* 13:117-119, 1974.
- [23] R.S. Dembo, S. Eisenstat and T. Steihaug. Inexact Newton Methods. *SIAM J. Numer. Anal.* 18(2):400-408, 1982.
- [24] J.E. Dennis and J.J. More. Quasi-Newton Methods, Motivation and Theory. *SIAM Review* 19:46 - 89, 1977.
- [25] J.J. Dongarra, J.R. Bunch, C.B. Moler and G.W. Stewart. *LINPACK User's Guide*. SIAM, Philadelphia, 1979.
- [26] I. Duff. A Survey of Sparse Matrix Research. *Proc. IEEE* 65:500-535, 1977.
- [27] S.C. Eisenstat, H.C. Elman and M.H. Schultz. Variational iterative methods for nonsymmetric systems of linear equations. *SIAM Journal on Numerical Analysis* 20:345-357, 1983.
- [28] S.C. Eisenstat, M.C. Gursky, M.H. Schultz and A.H. Sherman. *Yale sparse matrix package I: The symmetric codes*. Technical Report 112, Dept. of Computer Science, Yale Univ., 1977.
- [29] S.C. Eisenstat, M.C. Gursky, M.H. Schultz and A.H. Sherman. *Yale sparse matrix package II: The nonsymmetric codes*. Technical Report 114, Dept. of Computer Science, Yale Univ., 1977.
- [30] H.C. Elman. *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations*. Ph.D. Thesis, Yale University, 1982. Techreport # 229.
- [31] R. Fletcher and C.M. Reeves. Function Minimization by Conjugate Gradients. *Comput. J.* 7:149-154, 1964.
- [32] C.B. Garcia and W.I. Zangwill. *Pathways to Solutions, Fixed Points and Equilibria*. Prentice-Hall, Englewood Cliffs, N.J., 1981.
- [33] W. C. Gear and Y. Saad. Iterative Solution of Linear Equations in ODE Codes. *Siam J. Sci. Stat. Comp.* 4(4) December 1983.
- [34] A. George and J.W. Liu. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice Hall, New Jersey, USA, 1981.
- [35] K. Georg. On Tracing an Implicitly Defined Curve by Quasi-Newton Steps and Calculating Bifurcation by Local Perturbations. *Siam J. Sci. Stat. Comp.* 2(1) March 1981.
- [36] P.E. Gill, W. Murray and M. Wright. *Practical Optimization*. Academic Press, New York, 1981.
- [37] R. Glowinski, H.B. Keller and L. Reinhart. *Continuation-Conjugate Gradient Methods for the Least Squares Solution of Nonlinear Boundary Value Problems*. Technical Report, Institut National de Recherche en Informatique et en Automatique, 1982.

- [38] G.H. Golub, G.W. Stewart and V. Klema. *Rank Degeneracy and Least Squares Problems*. Technical Report STAN-CS-76-559, Computer Science Dept., Stanford University, 1976.
- [39] W. Hackbusch. Multi-Grid Solution of Continuation Problems. In R. Ansorge, T. Meis and W. Tornig, Editors, *Iterative Solution of Nonlinear Systems*, Springer-Verlag, Berlin, 1982 .
- [40] L.A. Hageman and D.M. Young. *Applied Iterative Methods*. Academic Press, New York, 1981.
- [41] M.R. Hestenes and E. Stiefel. Methods of Conjugate Gradient for solving Linear Systems. *Journal of Research of the National Bureau of Standards* 49:409-436, 1952.
- [42] A.S. Householder. *Theory of Matrices in Numerical Analysis*. Blaisdell Pub. Co., Johnson, Colo., 1964.
- [43] K.C. Jea. *Generalized Conjugate Gradient Acceleration of Iterative Methods*. Ph.D. Thesis, University of Texas at Austin, 1982.
- [44] A. Jepson and A. Spence. Singular Points and Their Computations. In T. Kupper, H. Mittelmann and H. Weber, Editors, *Numerical Methods for Bifurcation Problems*, Birkhauser Verlag, Basel, 1984 .
- [45] R.B. Kearfott. A Derivative-Free Arc Continuation Method and a Bifurcation Technique. In E.L. Allgower, K. Glashoff and H.-O. Peitgen, Editors, *Numerical Solution of Nonlinear Equations*, Springer Verlag, New York, 1981 .
- [46] H.B. Keller. Numerical Solution of Bifurcation and Nonlinear Eigenvalue Problems. In P. Rabinowitz, Editor, *Applications of Bifurcation Theory*, Academic Press, New York, 1977, pp. 359-384.
- [47] H.B. Keller. *Practical Procedures in Path Following Near Limit Points*. 1982. In *Computing Methods in Applied Sciences and Engineering*, eds. Glowinski and Lions, North-Holland Pubs. Inc.
- [48] H.B. Keller and R. Meyer-Spache. Numerical Study of Taylor-Vortex Flows Between Rotating Cylinders. *J. Comp. Phys.* 35(1):100-109, 1980.
- [49] H.B. Keller. The Bordering Algorithm and Path Following Near Singular Points of Higher Nullity. *SIAM J. Sci. and Stat. Comp.* 4(4) 1983.
- [50] H.B. Keller and R. Schreiber. Accurate Solutions for the Driven Cavity. *J. Comp. Phys.* 49(2):310-333, 1983.
- [51] M. Kubicek ad M. Holodniok. Numerical Determination of Bifurcation Points in Steady State and Periodic Solutions - Numerical Algorithms and Examples. In T. Kupper, H. Mittelmann and H. Weber, Editors, *Numerical Methods for Bifurcation Problems*, Birkhauser Verlag, Basel, 1984 .
- [52] T.A. Manteuffel. The Tchebychev Iteration for Nonsymmetric Linear Systems. *Numer. Math.* 28:307-327, 1977.
- [53] R.G. Melhem and W.C. Rheinboldt. A Comparison of Methods for Determining Turning Points of Nonlinear Equations. *Computing* 29:201-228, 1982.
- [54] H.D. Mittelmann. An Efficient Algorithm for Bifurcation Problems of Variational Inequalities. *Math. Comp.* 41:473-485, 1983.
- [55] H.D. Mittelmann and H. Weber, Editors. *Bifurcation Problems and their Numerical Solution*. Birkhauser, Basel, 1980.
- [56] H.D. Mittelmann and H. Weber. *Multi-Grid Solution of Bifurcation Problems*. Technical Report 65, Abteilung Math., Univ. Dortmund, 1983. To appear in Siam J. Sci. Stat. Comp.

- [57] G. Moore and A. Spence. The Calculation of Turning Points of Nonlinear Equations. *SIAM J. Numer. Anal.* 17:567-576, 1980.
- [58] D.P. O'Leary. A Discrete Newton Algorithm for Minimizing a Function of Many Variables. *Math. Progr.* 23:20 - 33, 1982.
- [59] J.M. Ortega and W.C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, New York, 1970.
- [60] G. Pönisch and H. Schwetlick. Computing Turning Points of Curves Implicitly Defined by Nonlinear Equations Depending on a Parameter. *Computing* 26:107-121, 1981.
- [61] M.J.D. Powell and Ph. L. Toint. On the Estimation of Sparse Hessian Matrices. *SIAM J. Numer. Anal.* 16:1060-1073, 1979.
- [62] G.W. Reddien. Computation of Turning and Bifurcation Points for Two Point Boundary Value Problems. In T. Kupper, H. Mittelmann and H. Weber, Editors, *Numerical Methods for Bifurcation Problems*, Birkhauser Verlag, Basel, 1984 .
- [63] W.C. Rheinboldt and J.V.Burkardt. A Locally Parameterized Continuation Process. *ACM Trans. Math. Soft.* 9(2):215-235, June 1983.
- [64] W.C. Rheinboldt. Computation of Critical Boundaries on Equilibrium Manifolds. *Siam J. Numer. Anal.* 19:653-669, 1982.
- [65] W.C. Rheinboldt. Numerical Analysis of Continuation Methods for Nonlinear Structural Problems. *Computers and Structures* 13:103-113, 1981.
- [66] Y. Saad. Krylov Subspace Methods for Solving Large Unsymmetric Linear Systems. *Math. Comp.* 37:105-126, 1981.
- [67] H. Schwetlick. Algorithms for Finite Dimensional Turning Point Problems from Viewpoint to Relationships with Constrained Optimization Methods. In T. Kupper, H. Mittelmann and H. Weber, Editors, *Numerical Methods for Bifurcation Problems*, Birkhauser Verlag, Basel, 1984 .
- [68] R. Seydel. Numerical Computation of Branch Points in Nonlinear Equations. *Numer. Math.* 33:339-352, 1979.
- [69] G.W. Stewart. On the Implicit Deflation of Nearly Singular Systems of Linear Equations. *SIAM J. Sci. Stat. Comp.* 2(2):136-140, 1981.
- [70] K. Stuben and U. Trottenberg. Multi-Grid Methods: Fundamental Algorithms, Model Problem Analysis and Applications. In W. Hackbusch adn U. Trottenberg, Editors, *Multigrid Methods*, Springer Verlag, Berlin, 1982 .
- [71] K.K. Tung, T.F. Chan and T. Kubota. Large Amplitude Internal Waves of Permanent Form. *Studies in Applied Math.* 66:1-44, February 1982.
- [72] P.K.W. Vinsome. ORTHOMIN, an iterative method for solving sparse sets of simultaneous linear equations. In *Proceedings of the Fourth Symposium on Reservoir Simulation*, Society of Petroleum Engineers of AIME, , 1976 , pp. 149-159.
- [73] H. Wacker, Editors. *Continuation Methods*. Academic Press, New York, 1978.
- [74] J.H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford Univ. Press, London, 1965.
- [75] David M. Young and Kang C. Jea. Generalized conjugate gradient acceleration of nonsymmetrizable iterative methods. *Linear Algebra and Its Applications* 34:159-194, 1980.

THE CALCULATION OF HIGH ORDER SINGULARITIES IN THE FINITE TAYLOR PROBLEM

K.A.Cliffe and A.Spence

1. INTRODUCTION

This paper is concerned with the numerical calculation of solutions of the Navier-Stokes equations for steady axisymmetric flow in the Taylor problem with small aspect ratio. A recent theoretical and experimental study was carried out by Benjamin and Mullin [1] and numerical results, which are in good agreement with those in [1], are given by Cliffe [3]. These results are summarized in figure 2.1 which gives the projection of paths of singular points on a certain parameter space.

There are certain symmetry properties in the Taylor problem which were not utilized in [3] and one aim of this paper is to describe how the results in [3] can be produced more efficiently using a simpler algorithm, derived in [10], which takes account of the underlying symmetry in the problem.

Also of interest in the Taylor problem are two 'high order' singularities (this term is made precise in section 3), which are important since they mark the onset and loss of hysteresis in the problem. These are denoted by B and C in figure 2.1 and correspond to (codimension one) singularities described in [5], [6]. The second aim of this paper is to derive and analyze numerical methods for the efficient computation of these high order singularities.

The plan of the paper is as follows. Section 2 contains a description of the Taylor problem with small aspect ratio and the details of the finite-element method used for its numerical solution. The main theoretical results are contained in section 3, where a general nonlinear two parameter problem with symmetry (see condition (3.2)) is considered. Numerical methods for the computation of certain high order singular points are derived and analyzed. Section 4 describes how the algorithms are implemented in the Taylor problem and numerical results are given in section 5.

2. THE TAYLOR PROBLEM

In the Taylor problem the annular gap between two concentric circular cylinders is filled with a viscous fluid. In the case we are considering the outer cylinder is fixed and the inner one rotates. The ends of the annulus are stationary. If the speed of the inner cylinder is sufficiently low the flow is axisymmetric. Let r_1 and r_2 be the radii of the inner and outer cylinders respectively and let ℓ be their length. Let Ω be the angular speed of the inner cylinder. We use cylindrical polar coordinates (r^*, ϕ, z^*) with origin midway between the ends and denote the velocity by $\mathbf{U}^* = (u_r^*, u_\phi^*, u_z^*)$. The equations for axisymmetric flow of a viscous fluid are

$$R\left(\Gamma u_r \frac{\partial u_r}{\partial r} + u_z \frac{\partial u_r}{\partial z} - \Gamma \frac{u_\phi^2}{r + \beta}\right) + \Gamma \frac{\partial p}{\partial r} - \frac{\Gamma}{r + \beta} \frac{\partial}{\partial r} 2(r + \beta) \frac{\partial u_r}{\partial r} - \frac{\partial}{\partial z} \left(\frac{1}{\Gamma} \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r}\right) + \Gamma \frac{u_r}{(r + \beta)^2} = 0, \quad (2.1)$$

$$R\left(\Gamma u_r \frac{\partial u_\phi}{\partial r} + u_z \frac{\partial u_\phi}{\partial z} + \Gamma \frac{u_r u_\phi}{r + \beta}\right) - \frac{\Gamma}{r + \beta} \frac{\partial}{\partial r} (r + \beta) \frac{\partial u_\phi}{\partial r} - \frac{1}{\Gamma} \frac{\partial^2 u_r}{\partial z^2} + \Gamma \frac{u_\phi}{(r + \beta)^2} = 0, \quad (2.2)$$

$$R\left(\Gamma u_r \frac{\partial u_z}{\partial r} + u_z \frac{\partial u_z}{\partial z}\right) + \frac{\partial p}{\partial z} - \frac{1}{r + \beta} \frac{\partial}{\partial r} (r + \beta) \left(\Gamma \frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z}\right) - \frac{2}{\Gamma} \frac{\partial^2 u_z}{\partial r^2} = 0, \quad (2.3)$$

$$\frac{\Gamma}{r + \beta} \frac{\partial}{\partial r} (r + \beta) u_r + \frac{\partial u_z}{\partial z} = 0. \quad (2.4)$$

In the above equations r, z, \mathbf{U} and p are given by

$$r = \frac{r^*}{d} - \beta, \quad z = \frac{z^*}{\ell}, \quad \mathbf{U} = \frac{\mathbf{U}^*}{r_1 \Omega}, \quad p = \frac{dp^*}{\mu r_1 \Omega},$$

where $d = r_1 - r_2$, $\beta = r_1/d = \eta/(1 - \eta)$, $\eta = r_1/r_2$, $\Gamma = \ell/d$ and $R = \rho r_1 \Omega d/\mu$ with ρ and μ the fluid density and viscosity respectively. Thus the problem has one dynamical parameter R , the *Reynolds number*, and two geometrical parameters Γ , the *aspect ratio*, and η , the *radius ratio*.

Equations (2.1)–(2.4) hold in the region

$$D = \{(r, z) | 0 \leq r \leq 1, -0.5 \leq z \leq 0.5\},$$

and, for later use, we define the subregion

$$D^+ = \{(r, z) | 0 \leq r \leq 1, 0 \leq z \leq 0.5\}.$$

The boundary conditions are that u_r and u_z are zero on the entire boundary of D , u_ϕ is zero on the outer cylinder ($r = 1$) and one on the inner cylinder ($r = 0$). On the ends ($z = \pm 0.5$) u_ϕ is zero except near the inner cylinder where it increases smoothly to one over a small distance (see [1] and [3] for more details).

For theoretical purposes, and also as a starting point for the finite-element discretization, it is convenient to convert equations (2.1)–(2.4) into an operator equation in an appropriate Hilbert space. We introduce the following notation: let $L^2(D)$ be the space of

functions which are square integrable over D ; let $W^{1,2}(D)$ be the space of functions whose generalized first derivatives lie in $L^2(D)$ and let $W_0^{1,2}(D)$ be that subspace of $W^{1,2}(D)$ whose elements vanish (weakly) on the boundary of D . $W^{1,2}(D)^3$ is the space of vector valued functions each component of which is in $W^{1,2}(D)$.

We introduce the following three functionals

$$\begin{aligned} a_1(\mathbf{U}; \mathbf{V}, \mathbf{W}) = R \int_D & \left[\left\{ (r + \beta) \left(\Gamma u_r \frac{\partial v_r}{\partial r} + u_z \frac{\partial v_r}{\partial z} \right) - \Gamma u_\phi v_\phi \right\} v_r \right. \\ & + \left\{ (r + \beta) (\Gamma u_r \frac{\partial v_\phi}{\partial r} + u_z \frac{\partial v_\phi}{\partial z}) + \Gamma u_r v_\phi \right\} w_\phi \\ & \left. + (r + \beta) \left(\Gamma u_r \frac{\partial v_z}{\partial r} + u_z \frac{\partial v_z}{\partial z} \right) v_z \right], \end{aligned} \quad (2.5)$$

$$\begin{aligned} a_0(\mathbf{U}, \mathbf{V}) = \int_D & \left[2\Gamma(r + \beta) \frac{\partial u_r}{\partial r} \frac{\partial v_r}{\partial r} + (r + \beta) \left(\frac{1}{\Gamma} \frac{\partial u_r}{\partial z} + \frac{\partial u_z}{\partial r} \right) \right. \\ & + \Gamma \frac{u_r v_r}{r + \beta} + \Gamma(r + \beta) \frac{\partial u_\phi}{\partial r} \frac{\partial v_\phi}{\partial r} + \frac{r + \beta}{\Gamma} \frac{\partial u_\phi}{\partial z} \frac{\partial v_\phi}{\partial z} \\ & \left. + \Gamma \frac{u_\phi v_\phi}{r + \beta} + (r + \beta) \left(\Gamma \frac{\partial u_z}{\partial r} + \frac{\partial u_r}{\partial z} \right) \frac{\partial v_z}{\partial r} + 2 \frac{r + \beta}{\Gamma} \frac{\partial u_z}{\partial z} \frac{\partial v_z}{\partial z} \right], \end{aligned} \quad (2.6)$$

$$b(q, \mathbf{V}) = - \int_D \left[\Gamma q \frac{\partial}{\partial r} (r + \beta) v_r + q(r + \beta) \frac{\partial v_z}{\partial z} \right]. \quad (2.7)$$

In (2.5)–(2.7) we have suppressed the dependence on R , Γ and β .

Let $\hat{\mathbf{U}} \in W^{1,2}(D)^3$ be a vector valued function which satisfies the boundary conditions of the problem. We can now define the operator

$$\mathbf{A} : W_0^{1,2}(D)^3 \times L^2(D) \times \mathbf{R} \times \mathbf{R} \times \mathbf{R} \rightarrow W_0^{1,2}(D)^3 \times L^2(D),$$

by $\mathbf{A}(\mathbf{U}, p, R, \Gamma, \beta) = (\mathbf{W}, r)$ where

$$\begin{aligned} a_1(\hat{\mathbf{U}}, \hat{\mathbf{U}}, \mathbf{V}) + a_1(\hat{\mathbf{U}}; \mathbf{U}, \mathbf{V}) + a_1(\mathbf{U}; \hat{\mathbf{U}}, \mathbf{V}) + a_1(\mathbf{U}; \mathbf{U}, \mathbf{V}) + \\ + a_0(\hat{\mathbf{U}}, \mathbf{V}) + a_0(\mathbf{U}, \mathbf{V}) + b(p, \mathbf{V}) = \langle \mathbf{W}, \mathbf{V} \rangle \quad \text{for all } \mathbf{V} \in W_0^{1,2}(D)^3, \end{aligned} \quad (2.8)$$

$$b(q, \hat{\mathbf{U}}) + b(q, \mathbf{U}) = \langle q, r \rangle \quad \text{for all } q \in L^2(D), \quad (2.9)$$

and $\langle \cdot, \cdot \rangle$ denotes the inner product in a Hilbert space. The problem of solving equations (2.1)–(2.4) is equivalent, under certain smoothness conditions, to solving the operator equation

$$\mathbf{A}(\mathbf{U}, p, R, \Gamma, \beta) = \mathbf{0}, \quad (2.10)$$

(where we reintroduce the dependence on R, Γ and β).

We turn now to the symmetry properties of the operator \mathbf{A} . We define $S \in \mathcal{L}(W^{1,2}(D)^3 \times L^2(D))$ by

$$S\{u_r(r,z), u_\phi(r,z), u_z(r,z), p(r,z)\} = \{u_r(r,-z), u_\phi(r,-z), -u_z(r,-z), p(r,-z)\} \quad (2.11)$$

for smooth functions, and use continuity to extend the definition to the whole space. It is easy to see that $S \neq I$ and that $S^2 = I$ (cf. condition (3.2)). We note that $\hat{\mathbf{U}}$ may be chosen so that

$$S\hat{\mathbf{U}} = \hat{\mathbf{U}}, \quad (2.12)$$

and we assume that this has been done. A simple change of variable in the integrals in (2.5)–(2.7) gives the following results

$$a_1(S\mathbf{U}; S\mathbf{V}, \mathbf{W}) = a_1(\mathbf{U}; \mathbf{V}, S\mathbf{W}), \quad (2.13)$$

$$a_0(S\mathbf{U}, \mathbf{V}) = a_0(\mathbf{U}, S\mathbf{V}), \quad (2.14)$$

$$b(Sp, \mathbf{U}) = b(p, S\mathbf{U}). \quad (2.15)$$

Similarly we have

$$\langle S\mathbf{U}, \mathbf{V} \rangle = \langle \mathbf{U}, S\mathbf{V} \rangle, \quad (2.16)$$

$$\langle Sp, q \rangle = \langle p, Sq \rangle. \quad (2.17)$$

Using (2.8),(2.9),(2.12)–(2.17) it follows that (cf. condition (3.2))

$$\mathbf{A}(S(\mathbf{U}, p), R, \Gamma, \beta) = S\mathbf{A}(\mathbf{U}, p, R, \Gamma, \beta). \quad (2.18)$$

Now consider the problem of obtaining a finite-element approximation to the solution of (2.1)–(2.4) (equivalently (2.18)). For each $h > 0$ let W_h and M_h be two finite-dimensional spaces such that $W_h \subset W^{1,2}(D)^3$, $M_h \subset L^2(D)$, and let $W_{h,0} = W_h \cap W_0^{1,2}(D)^3$. Let $\hat{\mathbf{U}}_h \in W_h$ be a vector valued function that approximates the boundary conditions of the problem and satisfies $b(1, \hat{\mathbf{U}}_h) = 0$. The finite-element approximation of \mathbf{A} ,

$$\mathbf{A}_h : W_{h,0} \times M_h \times \mathbf{R} \times \mathbf{R} \times \mathbf{R} \rightarrow W_{h,0} \times M_h,$$

is defined by $\mathbf{A}_h(\mathbf{U}_h, p_h, R, \Gamma, \beta) = (\mathbf{W}_h, r_h)$ where

$$\begin{aligned} & a_1(\hat{\mathbf{U}}_h, \hat{\mathbf{U}}_h, \mathbf{V}_h) + a_1(\hat{\mathbf{U}}_h; \mathbf{U}_h, \mathbf{V}_h) + a_1(\mathbf{U}_h; \hat{\mathbf{U}}_h, \mathbf{V}_h) + \\ & + a_0(\hat{\mathbf{U}}_h, \mathbf{V}_h) + a_0(\mathbf{U}_h, \mathbf{V}_h) + b(p_h, \mathbf{V}_h) = \langle \mathbf{W}_h, \mathbf{V}_h \rangle \quad \text{for all } \mathbf{V}_h \in W_{h,0}, \end{aligned} \quad (2.19)$$

$$b(q_h, \hat{\mathbf{U}}_h) + b(q_h, \mathbf{U}_h) = \langle q_h, r_h \rangle \quad \text{for all } q_h \in M_h. \quad (2.20)$$

If the finite-element mesh is symmetric it follows that $S \in \mathcal{L}(W_h \times M_h)$ and that

$$\mathbf{A}_h(S(\mathbf{U}_h, p_h), R, \Gamma, \beta) = S\mathbf{A}_h(\mathbf{U}_h, p_h, R, \Gamma, \beta). \quad (2.21)$$

It is easy to show, by a change of variable, that the following relationships hold

$$a_1(\mathbf{U}_h, \mathbf{V}_h, \mathbf{W}_h) = a_1^+(\mathbf{U}_h; \mathbf{V}_h, \mathbf{W}_h) + a_1^+(S\mathbf{U}_h; S\mathbf{V}_h, S\mathbf{W}_h), \quad (2.22)$$

$$a_0(\mathbf{U}_h, \mathbf{V}_h) = a_0^+(\mathbf{U}_h, \mathbf{V}_h) + a_0^+(S\mathbf{U}_h, S\mathbf{V}_h), \quad (2.23)$$

$$b(p_h, \mathbf{V}_h) = b^+(p_h, \mathbf{V}_h) + b^+(Sp_h, S\mathbf{V}_h), \quad (2.24)$$

where the superscript + indicates that the integral is over D^+ rather than D . An important consequence of (2.22)–(2.24) is that, if $\mathbf{U}_h, \mathbf{V}_h, \mathbf{W}_h, p_h$ are eigenvectors of S , that is if they are either

symmetric or antisymmetric, then the problem only involves integrals over D^+ and this fact can be used to reduce the number of degrees of freedom (see section 4).

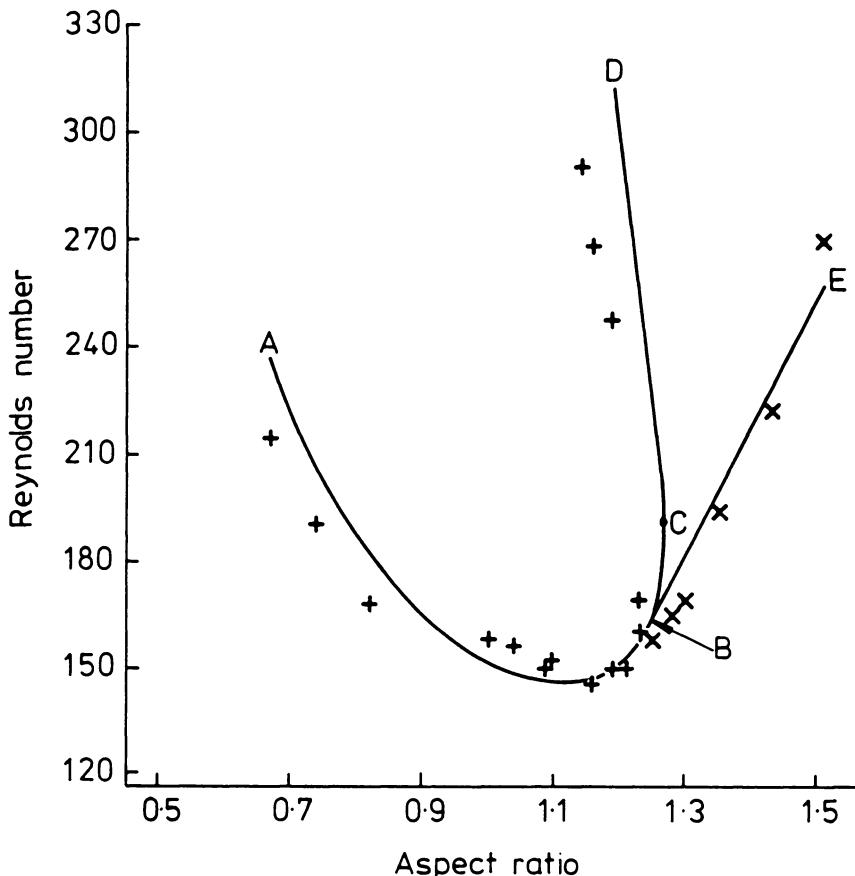


Figure (2.1). Bifurcation set for the Taylor problem: —, numerical predictions; +, experimental values for symmetry breaking bifurcation points; \times , experimental values for turning points (from Benjamin and Mullin [1]).

The particular aspect of the Taylor problem we are studying in this paper is the nature of the solution set when the length of the annular region is comparable to the difference in the cylinder radii. Benjamin and Mullin [1] have done some very elegant experiments for this case; they were able to vary the Reynolds number and the aspect ratio in their apparatus and, by flow visualization techniques, they measured the bifurcation set for the problem when $0.6 \leq \Gamma \leq 1.5$. In [1] the radius ratio was fixed at 0.615 which corresponds to $\beta = 1.597$; for the rest of this paper we suppress the dependence on β of equations like (2.18). The results of a numerical study [3] are

shown in figure 2.1. The line ABCD is the projection on the $R-\Gamma$ plane of a path of symmetry breaking bifurcations. The line BE is the projection of a path turning points which splits off from the path of symmetry breaking bifurcations at the point B. The bifurcation diagrams near B are like those in figure 3.2. At point C two symmetry breaking bifurcations coalesce giving bifurcation diagrams like those in figure 3.1. Numerically calculated diagrams are given in [3] and section 5.

The numerical methods used in [3] to calculate the symmetry breaking bifurcation points did not make explicit use of the symmetry. Also no attempt was made to calculate the higher codimension singularities occurring at B and C accurately. This paper extends [3] by using methods which exploit the symmetry and by developing and justifying algorithms for the singularities at B and C of figure 2.1.

3. SINGULAR POINTS OF TWO PARAMETER PROBLEMS WITH SYMMETRY

In this section we discuss the computation of singular points of a nonlinear two parameter dependent problem of the form

$$\mathbf{f}(\mathbf{x}, \lambda, \mu) = \mathbf{0}, \quad \mathbf{f} : \mathbf{X} \times \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{X}, \quad (3.1)$$

where \mathbf{x} represents state variable, λ and μ represent real parameters, and \mathbf{f} is a smooth mapping on a Banach space \mathbf{X} . (In section 4 \mathbf{X} is finite dimensional and thus isomorphic to \mathbf{R}^N ; \mathbf{f} is given by (4.7).) We assume that \mathbf{f} satisfies the symmetry relation

$$\begin{aligned} &\text{There exists an } S \in \mathcal{L}(\mathbf{X}) \text{ satisfying } S \neq I, \quad S^2 = I, \text{ and} \\ &\mathbf{f}(S\mathbf{x}, \lambda, \mu) = S\mathbf{f}(\mathbf{x}, \lambda, \mu), \quad \mathbf{x} \in \mathbf{X}, \quad \lambda, \mu \in \mathbf{R}. \end{aligned} \quad (3.2)$$

The approach is based on the theory of one parameter problems with symmetry described in [10] (see also [11]).

First we recall some of the results in [10]. The mapping S induces a natural decomposition of \mathbf{X} into

$$\mathbf{X} = \mathbf{X}_s \oplus \mathbf{X}_a, \quad (3.3a)$$

where

$$\mathbf{X}_s := \{\mathbf{x} \in \mathbf{X} | S\mathbf{x} = \mathbf{x}\}, \quad \mathbf{X}_a := \{\mathbf{x} \in \mathbf{X} | S\mathbf{x} = -\mathbf{x}\}, \quad (3.3b)$$

consist of the *symmetric* and *antisymmetric* elements of \mathbf{X} respectively. An easy consequence of (3.3) is that, for $\mathbf{x} \in \mathbf{X}_s$,

$$\mathbf{f}, \mathbf{f}_\lambda, \mathbf{f}_\mu \in \mathbf{X}_s, \quad (3.4a)$$

$$\mathbf{X}_s \text{ and } \mathbf{X}_a \text{ are invariant with regard to } \mathbf{f}_x, \mathbf{f}_\lambda, \mathbf{f}_\mu, \quad (3.4b)$$

$$\text{for } \mathbf{v}, \mathbf{w} \in \mathbf{X}_s \text{ or } \mathbf{v}, \mathbf{w} \in \mathbf{X}_a, \quad \mathbf{f}_{xx}\mathbf{v}\mathbf{w} \in \mathbf{X}_s, \quad (3.4c)$$

$$\text{for } \mathbf{v} \in \mathbf{X}_s, \mathbf{w} \in \mathbf{X}_a, \quad \mathbf{f}_{xx}\mathbf{v}\mathbf{w} \in \mathbf{X}_a. \quad (3.4d)$$

The point $(\mathbf{x}_0, \lambda_0, \mu_0)$ is a *singular* point of (3.1) if $\mathbf{f}_x^0 := \mathbf{f}_x(\mathbf{x}_0, \lambda_0, \mu_0)$ is singular, otherwise $(\mathbf{x}_0, \lambda_0, \mu_0)$ is a *regular* point. A singular point is *simple* if

$$\text{Null}(\mathbf{f}_x^0) = \text{Span}\{\phi_0\}; \quad \text{Range}(\mathbf{f}_x^0) = \{y \in X | \psi_0 y = 0\}, \quad \psi_0 \in X'. \quad (3.5)$$

In this paper we will only consider simple singular points. It is easy to show that, if $x_0 \in X_s$, then $\phi_0 \in X_a$, or $\phi_0 \in X_s$. We shall discuss only the latter case, namely, the *symmetry breaking* case

$$x_0 \in X_s, \quad \phi_0 \in X_a. \quad (3.6a)$$

Under the above assumptions it can be shown that

$$\psi_0 x = 0, \quad \text{for all } x \in X_s. \quad (3.6b)$$

Consider first the case μ fixed i.e. consider the one parameter problem

$$\mathbf{g}(x, \lambda) := \mathbf{f}(x, \lambda, \mu) = \mathbf{0}. \quad (3.7)$$

It is a straightforward consequence of (3.4) and (3.6) that

$$\psi_0 g_\lambda^0 = 0, \quad \psi_0 g_{xx}^0 \phi_0 \phi_0 = 0. \quad (3.8)$$

(Here $g_\lambda^0 := g_\lambda(x_0, \lambda_0)$, etc.) Standard bifurcation theory now gives that (x_0, λ_0) is a *bifurcation point* of (3.7) if and only if

$$b_\lambda := \psi_0(g_{\lambda x}^0 \phi_0 + g_{xx}^0 \phi_0 v_\lambda) \neq 0, \quad (3.9)$$

where

$$g_x^0 v_\lambda + g_\lambda^0 = 0, \quad v_\lambda \in X_s.$$

(Note that, since $\phi \in X_a$, g_x^0 is an isomorphism on X_s .) If (3.9) holds we call (x_0, λ_0) a *symmetry breaking bifurcation point* of \mathbf{g} . The main aim of [10] was to show how such points could be computed in a stable, efficient way. In that paper an *extended* system was introduced, namely,

$$\mathbf{G}(y) := \begin{cases} \mathbf{g}(x, \lambda), & y = (x, \phi, \lambda) \in Y, \\ g_x(x, \lambda, \mu)\phi, & Y := X_s \times X_a \times \mathbb{R}, \\ I\phi - 1, & G : Y \rightarrow Y, \end{cases} \quad (3.10)$$

where $I \in X'$ is chosen such that $I\phi_0 = 1$. The point (x_0, ϕ_0, λ_0) is a regular point of \mathbf{G} , regarded as a mapping on Y , if and only if (x_0, λ_0) is a symmetry breaking bifurcation point (Theorem 3.1 of [10]). Hence Newton's method can be applied directly to $\mathbf{G}(y) = \mathbf{0}$. Some examples of the use of this result, which show, in particular, how best to utilize the symmetry in a problem, are given in [10]. (See also section 4.)

Now let us reconsider the two parameter problem (3.1). First we introduce the extended system corresponding to (3.10), namely,

$$\mathbf{F}(y, \mu) := \begin{cases} \mathbf{f}(x, \lambda, \mu), & y = (x, \phi, \lambda) \in Y, \\ f_x(x, \lambda, \mu)\phi, & F : Y \times \mathbb{R} \rightarrow Y. \\ I\phi - 1, & \end{cases} \quad (3.11)$$

If (3.7) holds then (y_0, μ_0) is a regular point of (3.11) and the Implicit Function Theorem ensures that there is a neighbourhood of (y_0, μ_0) in which y can be parametrized by μ . Hence standard continuation methods can be used to compute a path of symmetry breaking bifurcation points (see

section 4). This idea of computing paths of singular points of (3.1) by first introducing an extended system, which characterizes the singular point, was used in [3], [9], [7] and is discussed in [8].

An alternative approach to analyzing the singularities of (3.1) and (3.7) is to try a Lyapunov–Schmidt reduction procedure. For example for (3.7) we can analyze the behaviour of the solutions of $\mathbf{g}(\mathbf{x}, \lambda) = \mathbf{0}$ near $(\mathbf{x}_0, \lambda_0)$ by making the substitutions

$$\lambda = \lambda_0 + \hat{\lambda}, \quad \mathbf{x} = \mathbf{x}_0 + \hat{\mathbf{x}}\Phi_0 + \mathbf{w}, \quad \Psi_0 \mathbf{w} = 0, \quad \hat{\mathbf{x}} \in \mathbb{R}, \quad (3.12)$$

to obtain the reduced equation

$$\hat{\mathbf{g}}(\hat{\mathbf{x}}, \hat{\lambda}) := \left(\frac{d}{6} \right) \hat{\mathbf{x}}^3 + b_\lambda \hat{\lambda} \hat{\mathbf{x}} + h.o.t. = 0, \quad \hat{\mathbf{g}} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}, \quad (3.13)$$

where b_λ is given by (3.9), and

$$d := \Psi_0(\mathbf{g}_{xxx}^0 \Phi_0 \Phi_0 \Phi_0 + 3\mathbf{g}_{xx}^0 \Phi_0 \mathbf{z}_0), \quad (3.14)$$

with

$$\mathbf{g}_x^0 \mathbf{z}_0 = - \mathbf{g}_{xx}^0 \Phi_0 \Phi_0. \quad (3.15)$$

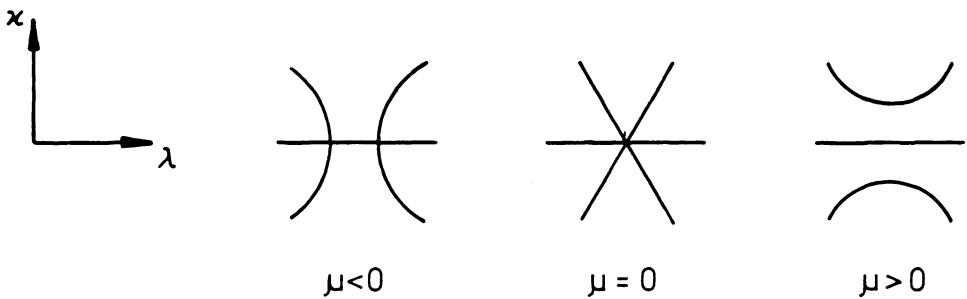
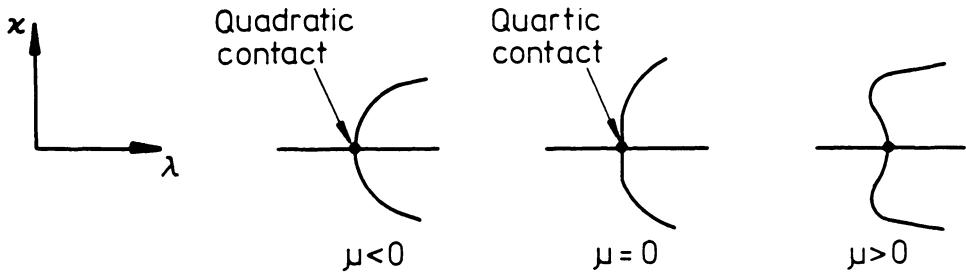
Note that $\hat{\mathbf{g}}(-\hat{\mathbf{x}}, \hat{\lambda}) = -\hat{\mathbf{g}}(\hat{\mathbf{x}}, \hat{\lambda})$, (see Lemma 8 of [2]) and so (3.13) exhibits $\mathbf{Z}_2 = \{\pm 1\}$ symmetry with respect to $\hat{\mathbf{x}}$. The results of [6] and [5] are particularly relevant now. If $d \neq 0$, $b_\lambda \neq 0$, then $\hat{\mathbf{g}}$ is contact equivalent to the canonical form $x^3 - \lambda x = 0$, and so is stable with respect to \mathbf{Z}_2 symmetry preserving perturbations, i.e. $\hat{\mathbf{g}}$ has codimension zero. If one of b_λ and d is zero then the singularity is of higher order (i.e. higher codimension), and we shall be concerned with these cases in this paper. The canonical forms for the codimension one singularities are $x^3 - \lambda^2 x = 0$ and $x^5 - \lambda x = 0$ with universal \mathbf{Z}_2 unfoldings $x^3 - \lambda^2 x + \mu x = 0$ and $x^5 - \mu x^3 - \lambda x = 0$ respectively. The solution diagrams for these two cases are given in figures 3.1 and 3.2 (cf. [6], [5]).

Now solution diagrams similar to these appear in the Taylor problem (see [1], [3]) and so it is reasonable to assume that the singularities there are of the above type. In the rest of this section we derive and analyze numerical methods for the computation of these high order singularities (codimension one singularities) for problems like (3.1) which satisfy (3.2). In fact the theory produces precise expressions which, in some cases, can be evaluated to confirm that the singularities are indeed of the above types. We test the methods on the Taylor problem and give numerical details and results in sections 4 and 5.

3.1 The point of coalescence

For the singularity in figure 3.1, which we call a *point of coalescence*, there are two symmetry breaking bifurcation points for $\mu < 0$, and none for $\mu > 0$. Thus if we consider the problem (3.1) and the the extended system (3.11) it is reasonable to ask if (3.11) has a turning point at (\mathbf{y}_0, μ_0) . Clearly we must assume that, at $(\mathbf{x}_0, \lambda_0, \mu_0)$,

$$b_\lambda = \Psi_0(\mathbf{f}_{xx}^0 \Phi_0 + \mathbf{f}_{xx}^0 \Phi_0 \mathbf{v}_\lambda) = 0, \quad (3.16)$$

Figure 3.1 $x^3 - \lambda^2 x + \mu x = 0$ Figure 3.2 $x^5 - \mu x^3 - \lambda x = 0$

since there is not a continuous path of regular solutions of (3.11) through μ_0 . We have the following theorem.

THEOREM 3.1 Assume (3.16) holds. (y_0, μ_0) is a simple turning point of $\mathbf{F}(y, \mu) = 0$ if and only if

$$b_\mu := \psi_0(\mathbf{f}_{\mu x}^0 \Phi_0 + \mathbf{f}_{xx}^0 \Phi_0 \mathbf{v}_\mu) \neq 0, \quad (3.17)$$

where

$$\mathbf{f}_x^0 \mathbf{v}_\mu + \mathbf{f}_\mu^0 = 0, \quad \mathbf{v}_\mu \in X_s. \quad (3.18)$$

Proof. The proof is straightforward. We only mention that, with

$$\mathbf{f}_x^0 \mathbf{w}_\lambda = -(\mathbf{f}_{\lambda x}^0 + \mathbf{f}_{xx}^0 \Phi_0 \Phi_0), \quad \mathbf{w}_\lambda \in X_s, \quad (3.19)$$

and

$$\zeta_0 \mathbf{f}_x^0 = -\psi_0 \mathbf{f}_{xx}^0 \Phi_0, \quad \zeta_0 \in (X')_s, \quad (3.20)$$

we can define Φ_0, Ψ_0 by

$$\Phi_0 = \begin{pmatrix} v_\lambda \\ w_\lambda \\ 1 \end{pmatrix}, \quad \Psi_0 = (\zeta_0, \psi_0, 0), \quad (3.21)$$

and these elements satisfy

$$Null(\mathbf{F}_y^0) = Span\{\Phi_0\}, \quad Range(\mathbf{F}_y^0) = \{z \in Y | \Psi_0 z = 0\}. \quad (3.22)$$

Condition (3.17) is precisely $\Psi_0 \mathbf{F}_y^0 \neq 0$. ■

The condition for the turning point to be quadratic (i.e. non-degenerate) is

$$\Psi_0 \mathbf{F}_{yy} \Phi_0 \Phi_0 \neq 0, \quad (3.23)$$

and this can be given in terms of the derivatives of f .

The above results provide us with several options for the computation of (x_0, λ_0, μ_0) . First, standard turning point methods could be applied directly to (3.11). Second we note that we could interchange the rôles of the parameters λ and μ , by rewriting (3.11) as

$$\hat{\mathbf{F}}(z, \lambda) := \mathbf{F}(y, \mu) = 0, \quad z = (x, \phi, \mu). \quad (3.24)$$

Now, using (3.13), $\hat{\mathbf{F}}_z^0$ has an isolated solution at $(z_0, \lambda_0) = (y_0, \mu_0)$ and the singularity has been 'removed'. To find the actual point of coalescence we need to find a zero of $\Psi_0(f_{xx}^0 \Phi_0 + f_{xx}^0 \Phi_0 v_\lambda)$. A third alternative is to set up a system which has (x_0, λ_0, μ_0) as part of an isolated root. Newton's method, or a Newton like method, could then be used with a starting value provided by a suitable point on the path of symmetry breaking bifurcation points. This is the approach we concentrate on here.

The system we use is obtained by adding to (3.11) equations equivalent to (3.16), namely,

$$\mathbf{H}(s) := \begin{pmatrix} f \\ f_x \phi \\ l\phi - 1 \\ f_x v_\lambda + f_\lambda \\ f_x w_\lambda + f_{xx} \phi + f_{xx} \phi v_\lambda \\ l w_\lambda \end{pmatrix} = \mathbf{0}, \quad s = (x, \phi, \mu, v_\lambda, w_\lambda, \lambda) \in Z, \quad Z = X_s \times X_a \times R \times X_s \times X_a \times R. \quad (3.25)$$

The last equation in (3.25) simply makes $w_\lambda \in X_a$ unique. We have the following theorem.

THEOREM 3.2 Let (x_0, λ_0, μ_0) satisfy (3.16) and (3.17). Then $\mathbf{H}(s) = 0$ has an isolated solution at $s_0 = (x_0, \phi_0, \mu_0, v_\lambda, w_\lambda, \lambda_0)$ if and only if $\Psi_0 \mathbf{F}_{yy}^0 \Phi_0 \Phi_0 = 0$ where Ψ_0 and Φ_0 are given by (3.21).

We omit the proof which is straightforward but tedious. System (3.25) was used to compute the point of coalescence in the Taylor problem (see section 5).

We end this subsection with the remark that, under assumptions (3.2), (3.16) and (3.17) we can show that the Lyapunov–Schmidt reduction applied to (3.1) gives

$$\hat{f}(\hat{x}, \hat{\lambda}, \hat{\mu}) := \left(\frac{d}{6} \right) \hat{x}^3 + \frac{1}{2} (\Psi_0 F_{yy}^0 \Phi_0 \Phi_0) \hat{\lambda}^2 \hat{x} + b_\lambda \hat{\mu} \hat{x} + h.o.t. = 0,$$

see (3.14) and (3.17). Hence, provided $d \cdot b_\mu \cdot \Psi_0 F_{yy}^0 \Phi_0 \Phi_0 \neq 0$, $\hat{f}(\hat{x}, \hat{\lambda}, \hat{\mu})$ is contact equivalent to $x^3 \pm \lambda^2 x - \mu x = 0$ (see figure 3.1) and so we can say that $f(x, \lambda, \mu) = 0$ is a universal Z_2 unfolding of $f(x_0, \lambda_0, \mu_0) = 0$, and that (x_0, λ_0, μ_0) is a point of coalescence of (3.1). Note that there are two types of solution behaviour (see figure 4.2 in [5]) depending on the sign of the $\lambda^2 x$ term. Also we can have confidence in the stability of any numerical method for the calculation of (x_0, λ_0, μ_0) provided the Z_2 symmetry is preserved. The actual implementation of Newton's method for (3.25) is a straightforward extension of that described in [10] and is discussed briefly in section 4.

3.2 The quartic bifurcation point

The singularity in figure 3.2, which we call a *quartic* symmetry breaking bifurcation point, differs from the point of coalescence since (3.9) holds and so F_y^0 is nonsingular at (x_0, λ_0, μ_0) . In fact the condition that the contact is quartic (see figure 3.2) is

$$d := \psi_0 (F_{xx}^0 \Phi_0 \Phi_0 \Phi_0 + 3F_{xx}^0 \Phi_0 z_0) = 0, \quad (3.26)$$

with z_0 given by (3.15). Clearly one could use this condition to provide the basis for a numerical procedure to find a quartic bifurcation point. However, as in the previous section we prefer to introduce a system of equations which has an isolated root at the quartic bifurcation point.

First we remark that if a Lyapunov–Schmidt decomposition of (3.1) is carried out under assumptions (3.2) and (3.26) we obtain

$$\hat{f}(\hat{x}, \hat{\lambda}, \hat{\mu}) = e\hat{x}^5 + t\hat{\mu}\hat{x}^3 + b_\lambda \hat{x} = 0, \quad (3.27)$$

where e and t are known quantities. (We note that to simplify the analysis it is convenient first to make a transformation of (λ, μ) to $(\bar{\lambda}, \bar{\mu})$ where

$$\psi_0 F_{\bar{\mu}x}^0 \Phi_0 + \psi_0 F_{xx}^0 \Phi_0 v_{\bar{\mu}} = 0.) \quad (3.28)$$

Hence $f(x, \lambda, \mu) = 0$ is a universal Z_2 unfolding of $f(x_0, \lambda_0, \mu_0) = 0$ if $t \cdot e \neq 0$.

Now the system we use to compute (x_0, λ_0, μ_0) is

$$H(s) := \begin{pmatrix} f \\ f_x \phi \\ l \phi - 1 \\ f_x z + f_{xx} \phi \phi \\ f_x q + f_{xx} \phi z + \frac{1}{2} f_{xxx} \phi \phi \phi \\ l q \end{pmatrix} = 0, \quad s = (x, \phi, \lambda, z, q, \mu) \in Z, \quad Z = X_s \times X_a \times R \times X_s \times X_a \times R. \quad (3.29)$$

One can prove the following theorem.

THEOREM 3.3 Let (x_0, λ_0, μ_0) satisfy (3.9) and (3.26). Then $H(s) = 0$, given by (3.29), has an isolated solution if and only if $t \neq 0$, where t is the multiplier for $\hat{\mu}, \hat{x}$ in (3.27). If, in addition in (3.27), $e \neq 0$ then (x_0, λ_0, μ_0) is a quartic symmetry breaking bifurcation point.

Again, the implementation of Newton's method for (3.29) is an extension of the approach in [10]. Numerical results illustrating the use of system (3.29) are given in section 5.

4. NUMERICAL IMPLEMENTATION

In this section we discuss how the algorithms developed in section 3 may be applied to the finite-element discretization of the Navier-Stokes equations described in section 2. In particular we shall explain how the symmetry is used to reduce the number of degrees of freedom in the problem.

Let $\hat{\mathbf{X}} = W_{h,0} \times M_h$ and denote the pair (\mathbf{U}_h, p_h) by \mathcal{U} . We denote the Fréchet derivative (with respect to \mathcal{U}) of \mathbf{A}_h evaluated at (\mathcal{U}, R, Γ) by $\mathbf{D}_{\mathcal{U}} \mathbf{A}_h(\mathcal{U}, R, \Gamma)$. Similarly we denote the second Fréchet derivative by $\mathbf{D}_{\mathcal{U}\mathcal{U}}^2 \mathbf{A}_h(\mathcal{U}, R, \Gamma)$ etc. Exact forms for these derivatives may readily be found and we have

$$\mathbf{D}_{\mathcal{U}} \mathbf{A}_h(\mathcal{U}, R, \Gamma) \mathcal{U}^{(1)} = \mathcal{W}(\equiv (\mathbf{W}_h, r_h)), \quad (4.1)$$

where

$$\begin{aligned} & a_1(\hat{\mathbf{U}}_h; \mathbf{U}_h^{(1)}, \mathbf{V}_h) + a_1(\mathbf{U}_h; \mathbf{U}_h^{(1)}, \mathbf{V}_h) + a_1(\mathbf{U}_h^{(1)}; \mathbf{U}_h, \mathbf{V}_h) + \\ & + a_0(\mathbf{U}_h^{(1)}, \mathbf{V}_h) + b(p_h^{(1)}, \mathbf{V}_h) = \langle \mathbf{W}_h, \mathbf{V}_h \rangle \quad \text{for all } \mathbf{V}_h \in W_{h,0}, \\ & b(q_h, \mathbf{U}_h^{(1)}) = \langle r_h, q_h \rangle \quad \text{for all } q_h \in M_h, \end{aligned}$$

$$\mathbf{D}_{\mathcal{U}\mathcal{U}}^2 \mathbf{A}_h(\mathcal{U}, R, \Gamma) \mathcal{U}^{(1)} \mathcal{U}^{(2)} = \mathcal{W}(\equiv (\mathbf{W}_h, 0)), \quad (4.2)$$

where

$$a_1(\mathbf{U}_h^{(1)}; \mathbf{U}_h^{(2)}, \mathbf{V}_h) + a_1(\mathbf{U}_h^{(2)}; \mathbf{U}_h^{(1)}, \mathbf{V}_h) = \langle \mathbf{W}_h, \mathbf{V}_h \rangle \quad \text{for all } \mathbf{V}_h \in W_{h,0}.$$

Note also that $\mathbf{D}_{\mathcal{U}\mathcal{U}\mathcal{U}}^3 \mathbf{A}_h \equiv \mathbf{0}$.

The algorithms discussed in section 3 all involve \mathbf{A}_h and its derivatives evaluated for vectors which have a definite symmetry (that is, they are eigenvectors of S). The results (2.22)–(2.24) show that these quantities may be evaluated as integrals over D^+ and that they only involve functions defined on D^+ . This means that we can halve the number of degrees of freedom in the problem. We now explain how this is done.

The finite-element space $W_{h,0}$ is generated by nine-node isoparametric quadrilateral elements with biquadratic interpolation. The space M_h is generated by piecewise linear interpolation on the same elements, the interpolation being, in general, discontinuous across element boundaries [3]. A natural basis is associated with this finite-element approximation in which each coefficient appearing in any linear combination of the basis functions is either the value of the velocity or pressure at a node in the grid. We denote this basis by

$$\{Q_i | i \in N\}, \quad (4.3)$$

where N is an index set containing the labels of the degrees of freedom in the problem. Any $\mathcal{U} \in \hat{\mathbf{X}}$ has a unique representation of the form

$$\mathcal{U} = \sum_{i \in N_s} x_i Q_i. \quad (4.4)$$

This representation induces a mapping $\mathcal{U} \rightarrow \mathbf{x} (\equiv \{x_i\})$ which is an isomorphism from $\hat{\mathbf{X}}$ onto $\mathbf{X} (\equiv \mathbb{R}^{n(N)})$ where $n(N)$ is the number of elements in N .

For the purposes of implementing the algorithms of section 3 we can define bases for $\hat{\mathbf{X}}$, and $\hat{\mathbf{X}}_a$ which only involve degrees of freedom associated with nodes in D^+ . If N_s contains the labels of all the degrees of freedom associated with nodes in D^+ except those corresponding to the z -component of velocity on the line $z = 0$, then any $\mathcal{U} \in \hat{\mathbf{X}}_s$ may be evaluated on D^+ by the expression

$$\mathcal{U} = \sum_{i \in N_s} x_i Q_i. \quad (4.5)$$

Similarly for any $\mathcal{U} \in \mathbf{X}_a$

$$\mathcal{U} = \sum_{i \in N_a} x_i Q_i, \quad (4.6)$$

where N_a contains all those degrees of freedom associated with nodes in D^+ except those corresponding to the r - and ϕ -components of velocity on the line $z = 0$. It is easy to see that $n(N_s), n(N_a) \approx n(N)/2$.

We can now show how this finite-element method fits into the framework of section 3. We define $\mathbf{f} : \mathbf{X} \times \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{X}$ (cf. (3.1)) by

$$\{\mathbf{f}(\mathbf{x}, R, \Gamma)\}_i = \langle \mathbf{A}_h(\mathcal{U}, R, \Gamma), Q_i \rangle, \quad i \in N_s, \quad (4.7)$$

where $\mathcal{U} = \sum_{i \in N_s} x_i Q_i$. We recall that the algorithm in [10] involves the *symmetric* and *antisymmetric*

Jacobian matrices which may be calculated as

$$\{\mathbf{f}_x^s\}_{i,j} = \langle \mathbf{D}_{\mathcal{U}} \mathbf{A}_h Q_j, Q_i \rangle, \quad i, j \in N_s,$$

$$\{\mathbf{f}_x^a\}_{i,j} = \langle \mathbf{D}_{\mathcal{U}} \mathbf{A}_h Q_j, Q_i \rangle, \quad i, j \in N_a.$$

The algorithm also requires quantities like $\mathbf{f}_{xx} \mathbf{x}^{(1)} \mathbf{x}^{(2)}$ which are obtained as

$$\{\mathbf{f}_{xx} \mathbf{x}^{(1)} \mathbf{x}^{(2)}\}_i = \langle \mathbf{D}_{\mathcal{U}\mathcal{U}}^2 \mathbf{A}_h \mathcal{U}^{(1)} \mathcal{U}^{(2)}, Q_i \rangle, \quad i \in N_s \text{, or } i \in N_a.$$

It is important to note that the integrals involved in evaluating $\mathbf{f}_{xx} \mathbf{x}^{(1)} \mathbf{x}^{(2)}$ are no more complicated than those involved in evaluating \mathbf{f} . This is particularly important in the algorithms for the high order (codimension one) singularities which involve several such terms.

All the integrals are evaluated as a sum of integrals over elements where a 9-point Gaussian quadrature is used. The linear systems of equations arising in Newton's method are solved by performing an *LU* decomposition, followed by the solution of two triangular systems, using the frontal method [4]. We recall that the algorithm in [10] involves making rank one modifications to the antisymmetric Jacobian matrix and note that the frontal method can handle these at very little extra cost.

The solution of the extended systems (3.25) and (3.29) by Newton's method is straightforward being an obvious extension of the techniques used in [10].

We conclude this section by indicating the amount of computational work involved in one iteration of the algorithm in [10] and the algorithms for the codimension one singularities.

Symmetry breaking bifurcation (codimension zero)

- 1 LU decomposition of \mathbf{f}_x^s
- 1 LU decomposition of \mathbf{f}_x^u
- 3 forward substitutions
- 3 back substitutions
- 4 right hand side evaluations.

Codimension one singularity

- 1 LU decomposition of \mathbf{f}_x^s
- 1 LU decomposition of \mathbf{f}_x^u
- 9 forward substitutions
- 9 back substitutions
- 12 right hand side evaluations.

One of the characteristics of the finite-element method is that the cost of calculating the right hand sides and the matrices is high relative to finite-difference methods and as a consequence the cost for the codimension one singularities is dominated by these evaluations rather than the LU decomposition, at least for moderately sized problems (having a few thousand degrees of freedom).

5. RESULTS

In this section we present some results obtained for the Taylor problem, described in section 2, using the methods described in this paper.

Iteration	R	Γ	$ \delta R $	$ \delta\Gamma $	$\ \delta\mathbf{x}\ _\infty$
0	151.27000000	1.2000000000			
1	158.17933743	1.2470657027	0.69E+01	0.48E-01	0.32E+01
2	159.27828024	1.2356022236	0.11E+01	0.12E-01	0.40E+00
3	159.33525031	1.2357042905	0.57E-01	0.10E-03	0.24E-01
4	159.33531224	1.2357042260	0.62E-04	0.65E-07	0.25E-04
			0.38E-10	0.38E-13	0.18E-10

Table 5.1

Part of the path of symmetry breaking bifurcations shown in figure (2.1) was calculated using Euler–Newton continuation applied to the system (3.11). In the neighbourhood of point C the aspect ratio was used as the bifurcation parameter so that \mathbf{F}_y was not singular. We note here that, if one is prepared to interchange the rôles of the parameters in this fashion it is not necessary to use a psuedo-arc length technique when following the path of symmetry breaking bifurcation points. The parameter values obtained using this method were the same as in [3] where a different technique (not exploiting the symmetry) was used.

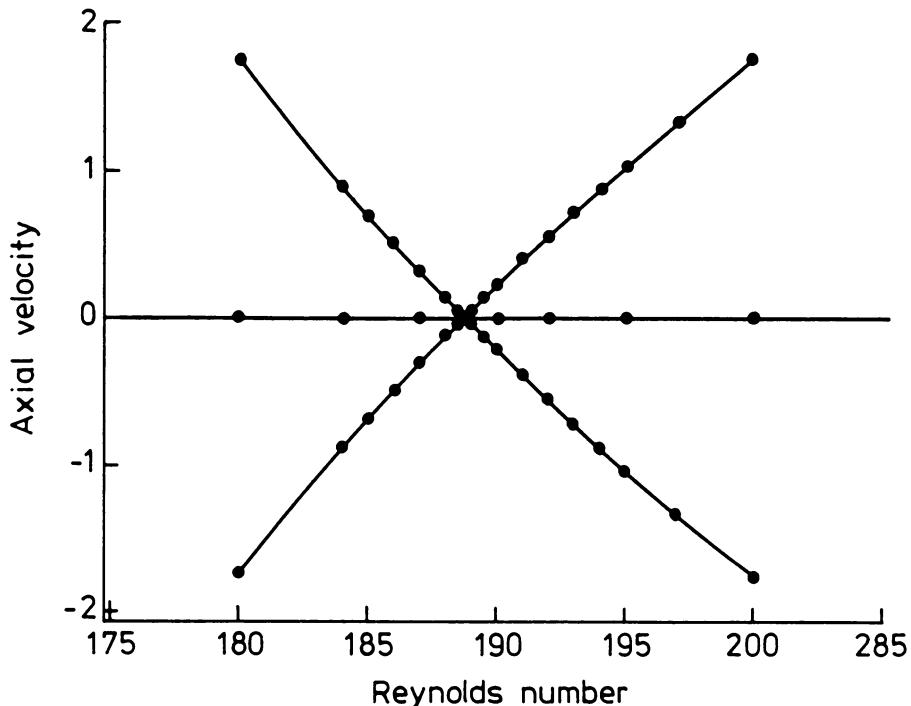


Figure (5.1). Numerically calculated bifurcation diagram at the coalescence point (C). The z -component of velocity at point $(0.75, 0)$ is plotted against Reynolds number; the solid dots indicate the calculated points.

Points on the curve near B and C were used to provide initial guesses for the quartic symmetry-breaking bifurcation point and the coalescence point respectively. Table 5.1 contains details of the Newton iterates for the quartic bifurcation point (the coalescence point gives very similar results). The quadratic convergence is evident. The coordinates of the coalescence point (point C) in the $(R-\Gamma)$ plane are $(188.73, 1.2644)$. A numerically calculated bifurcation diagram at this point is shown in figure (5.1).

The above calculations were done on a grid having approximately 1000 degrees of

freedom. The finite-element program was run on a CRAY-1S computer with an I/O processor. The CPU time taken for one iteration of a codimension one singularity algorithm as described in section 3 was 2.5 seconds compared with 1.6 seconds for one iteration of the symmetry breaking bifurcation algorithm, [10].

REFERENCES

1. Benjamin, T.B. & Mullin, T. 1981 Anomalous modes in the Taylor experiment. *Proc. Roy. Soc. Lond. A* **359**, 24–43.
2. Brezzi, F., Rappaz, J. & Raviart, P.A. 1981 Finite dimensional approximations of nonlinear problems. Part III. Simple bifurcation points. *Numer. Math.* **38**, 1–30.
3. Cliffe, K.A. 1983 Numerical calculations of two-cell and single-cell Taylor flows. *J. Fluid Mech.* **135**, 219–233.
4. Duff, I.S. 1981 MA32 – A package for solving sparse unsymmetric systems using the frontal method. *Harwell Report AERE R-10079*. HMSO.
5. Golubitsky, M. & Langford, W.F. 1981 Classification and unfoldings of degenerate Hopf bifurcations. *J. Diff. Eqn.* **41**, 375–415.
6. Golubitsky, M. & Schaeffer, D. 1979 Imperfect bifurcation in the presence of symmetry. *Commun. Math. Phys.* **67**, 205–232.
7. Jepson, A.D. & Spence, A. 1982 Folds in solutions of two parameter systems and their calculation: Part I. *Stanford University Technical Report (submitted to SIAM J. Numer. Anal.)*
8. Spence, A. & Jepson, A.D. The calculation of cusps, bifurcation points and isolas in two parameter problems, these proceedings.
9. Spence, A. & Werner, B. Non-simple turning points and cusps. *I.M.A. J. Numer. Anal.* **2**, 413–427.
10. Werner, B. & Spence, A. The computation of symmetry-breaking bifurcation points. *SIAM J. Numer. Anal.* (To appear).
11. Werner, B. Regular systems for bifurcation points with underlying symmetries, these proceedings.

K. Andrew Cliffe
 Theoretical Physics Division
 AERE Harwell
 Oxfordshire, OX11 0RA
 U.K.

Alastair Spence
 School of Mathematics
 University of Bath
 Claverton Down
 Bath, BA2 7AY
 U.K.

ON HOPF AND SUBHARMONIC BIFURCATIONS

Jean DESCLOUX

1. INTRODUCTION

Let $(\lambda, x) \in \mathbb{R} \times \mathbb{R}^n \rightarrow f(\lambda, x) \in \mathbb{R}^n$ be a given function that, for simplicity, we shall assume to be of classe C^∞ . We also assume that the partial derivative $D_x f(\lambda, x)$ is bounded, uniformly with respect to $(\lambda, x) \in \mathbb{R} \times \mathbb{R}^n$.

We are interested by periodic solutions of the differential equation $\dot{u}(t) + f(\lambda, u(t)) = 0$. By the standard modification of the independent variable t , we are lead to the problem of finding scalar parameters ρ, λ and 2π -periodic functions $u(t)$ such that

$$\dot{u}(t) + \rho f(\lambda, u(t)) = 0, \quad u \text{ 2}\pi\text{-periodic}; \quad (1.1)$$

in this paper the "dot" will always denote the derivative with respect to time.

Let $(t; \rho, \lambda, a) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow Z(t; \rho, \lambda, a)$ be the "solution operator" of the differential equation (1.1) with initial conditions, i.e. Z is defined by the relations

$$\dot{Z}(t; \rho, \lambda, a) + \rho f(\lambda, Z(t; \rho, \lambda, a)) = 0, \quad Z(0; \rho, \lambda, a) = a. \quad (1.2)$$

Since f is a smooth map, Z will be a C^∞ function of all its arguments. Let $F: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be defined by

$$F(\rho, \lambda, a) = Z(2\pi; \rho, \lambda, a) - a; \quad (1.3)$$

by setting $a = u(0)$, we see that Problem (1.1) is equivalent to the equation $F(\rho, \lambda, a) = 0$.

Let $(\rho_0, \lambda_0, u_0(t))$ be a solution of (1.1); we shall be interested in studying (and computing) solutions of (1.1) in a "neighborhood" of $(\rho_0, \lambda_0, u_0(t))$. Since the differential equation is autonomous, for any value of the "phase τ ", $(\rho_0, \lambda_0, u_0(t-\tau))$ is also a solution; however this multiplicity of solutions is

clearly uninteresting and we want to eliminate it; to achieve this goal we "anchor" the phase by adding an "anchorage" equation. Let a_0 be equal, exactly or approximately, to $u_0(0)$; we consider the problem of finding $(\rho, \lambda, a) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$ such that

$$F(\rho, \lambda, a) = 0, \quad (1.4)$$

$$F: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}, \quad F(\rho, \lambda, a) = (F(\rho, \lambda, a), \quad b_0^T(a - a_0)); \quad (1.5)$$

here b_0^T , transposed of b_0 , is a row vector that we shall choose from case to case. Note that, at least for situations we shall consider in this paper, small changes of b_0 and a_0 in (1.5) will only affect the phase.

As far as numerical computations are concerned, the formulation (1.4) of the "exact" problem clearly suggests the use of shooting methods. Let $Z_h(t_k; \rho, \lambda, a)$ be a numerical approximation of $Z(t_k; \rho, \lambda, a)$ obtained by any standard method for $0 = t_0 < t_1 < t_2 \dots < t_M = 2\pi$, where $h = \max(t_i - t_{i-1})$. We suppose that Z_h is a smooth function of ρ, λ, a . Setting

$$F_h(\rho, \lambda, a) = Z_h(2\pi; \rho, \lambda, a) - a, \quad (1.6)$$

we consider the problem of finding $(\rho, \lambda, a) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$ such that

$$F_h(\rho, \lambda, a) = 0, \quad (1.7)$$

$$F_h: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}, \quad F_h(\rho, \lambda, a) = (F_h(2\pi; \rho, \lambda, a), \quad b_0^T(a - a_0)). \quad (1.8)$$

The purpose of this paper is to apply the general results of [1], [2] to the problems (1.4) and (1.7) for showing existence of solution branches of the exact and the approximate problems, for establishing convergence results and error estimates.

In Section 2, we shall recall the results of [1], [2] which are relevant to our purpose. Section 3 will be devoted to the Hopf bifurcation situation. In Section 4 we shall analyse three cases of subharmonic bifurcation (for the theory of subharmonic bifurcation, see Iooss and Joseph [3]). The proofs of the results contained in Sections 3 and 4 are rather elementary and repetitive; for this reason we shall only sketch them in Section 5.

In this paper we shall work with the following approximation assumption: For $k = 0, 1, 2, \dots$ and for any $(\rho_0, \lambda_0, a_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^N$ there is a neighborhood N of this point such that

$$\| F^{(k)}(\rho, \lambda, a) - F_h^{(k)}(\rho, \lambda, a) \| = O(h^q) \text{ uniformly in } N, \quad (1.9)$$

where q is a positive integer independent of $k, \rho_0, \lambda_0, a_0$. $F^{(k)}$ denotes the total k -th Frechet derivative of F and $\| \cdot \|$ represents the norm in $\mathcal{L}_p(\mathbb{R}^{n+2}, \mathbb{R}^n)$, the set of p -linear maps from \mathbb{R}^{n+2} into \mathbb{R}^n .

In fact Hypothesis (1.9) is satisfied for any Runge-Kutta method of order q or for any stable multistep method of order q with a convenient starting procedure (Runge-Kutta for example); an idea of the proof can be found in Section 5.

Remark 1.1. Since it does not involve the introduction of functional spaces, (1.4) is the simplest formulation of Problem (1.1). In the same way Formulation (1.7) of the approximate problem reduces to a minimum the size of the nonlinear system of equations to solve; however, as mention in [4], the drawback of this simplicity may be, in some cases, the presence of numerical instabilities due likely to the fact that the anchorage acts only on the initial condition a .

Remark 1.2. The analysis contained in this paper can be transposed without too much difficulty to formulations which consider (1.1) as a two-point boundary value problem; the anchorage can be chosen of the form $\int_0^{2\pi} \psi^T(t)(u(t) - u_0(t))dt = 0$ for a given $\psi: [0, 2\pi] \rightarrow \mathbb{R}^n$. For an example of discretization, see [4].

Remark 1.3. The analysis of approximations of Hopf bifurcation for partial differential equations is a much more complicated task as shown in [5]. See also [6] where we apply the results of [1].

2. SOME BASIC ESTIMATES

In this section we quote some results of [1], [2] which are directly relevant to the purposes of this paper.

Let $G: \mathbb{R}^{N+1} \rightarrow \mathbb{R}^N$ a C^∞ -mapping and $x_0 \in \mathbb{R}^{N+1}$ such that $G(x_0) = 0$. We are interested by branches of solutions of the equation $G(x) = 0$ in the neighborhood of x_0 . G is approximated by C^∞ -mappings $G_h: \mathbb{R}^{N+1} \rightarrow \mathbb{R}^N$. We suppose there exists a positive integer q such that for any $h = 0, 1, 2, \dots$ and any $\xi \in \mathbb{R}^{N+1}$ there is a neighborhood $N(\xi)$ of ξ for which

$$\|G^{(k)}(x) - G_h^{(k)}(x)\| = O(h^q) \text{ uniformly in } N(\xi). \quad (2.1)$$

Let $(p+1)$, $p > 0$, be the dimension of the kernel K of $F'(x_0): \mathbb{R}^{N+1} \rightarrow \mathbb{R}^N$. Then there exist p linearly independent row vectors $\psi_1^T, \psi_2^T, \dots, \psi_p^T$ such that the range R of $F'(x_0)$ is given by $R = \{y \in \mathbb{R}^N \mid \psi_k^T y = 0, k = 1, 2, \dots, p\}$.

An element $\sigma_0 \in K$, $\sigma_0 \neq 0$, will be called a characteristic ray if

$$\psi_k^T G''(x_0)[\sigma_0, \sigma_0] = 0, \quad k = 1, 2, \dots, p; \quad (2.2)$$

the characteristic ray σ_0 will be said "non-degenerate" if

$$\left. \begin{array}{l} \text{the relations } \sigma \in K, \psi_k^T G''(x_0)[\sigma_0, \sigma] = 0, k = 1, \dots, p \text{ implies } \\ \text{the existence of } \alpha \in \mathbb{R} \text{ such that } \sigma = \alpha \sigma_0. \end{array} \right\} \quad (2.3)$$

Theorem 2.1. Suppose that Hypothesis (2.1) is satisfied and that $\sigma_0 \in K$ is a non-degenerate characteristic ray, i.e. (2.2) and (2.3) hold. Then there exist $\xi_0 > 0$, $h_0 > 0$, $\xi_h = O(h^{q/2}) > 0$ such that

a) there exists a C^∞ -mapping $\xi \in (-\xi_0, \xi_0) \rightarrow x(\xi)$ such that

$$G(x(\xi)) = 0, \quad |\xi| < \xi_0; \quad x(0) = x_0; \quad x'(0) = \sigma_0; \quad (2.4)$$

b) for $h < h_0$ there exists a C^∞ -mapping $\xi \in (-\xi_0, -\xi_h) \cup (\xi_h, \xi_0) \rightarrow x_h(\xi)$ such that

$$G_h(x_h(\xi)) = 0, \quad \xi_h < |\xi| < \xi_0; \quad \sup_{\xi_h < |\xi| < \xi_0} \|x(\xi) - x_h(\xi)\| = O(h^{q/2}); \quad (2.5)$$

c) if furthermore for $h < h_0$ there exists $n_h \in \mathbb{R}^{N+1}$ such that $\lim_{h \rightarrow 0} n_h = 0$, $G(x_0 + n_h) = 0$ and dimension $\text{Ker}(G'(x_0 + n_h)) = p+1$, then there exists a C^∞ -mapping $\xi \in (-\xi_0, \xi_0) \rightarrow x_h(\xi)$, $h < h_0$, such that

$$G_h(x_h(\xi)) = o, \quad |\xi| < \xi_0; \quad \sup_{|\xi| < \xi_0} \|x(\xi) - x_h(\xi)\| = O(h^q) + \|n_h\|. \quad (2.6)$$

If $p = 0$, we shall say that x_0 is a *regular point* and we have:

Theorem 2.2. Suppose that (2.1) is satisfied and that x_0 is a regular point. Let $\sigma_0 \in K$, $\sigma_0 \neq o$. Then there exist $\xi_0 > 0$ and two C^∞ -mappings $x(\xi), x_h(\xi)$, $|\xi| < \xi_0$ such that $x(o) = x_0$, $x'(o) = \sigma_0$

$$G(x(\xi)) = G_h(x_h(\xi)) = o \text{ for } |\xi| < \xi_0, \quad \sup_{|\xi| < \xi_0} \|x(\xi) - x_h(\xi)\| = O(h^q). \quad (2.7)$$

If $p = 1$ and if there exists a non-degenerate characteristic ray σ_0 , we shall say that x_0 is a *simple bifurcation point* and we have

Theorem 2.3. Suppose that (2.1) is satisfied and that x_0 is a simple bifurcation point. Then:

- a) there exist exactly two linearly independent characteristic rays σ_0 and σ_1 (defined up to a non-vanishing multiplicative constant); both are non-degenerate so that we can apply Theorem 2.1 a,b for each of them;
- b) let $x(\xi)$ be a C^∞ -map, $|\xi| < \xi_0 > 0$, such that $G(x(\xi)) = o$ for $|\xi| < \xi_0$, $x(o) = o$, $x'(o) = \sigma_0$; suppose that there exist a C^∞ -map $x_h(\xi)$, $|\xi| < \xi_0$ such that $G_h(x_h(\xi)) = o$ and $\lim_{h \rightarrow 0} \sup_{|\xi| < \xi_0} \|x(\xi) - x_h(\xi)\| = o$; then the hypothesis of Theorem 2.1c is satisfied with $\|n_h\| = O(h^q)$.

Remark 2.1. When x_0 is a regular point or a simple bifurcation point, it is possible to give results concerning uniqueness. For the general case see [1], Remarks 4.3, 4.4.

Remark 2.2. The hypothesis of Theorem 2.1c means essentially that G_h possesses near x_0 a singularity similar to the one of G in x_0 ; if this assumption is not satisfied we must expect only an "imperfect bifurcation" for the approximate problem.

Remark 2.3. If x_0 is a simple bifurcation point, by using Morse's lemma, it is

possible to improve the information given in Theorem 2.1b; see [7], [8], [9].

Remark 2.4. The framework defined in the beginning of this section may also be interesting from a computational point of view. Suppose that σ_0 is a non-degenerate ray in the sense of (2.2), (2.3), normalized by the condition $\sigma_0^T \sigma_0 = 1$ and consider the following Newton-Chord algorithm:

$$y_0 = x_0 + \alpha \sigma_0, \quad y_{k+1} = y_k - (H_\alpha'(y_k))^{-1} H_\alpha(y_k), \quad k = 0, 1, 2, \dots,$$

where $H_\alpha : \mathbb{R}^{N+1} \rightarrow \mathbb{R}^N \times \mathbb{R}$ is defined by $H_\alpha(x) = (G(x), \sigma_0^T(x-x_0) - \alpha)$. In [10], we prove the existence of $\alpha_0 > 0$ such that for $0 < |\alpha| < \alpha_0$ the sequence y_k converges to a solution of the equation $H_\alpha(x) = 0$.

3. HOPF BIFURCATION

We consider the situation described in Section 1. In particular Z, F and F_h are defined by (1.2), (1.3) and (1.6). We assume that Hypothesis (1.9) is satisfied.

Furthermore we suppose in this section that the classical Hopf assumptions are fulfilled:

a) $f(\lambda, 0) = 0 \quad \forall \lambda \in \mathbb{R}; \quad (3.1)$

b) there exists $\lambda_0 \in \mathbb{R}$, $\rho_0 \in \mathbb{R}$, $\rho_0 > 0$ and a vector $\varphi_0 \in \mathbb{C}^n$, $\varphi_0 \neq 0$ such that $\rho_0 D_x f(\lambda_0, 0) \varphi_0 = i \varphi_0; \quad (3.2)$

c) i is an algebraically simple eigenvalue of $\rho_0 D_x f(\lambda_0, 0); \quad (3.3)$

d) for $k \in \mathbb{Z} - \{\pm 1\}$, ki is not an eigenvalue of $\rho_0 D_x f(\lambda_0, 0); \quad (3.4)$

e) $\operatorname{Re} \mu'(\lambda_0) \neq 0$, where $\mu(\lambda)$ is the eigenvalue of $\rho_0 D_x f(\lambda, 0)$ with $\mu(\lambda_0) = i. \quad (3.5)$

Clearly by (3.1), $Z(t, \rho, \lambda, 0) = 0$. Without much restriction on the "numerical solver" Z_h , we moreover shall assume that

$$Z_h(2\pi, \rho, \lambda, 0) = 0 \quad \text{for all } \rho, \lambda \in \mathbb{R}. \quad (3.6)$$

Let $c_0 = 2 \operatorname{Re} \varphi_0 \in \mathbb{R}^n$, $b_0 \in \mathbb{R}^n$ be such that $b_0 \neq 0$, $b_0 \in \operatorname{span}(\operatorname{Re} \varphi_0, \operatorname{Im} \varphi_0)$,

$b_0^T c_0 = 0$; furthermore let $\psi_0 \in C^n$ be such that $\psi_0^T D_x f(\lambda_0, o) = i\psi_0^T$. We set

$$F: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}, \quad F(\rho, \lambda, a) = (F(\rho, \lambda, a), b_0^T a), \quad (3.7)$$

$$F_h: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}, \quad F_h(\rho, \lambda, a) = (F_h(\rho, \lambda, a), b_0^T a). \quad (3.8)$$

Clearly $F(\rho_0, \lambda_0, o) = 0$. Let $K \subset \mathbb{R}^{n+2}$ be the kernel of $F'(\rho_0, \lambda_0, o)$, the "total" derivative of F at (ρ_0, λ_0, o) and $R \subset \mathbb{R}^{n+1}$ be its range. We have

Theorem 3.1. K has dimension 3, R has codimension 2. They are given by

$$K = \{(\alpha, \beta, \gamma c_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \mid \alpha, \beta, \gamma \in \mathbb{R}\},$$

$$R = \{(a, \lambda) \in \mathbb{R}^n \times \mathbb{R} \mid \psi_0^T a = 0\}.$$

Applying the definitions (2.2) and (2.3) to $G = F$, $x_0 = (\rho_0, \lambda_0, o)$, $N = n+1$, we have

Theorem 3.2. a) $\sigma_0 \in K$ is a characteristic ray if and only if it is of the form $\sigma_0 = (\alpha, \beta, o)$, $\alpha^2 + \beta^2 > 0$ or $\sigma_0 = (o, o, \gamma c_0)$, $\gamma \neq o$. b) The only non-degenerate (defined up to a non-vanishing scalar factor) characteristic ray is (o, o, c_0) .

As far as the approximate problem $F_h(\rho, \lambda, a) = 0$ is concerned we have

Theorem 3.3. There exist $h_0 > 0$, $\rho_{oh} \in \mathbb{R}$, $\lambda_{oh} \in \mathbb{R}$ such that a) $|\rho_o - \rho_{oh}| + |\lambda_o - \lambda_{oh}| = O(h^q)$, b) the dimension of the kernel of $F'_h(\rho_{oh}, \lambda_{oh}, o)$ is equal to 3 for $h < h_0$.

By using Theorem 3.2 and 3.3, we deduce immediately from Theorem 2.1a,c, the main result of this section:

Theorem 3.4. There exist scalars $h_0 > 0$, $\xi_0 > 0$ and C^∞ -mappings $\xi \in (-\xi_0, \xi_0) \rightarrow (\rho(\xi), \lambda(\xi), a(\xi))$, $\xi \in (-\xi_0, \xi_0) \rightarrow (\rho_h(\xi), \lambda_h(\xi), a_h(\xi))$ for $h < h_0$ such that

$$F(\rho(\xi), \lambda(\xi), a(\xi)) = 0, \quad F_h(\rho_h(\xi), \lambda_h(\xi), a_h(\xi)) = 0, \quad |\xi| < \xi_0,$$

$$(\rho(o), \lambda(o), a(o)) = (\rho_0, \lambda_0, o), \quad (\rho'(o), \lambda'(o), a'(o)) = (o, o, c_0),$$

$$(\rho_h(o), \lambda_h(o), a_h(o)) = (\rho_{oh}, \lambda_{oh}, o),$$

$$\sup_{|\xi| < \xi_0} \{ |\rho(\xi) - \rho_h(\xi)| + |\lambda(\xi) - \lambda_h(\xi)| + \|a(\xi) - a_h(\xi)\| \} = O(h^q).$$

Remark 3.1. Theorem 3.4 gives results of existence but leaves aside the non trivial questions of uniqueness.

Remark 3.2. There exists in the literature a great number of proofs for Hopf bifurcation. Let us mention, in a context close to ours, the constructive approach (i.e. leading to algorithms) of Weber [11] which does not however take in consideration discretization errors.

4. SUBHARMONIC BIFURCATION

Let $(\rho_0, \lambda_0, u_0(t))$ be a non-stationary solution of problem (1.1), i.e. $\rho_0 > 0$, $\dot{u}_0(t) \neq 0 \forall t \in \mathbb{R}$. Looking for other solutions in the neighborhood of this one, we are led, following Section 1, to consider Problem (1.4) and (1.7). A natural choice for the anchorage equation $b_0^T(a-a_0) = 0$ would be $b_0 = \dot{u}_0(0)$ and $a_0 = u_0(0)$; however from a computational point of view the difficulty arises from the fact that the approximate problem $F_h(\rho, \lambda, a) = 0$ contains the data a_0 and b_0 which are in fact not known explicitly; that means that a_0 and b_0 should be only approximations of $u_0(0)$ and $\dot{u}_0(0)$. For the sake of simplicity in this section we shall forget this difficulty and simply note the following fact: since $\dot{u}_0(0) \neq 0$, there exist $\varepsilon > 0$ and $\delta > 0$ such that, if $\|b_0 - \dot{u}_0(0)\| + \|a_0 - u_0(0)\| < \varepsilon$, then there exists an unique $t_0 \in (-\delta, \delta)$ such that $b_0^T(u(t_0) - a_0) = 0$, which implies that $F(\rho_0, \lambda_0, u(t_0)) = 0$.

From now on we consider the situation described in Section 1. In particular Z , F and F_h are defined by (1.2), (1.3) and (1.6) and we suppose that Hypothesis (1.9) is satisfied. On the basis of the above analysis we introduce $(\rho_0, \lambda_0, a_0) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$ for which we suppose

$$F(\rho_0, \lambda_0, a_0) = 0, \quad \rho_0 > 0, \quad f(\lambda_0, a_0) \neq 0. \quad (4.1)$$

Let $b_0 := -\rho_0 f(\lambda_0, a_0) = Z(0; \rho_0, \lambda_0, a_0)$, and set

$$F: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}, \quad F(\rho, \lambda, a) = (F(\rho, \lambda, a), b_0^T(a - a_0)), \quad (4.2)$$

Let $A_0 = D_a Z(2\pi; \rho_0, \lambda_0, a_0)$ (partial derivative of Z with respect to a). A_0 is a square matrix of order n ; the eigenvalues of A_0 are called the "*Floquet multipliers*" associated to the solution $Z(t; \rho_0, \lambda_0, a_0)$ of Problem (1.1) for $\rho = \rho_0$, $\lambda = \lambda_0$. Since the differential equation in (1.1) is autonomous we have

Theorem 4.1. 1 is eigenvalue of A_0 with eigenvector b_0 .

Following the definitions introduced in Section 2, (ρ_0, λ_0, a_0) will be a regular point of the equation $F(\rho, \lambda, a) = 0$ if the kernel of $F'(\rho_0, \lambda_0, a_0)$ has dimension 1. We have the following characterization:

Theorem 4.2. (ρ_0, λ_0, a_0) is a regular point if and only if one of the two situations holds: a) 1 is an algebraically simple eigenvalue of A_0 , b) 1 is an algebraically multiple but geometrically simple eigenvalue of A_0 and $D_\lambda Z(2\pi; \rho_0, \lambda_0, a_0)$ does not belong to the range of A_0 . In case a) the kernel of $F'(\rho_0, \lambda_0, a_0)$ is spanned by a vector of the form $(\alpha, 1, c) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$; in case b) it is spanned by a vector of the form $(1, 0, c) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$.

Without stating an explicit theorem, we note that Theorem 2.2 applies when (ρ_0, λ_0, a_0) is a regular point. Remark furthermore that in case b of Theorem 4.2, (ρ_0, λ_0, a_0) is a "limit point" with respect to the parameter λ ; if this limit point is a "turning point", one can show that the corresponding approximate branch will also have a turning point (see for example [1] or [2]).

In the remaining part of this section we shall "translate" some of the situations described in [3], chapter XI, in terms of the notions introduced in Section 2. First we complete Hypothesis (4.1) by assuming the existence of a C^∞ -map $\lambda \in I \rightarrow (\rho_0(\lambda), \lambda, a_0(\lambda))$, where I is an open interval containing λ_0 , such that

$$F(\rho_0(\lambda), \lambda, a_0(\lambda)) = 0 \text{ for } \lambda \in I, \quad \rho_0(\lambda_0) = \rho_0, \quad a_0(\lambda_0) = a_0, \quad (4.4)$$

i.e. we suppose the existence of a smooth branch of solutions for the equation $F(\rho, \lambda, a) = 0$ which is parametrized by λ and passes through (ρ_0, λ_0, a_0) . We furthermore define

$$F_\ell : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R}, \quad F_\ell(\rho, \lambda, a) = (Z(2\pi\ell; \rho, \lambda, a) - a, \quad b_0^T(a - a_0)), \quad \ell = 1, 2, 3, \dots, \quad (4.5)$$

$$F_{h\ell}: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}, \quad F_{h\ell}(p, \lambda, a) = (Z_h(2\pi\ell; p, \lambda, a) - a, b_o^T(a - a_0)), \quad \ell = 1, 2, 3, \dots, \quad (4.6)$$

$$A_0(\lambda) = D_a Z(2\pi; p_0(\lambda), \lambda, a_0(\lambda)), \quad b_0(\lambda) = \dot{Z}(o; p_0(\lambda), \lambda, a_0(\lambda)), \quad \lambda \in I, \quad (4.7)$$

where $Z_h(2\pi\ell; p, \lambda, a)$ is the approximation of $Z(2\pi\ell; p, \lambda, a)$ defined by the following "semi-group" property

$$Z_h(2\pi\ell; p, \lambda, a) = Z_h(2\pi; p, \lambda, Z_h(2\pi(\ell-1); p, \lambda, a)), \quad \ell = 2, 3, \dots \quad (4.8)$$

Clearly we have $F_1 = F$, $F_{h1} = F_h$, $A_0(\lambda_0) = A_0$, $b_0(\lambda_0) = b_0$. Hypothesis (1.9) and Definition (4.8) imply for any $k = 0, 1, 2, \dots$ and any $\ell = 1, 2, 3, \dots$ the existence of a neighborhood N of (p_0, λ_0, a_0) such that

$$\| F_\ell^{(k)}(p, \lambda, a) - F_{h\ell}^{(k)}(p, \lambda, a) \| = O(h^q) \text{ uniformly in } N. \quad (4.9)$$

By (4.1), $b_0 \neq o$ and without restriction of generality we can assume that $b_0(\lambda) \neq o$ for all $\lambda \in I$; Theorem 4.1 generalizes immediately in

Theorem 4.3. For all $\lambda \in I$, 1 is an eigenvalue of $A_0(\lambda)$ with eigenvector $b_0(\lambda)$.

We now consider a first case of "subharmonic bifurcation". Suppose that 1 is an algebraically simple eigenvalue of $A_0(\lambda)$ for $\lambda \neq \lambda_0$ and an algebraically double eigenvalue of $A_0(\lambda_0)$. By using Theorem 4.3 and elementary arguments of linear algebra, one shows the existence of a smooth eigenvalue $\mu(\lambda)$ and of a smooth corresponding eigenvector $\varphi(\lambda)$ of $A_0(\lambda)$ such that $\mu(\lambda_0) = 1$, $\mu(\lambda) \neq 1$ if $\lambda \neq \lambda_0$; moreover $\varphi(\lambda_0)$ and $b_0(\lambda_0)$ will be linearly dependent if and only if 1 is a geometrically simple eigenvalue of $A_0(\lambda_0)$. The following theorem holds:

Theorem 4.4. Suppose that 1 is an algebraically double eigenvalue of A_0 and that there exists a smooth eigenvalue $\mu(\lambda)$ of $A(\lambda)$ such that $\mu(\lambda_0) = 1$ and $\mu'(\lambda_0) \neq o$. Then (p_0, λ_0, a_0) is a simple bifurcation point of the equation $F(p, \lambda, a) = o$.

If the Hypotheses of Theorem 4.4 are fulfilled we can apply to F Theorem 2.3a and Theorem 2.1a,b. The exact problem will have two branches of solutions emerging from (p_0, λ_0, a_0) one of them being obviously the branch $(p_0(\lambda), \lambda, a_0(\lambda))$

assumed by Hypothesis (4.4); in our opinion there is no reason that the approximate problem bifurcates exactly i.e. there is no reason that Theorem 2.1c or Theorem 2.3b could be applied.

From physical observations and numerical computations, the next situation we shall consider seems particularly important.

Theorem 4.5. Suppose that 1 is an algebraically simple eigenvalue of $A_0(\lambda_0)$ and that there exists a smooth algebraically simple eigenvalue $\mu(\lambda)$ of $A_0(\lambda)$ such that $\mu(\lambda_0) = -1$, $\mu'(\lambda_0) \neq 0$. Then a) (ρ_0, λ_0, a_0) is a regular point for equation $F(\rho, \lambda, a) = 0$; b) (ρ_0, λ_0, a_0) is a simple bifurcation point for the equation $F_2(\rho, \lambda, a) = 0$ with characteristic directions $(\rho'(\lambda_0), 1, a'(\lambda_0))$ and $(0, 0, \varphi_0 + \alpha b_0)$, where $\alpha \in \mathbb{R}$ and φ_0 is an eigenvector of $A_0(\lambda_0)$ for the eigenvalue -1.

Theorem 4.5 deserves some comments. Since (ρ_0, λ_0, a_0) is a regular point for the equation $F(\rho, \lambda, a) = 0$, by Theorem 2.2, the solution branch $(\rho_0(\lambda), \lambda, a_0(\lambda))$ will be approximated by a solution branch $(\rho_{oh}(\lambda), \lambda, a_{oh}(\lambda))$ of the approximate problem $F_h(\rho, \lambda, a) = 0$ (the fact that the approximate branch can be parametrized by λ follows from results of [1] or [2] for example); because of (4.8), we shall have $F_{h2}(\rho_{oh}(\lambda), \lambda, a_{oh}(\lambda)) = 0$ identically in a neighborhood of $\lambda = \lambda_0$ and h small enough. Since $F_2(\rho_0(\lambda), \lambda, a_0(\lambda)) = 0$, we can apply Theorem 2.3a, b and Theorem 2.1a, c; that means that the approximate problem $F_{h2}(\rho, \lambda, a) = 0$ will bifurcate exactly in two branches which will approximate the two branches of the exact problem $F_2(\rho, \lambda, a) = 0$ with errors in $O(h^q)$ (see 4.9). Solving the equation $F_2(\rho, \lambda, a) = 0$ is equivalent to find 4π -periodic solutions of the differential equation (1.1); by (4.4) the branch $(\rho_0(\lambda), \lambda, a_0(\lambda))$ will satisfy the supplementary "symmetry" property of being 2π -periodic; consequently (ρ_0, λ_0, a_0) appears as a point of "symmetry-breaking" bifurcation; however the formulation " $F_2(\rho, \lambda, a) = 0$ " of the problem does not seem very appropriate for putting in evidence this property in a formal way.

At first glance at least, the last situation we shall consider in this section may appear rather academic since there is little "chance" that it can happen.

Theorem 4.6. Suppose that 1 is an algebraically simple eigenvalue of $A_0(\lambda_0)$ and that there exists a smooth complex, algebraically simple, eigenvalue $\mu(\lambda)$ of $A_0(\lambda)$ such that $\mu^3(\lambda_0) = 1$ and $\mu'(\lambda_0) \neq 0$. Let φ_0 and ψ_0 be non-vanishing complex vectors such that $A_0\varphi_0 = \mu(\lambda_0)\varphi_0$, $\psi_0^T A_0 = \mu(\lambda_0)\psi_0^T$. We suppose that $\psi_0^T D_{aa}^2 F(\rho_0, \lambda_0, a_0)[\bar{\varphi}_0, \bar{\varphi}_0] \neq 0$ where $\bar{\varphi}_0$ is the complex-conjugate of φ_0 . Then:
a) (ρ_0, λ_0, a_0) is a regular point of the equation $F(\rho, \lambda, a) = 0$; b) the kernel of $F'_3(\rho_0, \lambda_0, a_0)$ has dimension 3 and contains exactly 4 characteristic rays; they are non degenerate and of the form $\sigma_k = (\rho'(\lambda_0), 1, a'(\lambda_0) + d_k)$, $k = 0, 1, 2, 3$, where $d_0 = 0$ and d_1, d_2, d_3 are linear combinations of b_0 , $\operatorname{Re}\varphi_0$ and $\operatorname{Im}\varphi_0$; moreover the three branches of solutions of the equation $F_3(\rho, \lambda, a) = 0$ passing through (ρ_0, λ_0, a_0) with directions $\sigma_1, \sigma_2, \sigma_3$ correspond to a single branch of 6π -periodic solutions of the differential equation (1.1) with different phases.

For the solution branches of the equation $F_{h3}(\rho, \lambda, a) = 0$ relative to the characteristic rays $\sigma_1, \sigma_2, \sigma_3$, we can only apply Theorem 2.1a,b; as far as the characteristic rays σ_0 is concerned, we note that by (4.8), the equations $F_{3h} = 0$ and $F_h = 0$ have a common branch of solutions for which we can apply

Theorem 2.2.

Remark 4.1. In [3], Iooss and Joseph analyse more generally the case where $A_0(\lambda)$ possesses a simple eigenvalue $\mu(\lambda)$ such that, for some integer ℓ , $\mu^\ell(\lambda_0) = 1$; in this paper we have considered the cases $\ell = 1, 2, 3$; the case $\ell = 4$ cannot be treated in the framework of Section 2 or even by the more general theory developped in [1], [2]. Following the authors of [3], for $\ell \geq 5$, one cannot expect to obtain subharmonic bifurcations.

5. ELEMENTS OF PROOF

As in the preceding sections, we shall assume here that $f(\lambda, x): \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a C^∞ -function and that $D_x f(\lambda, x)$ is uniformly bounded; this garanties the existence of the C^∞ -map $Z(t; \rho, \lambda, a)$ defined by (1.2).

We first justify the basic hypothesis (1.9). We introduce the compact notations $\alpha = (\rho, \lambda, a)$, $g(\alpha, x) = \rho f(\lambda, x)$, $w_0(\alpha) = a$, $W_0(t, \alpha) = Z(t; \rho, \lambda, a)$. For a

fixed argument $\xi \in \mathbb{R}^{n+2}$, we set $W_1(t, \alpha) = D_\alpha W_0(t, \alpha)\xi$, $W_2(t, \alpha) = D_{\alpha\alpha}^2 W_0(t, \alpha)[\xi, \xi]$, $W_3(t, \alpha) = D_{\alpha\alpha\alpha}^3 W_0(t, \alpha)[\xi, \xi, \xi]$, ..., $w_1(\alpha) = w'_0(\alpha)$, $w_2(\alpha) = w''_0(\alpha)[\xi, \xi]$, $w_3(\alpha) = w'''_0(\alpha)[\xi, \xi, \xi]$, ... Let k be a given positive integer; by differentiating k times the differential equation $W_0(t, \alpha) + g(W_0(t, \alpha)) = 0$ and the initial condition $W_0(0, \alpha) = w_0(\alpha)$, we obtain a system of differential equations of the form

$$\left. \begin{aligned} \dot{W}_0(t, \alpha) + g(W_0(t, \alpha)) &= 0, \quad W_0(0, \alpha) = w_0(\alpha), \\ \dot{W}_1(t, \alpha) + g_1(W_0(t, \alpha), W_1(t, \alpha)) &= 0, \quad W_1(0, \alpha) = w_1(\alpha), \\ \dots & \\ \dot{W}_k(t, \alpha) + g_k(W_0(t, \alpha), W_1(t, \alpha), \dots, W_{k-1}(t, \alpha)) &= 0, \quad W_k(0, \alpha) = w_k(\alpha). \end{aligned} \right\} \quad (5.1)$$

Consider now a standard explicit Runge-Kutta method of order q that we apply to (5.1) for obtaining approximation $W_{\ell h}(t, \alpha)$, $\ell = 0, 1, \dots, k$, at nodes $0 = t_0 < t_1 < t_2 < \dots < t_M = 2\pi$ where $h = \max_{i=1, \dots, M} (t_i - t_{i-1})$. By elementary explicit calculations, one obtains

Lemma 5.1. $W_{\ell h}(t_1, \alpha) = D_{\alpha, \dots, \alpha}^\ell W_{0h}(t_1, \alpha)[\xi, \dots, \xi]$ $\ell = 1, 2, \dots, k$.

By applying recursively Lemma 5.1, we obtain $W_{\ell h}(2\pi, \alpha) = D_{\alpha, \dots, \alpha}^\ell W_{0h}(2\pi, \alpha)[\xi, \dots, \xi]$, $\ell = 1, 2, \dots, k$. From the theory of Runge-Kutta methods we get the error estimate $\| D_{\alpha, \dots, \alpha}^k W_0(2\pi, \alpha)[\xi, \dots, \xi] - D_{\alpha, \dots, \alpha}^k W_{0h}(2\pi, \alpha)[\xi, \dots, \xi] \| = O(h^q)$ uniformly with respect to the variables α and ξ in a any compact set; taking in account the fact that Frechet derivatives are symmetric multilinear functions (see [12] for the properties of symmetric multilinear functions), we obtain (1.9).

The following identities will play an important role for proving the results of Sections 3 and 4.

Lemma 5.2. For any t, ρ, λ, a and $\xi_1, \xi_2 \in \mathbb{R}^n$ we have

$$\dot{Z}(t; \rho, \lambda, a) = D_a Z(t; \rho, \lambda, a) \dot{Z}(0; \rho, \lambda, a); \quad (5.2)$$

$$D_\rho Z(t; \rho, \lambda, a) = \frac{t}{\rho} \dot{Z}(t; \rho, \lambda, a), \quad \rho \neq 0; \quad (5.3)$$

$$\begin{aligned} D_{aa}^2 Z(t; \rho, \lambda, a)[\xi_1, \xi_2] &= -D_a Z(t; \rho, \lambda, a) \rho \int_0^t (D_a Z(\tau; \rho, \lambda, a))^{-1} \cdot \\ &\cdot D_{xx}^2 f(\lambda, Z(\tau; \rho, \lambda, a)) [D_a Z(\tau; \rho, \lambda, a) \xi_1, D_a Z(\tau; \rho, \lambda, a) \xi_2] d\tau. \end{aligned} \quad (5.4)$$

For obtaining (5.4), one differentiates twice (1.2) with respect to a for the arguments ξ_1, ξ_2 , multiplies at the left the resulting equation by $(D_a Z(t; \rho, \lambda, a))^{-1}$, integrates by parts between o and t , takes in account the identity: $-\dot{V}(t) + V(t)R(t) = o$, where $V(t) = (D_a Z(t; \rho, \lambda, a))^{-1}$, $R(t) = \rho D_x f(\lambda, Z(t; \rho, \lambda, a))$.

We next consider the situation of Hopf bifurcation described in Section 3; we use the same notations and adopt the same hypotheses. We set $D(\lambda) = \rho_0 D_x f(\lambda, o)$; by (3.2) - (3.4), there exists a smooth vector $\varphi(\lambda)$ such that $D(\lambda)\varphi(\lambda) = \mu(\lambda)\varphi(\lambda)$, $\varphi(\lambda_0) = \varphi_0$. Without restriction of generality, we can suppose that $\psi_o^T \varphi_0 = 1$ and $\psi_o^T \bar{\varphi}_0 = o$, where we recall that $\psi_o^T D(\lambda_0) = i\psi_o^T$ and $\bar{\varphi}_0$ is the complex conjugate of φ_0 . We define

$$Q: \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{C}, Q(\alpha, \beta, \gamma), (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})) = \psi_o^T F''(\rho_0, \lambda_0, o)[(\alpha, \beta, \gamma c_0), (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma} c_0)]. \quad (5.5)$$

By Theorem 3.1, which can be easily verified, $\sigma_0 = (\alpha, \beta, \gamma c_0) \neq o$ is a characteristic ray if and only if $Q((\alpha, \beta, \gamma), (\alpha, \beta, \gamma)) = o$; furthermore a characteristic ray $(\alpha, \beta, \gamma c_0)$ will be non-degenerate if and only if $Q((\alpha, \beta, \gamma), (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})) = o$ implies that $(\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})$ and (α, β, γ) are linearly dependent. Taking in account (3.5), Theorem (3.2) is a direct consequence of the following

Lemma 5.3. $Q((\alpha, \beta, \gamma), (\tilde{\alpha}, \tilde{\beta}, \tilde{\gamma})) = -2\pi \left\{ \frac{i}{\rho_0} (\alpha \tilde{\gamma} + \tilde{\alpha} \gamma) + \mu'(\lambda_0) (\beta \tilde{\gamma} + \tilde{\beta} \gamma) \right\}.$

Proof. Clearly, $D_{\rho\rho}^2 F$, $D_{\rho\lambda}^2 F$ and $D_{\lambda\lambda}^2 F$ vanish at (ρ_0, λ_0, o) . We show that $\psi_o^T D_{aa}^2 F(\rho_0, \lambda_0, o)[c_0, c_0] = o$; indeed $D_{aa}^2 F = D_{aa}^2 Z$ and we can use (5.4) with $t = 2\pi$, $\rho = \rho_0$, $\lambda = \lambda_0$, $a = o$, $\xi_1 = \xi_2 = c_0 = \varphi_0 + \bar{\varphi}_0$; by (3.1) $Z(t; \rho_0, \lambda_0, o) = o$ so that $D_{xx}^2 f(\lambda, Z(t; \rho_0, \lambda_0, o))$ is independent of t ; since $D_a Z(t; \rho_0, \lambda_0, o) = \exp(-tD(\lambda_0))$, the right member of (5.4) multiplied by ψ_o^T can be evaluated explicitly as a linear combination of integrals of the form $\int_0^{2\pi} e^{imt}$, $m \neq o$, which all vanish. By (5.2), (5.3), we have $D_\rho Z(2\pi; \rho_0, \lambda_0, a) = -2\pi D_a Z(2\pi; \rho_0, \lambda_0, a)F(\lambda_0, a)$; we differentiate this relation with respect to a , multiply at the left by ψ_o^T , at the right by $c_0 = \varphi_0 + \bar{\varphi}_0$ and obtain $\psi_o^T D_{\rho a} Z(2\pi; \rho_0, \lambda_0, o) = c_0 = -2\pi i/\rho_0$ (here we use the relations $\psi_o^T \varphi_0 = 1$, $\psi_o^T \bar{\varphi}_0 = 1$). It remains to show that $\psi_o^T D_{\lambda a} Z(2\pi; \rho_0, \lambda_0, o)c_0 = -2\pi \mu'(\lambda_0)$; to this end, if suffices to differentiate the relation $\psi_o^T D_a Z(2\pi; \rho_0, \lambda_0, o)(\varphi(\lambda) + \bar{\varphi}(\lambda)) =$

$\exp(-2\pi i \mu(\lambda)) \psi_o^T \varphi(\lambda) + \exp(-2\pi i \bar{\mu}(\lambda)) \psi_o^T \bar{\varphi}(\lambda)$ with respect to λ .

Proof of Theorem 3.3. By using an "approximate Lyapunov-Schimdt reduction" as in [1] or [2] it is easy show that we can replace equivalently Theorem 3.3b by: the dimension of the kernel of $D_h F(\rho_{oh}, \lambda_{oh}, o)$ is larger equal to 3 for $h < h_0$. Because of Hypothesis (3.6), we have $D_\rho F_h(\rho, \lambda, o) = D_\lambda F_h(\rho, \lambda, o) = o$ for all ρ and λ . Consequently it is sufficient to show the existence of $h_0 > 0$, $\rho_{oh}, \lambda_{oh} \in \mathbb{R}$ such that a) $|\rho_o - \rho_{oh}| + |\lambda_o - \lambda_{oh}| = O(h^q)$, b) the kernel of $D_a F_h(\rho_{oh}, \lambda_{oh}, o)$ has at least dimension 1 for $h < h_0$. To this end we define the maps

$$H : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}, H(\rho, \lambda, c) = (D_a F(\rho, \lambda, o)c, b_o^T c, c_o^T(c - c_o)),$$

$$H_h : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}, H_h(\rho, \lambda, c) = (D_a F_h(\rho, \lambda, o)c, b_o^T c, c_o^T(c - c_o)).$$

By using the same arguments as in the proof of Lemma 5.3 (in particular Hypothesis (3.5)), one can verify that $H'(\rho_o, \lambda_o, c_o) : \mathbb{R}^{n+2} \rightarrow \mathbb{R}^{n+2}$ is injective and consequently is an isomorphism; moreover by Theorem 3.1, $H(\rho_o, \lambda_o, c_o) = o$; taking in account Hypothesis (1.9), we apply the discrete implicit function theorem of [1] which implies, for h small enough, the existence of $(\rho_{oh}, \lambda_{oh}, c_{oh}) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}^n$ such that $H_h(\rho_{oh}, \lambda_{oh}, c_{oh}) = o$ with $|\rho_o - \rho_{oh}| + |\lambda_o - \lambda_{oh}| + \|c_o - c_{oh}\| = O(h^q)$.

The remaining part of this section will be devoted to subharmonic bifurcations; we use the notations of Sectio 4 and suppose that its hypotheses are fulfilled. We first note that Theorem 4.1, 4.2 and 4.3 are easy consequences of the identities (5.2), (5.3).

Proof of Theorem 4.4. By (5.3), $D_\rho F(\rho_o, \lambda_o, a_o)$ does not vanish and is colinear with b_o which belongs (Theorem 4.3) to the kernel of $D_a F(\rho_o, \lambda_o, a_o) = A_o - I$; since o is an algebraically double eigenvalue of $(A_o - I)$, it follows that the matrix $F'(\rho_o, \lambda_o, a_o)$ has rank $(n-1)$; we easily conclude that the kernel of $F'(\rho_o, \lambda_o, a_o)$ has dimension 2. Hypothesis (4.4) implies that $\sigma_o = (\sigma'_o(\lambda_o), 1, a'_o(\lambda_o))$ is a characteristic direction; by definition of the simple bifurcation point, it suffices to prove that σ_o is non degenerate. Suppose first that 1 is a geometrically double eigenvalue of A_o ; let ψ_o and $\varphi(\lambda)$ be such that $\psi_o^T A_o = \psi_o^T$, $A(\lambda) \varphi(\lambda) = \mu(\lambda) \varphi(\lambda)$, $\psi_o^T \varphi(\lambda_o) = 1$, $\psi_o^T b_o = o$ (see the comment preced-

ing Theorem 4.4). Then, for some $\varepsilon \in \mathbb{R}$,

$$\begin{aligned} \text{Kernel } F'(\rho_0, \lambda_0, a_0) &= \text{span}\{(\rho'_0(\lambda_0), 1, a'_0(\lambda_0)), (0, 0, \varphi(\lambda_0) + \varepsilon b_0)\}, \\ \text{Range } F'(\rho_0, \lambda_0, a_0) &= \{(a, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid \psi_0^T a = 0\}. \end{aligned}$$

σ_0 will be non-degenerate if and only if $I := \psi_0^T F''(\rho_0, \lambda_0, a_0)[(\rho'_0(\lambda_0), 1, a'_0(\lambda_0)), (0, 0, \varphi(\lambda_0) + \varepsilon b_0)] \neq 0$; but $I = \frac{d}{d\rho} \psi_0^T D_a F(\rho_0(\lambda), \lambda, a_0(\lambda))(\varphi(\lambda_0) + \varepsilon b_0)_{/\lambda=\lambda_0} = \mu'(\lambda_0) \neq 0$. Next suppose that k is a geometrically simple eigenvalue of A_0 ; then there exists a smooth vector $e(\lambda)$ such that the span of $(b_0(\lambda); e(\lambda))$ is an invariant subspace of $A_0(\lambda)$, $e_0 := e(\lambda_0) = A_0 e_0 - b_0$ and $b_0^T e_0 = 0$; there exists furthermore the vector ψ_0 such that $\psi_0^T A_0 = \psi_0^T$ and $\psi_0^T e_0 = 1$; note that $\psi_0^T b_0 = 0$. We deduce from (5.3) that

$$\begin{aligned} \text{Kernel } F'(\rho_0, \lambda_0, a_0) &= \text{span}\{(\rho'_0(0), 1, a'_0(0)), (-\frac{\rho_0}{2\pi}, 0, e_0)\} \\ \text{Range } F'(\rho_0, \lambda_0, a_0) &= \{(a, \gamma) \in \mathbb{R}^n \times \mathbb{R} \mid \psi_0^T a = 0\}. \end{aligned}$$

σ_0 will be non-degenerate if and only if $I := \psi_0^T F''(\rho_0, \lambda_0, a_0)[(\rho'_0(\lambda_0), 1, a'_0(\lambda_0)), (-\frac{\rho_0}{2\pi}, 0, e_0)] \neq 0$; but $I = \psi_0^T \frac{d}{d\lambda} F'(\rho_0(\lambda), \lambda, a_0(\lambda))(-\frac{\rho_0}{2\pi}, 0, e_0)_{/\lambda=\lambda_0} = \psi_0^T \frac{d}{d\lambda} \{-\frac{\rho_0}{\rho(\lambda)} b_0(\lambda) + (A_0(\lambda) - I)e_0\}_{/\lambda=\lambda_0} = \psi_0^T (-b'_0(\lambda_0) + A'_0(\lambda_0)e_0) = \mu'(\lambda_0) \neq 0$. ■

Proof of Theorem 4.5. Without explicit reference we shall repetitively use the Floquet theory (see for example [3]). Theorem 4.5a is a consequence of Theorem 4.2a. We note that (4.4) implies that $F_2(\rho_0(\lambda), \lambda, a_0(\lambda)) = 0$. Let φ_0 and ψ_0^T be right and left eigenvectors of A_0 for the eigenvalue $\mu(\lambda_0) = -1$. Since 1 is an algebraically and geometrically double eigenvalue of $D_a Z(4\pi; \rho_0, \lambda_0, a_0)$ with eigenvector ψ_0 and b_0 , we can apply Theorem 4.4 for F_2 instead of F . By Theorem 2.3, it remains to show the existence of a characteristic ray of the form $\sigma = (0, 0, \varphi_0 + \alpha b_0)$; first σ belongs to the kernel of $F'_2(\rho_0, \lambda_0, a_0)$ if and only if $b_0^T (\varphi_0 + \alpha b_0) = 0$, which is a condition that determines α uniquely; proving that σ is a characteristic ray then amounts to show that $I := \psi_0^T D_{aa} Z(4\pi; \rho_0, \lambda_0, a_0)[\varphi_0 + \alpha b_0, \varphi_0 + \alpha b_0] = 0$. Let ξ_0 be any eigenvector of $D_a Z(4\pi; \rho_0, \lambda_0, a_0)$ for the eigenvalue 1; we first establish the relation

$$\psi_0^T D_{aa} Z(4\pi; \rho_0, \lambda_0, a_0)[\xi_0, b_0] = 0. \quad (5.6)$$

The matrix $D_a Z(2\pi; \rho_0, \lambda_0, Z((t; \rho_0, \lambda_0, a_0)))$ has a spectrum which is independent of

$t \in \mathbb{R}$; consequently there exists a smooth vector $\psi_0(t)$ such that $\psi_0(0) = \psi_0$ and

$$\psi_0^T(t) D_a Z(4\pi; \rho_0, \lambda_0, Z(t; \rho_0, \lambda_0, a_0)) = \psi_0^T(t), \quad t \in \mathbb{R};$$

by multiplying this relation by ξ_0 and differentiating with respect to t at $t = 0$, one gets (5.6.). Consequently I reduces to $I = \psi_0^T D_{aa} Z(4\pi; \rho_0, \lambda_0, a_0) [\psi_0, \psi_0]$, expression we compute with Formula (5.4); we note that $D_a Z(\tau + 2\pi; \rho_0, \lambda_0, a_0) \psi_0 = -D_a Z(\tau; \rho_0, \lambda_0, a_0) \psi_0$ and $\psi_0^T (D_a Z(\tau + 2\pi; \rho_0, \lambda_0, a_0))^{-1} = -\psi_0^T (D_a Z(\tau; \rho_0, \lambda_0, a_0))^{-1}$; since $D_{xx}^2 f(\lambda_0, Z(\tau; \rho_0, \lambda_0, a_0))$ is 2π -periodic with respect to τ , we obtain that $I = 0$.

The proof of Theorem 4.6 is rather long but can be achieved with the same arguments already used in this section. We shall not present it here.

REFERENCES

- [1] J. Descloux, J. Rappaz: Approximation of solution branches of nonlinear equations. RAIRO, Analyse numérique, 16, 319-349, 1982.
- [2] J. Descloux, J. Rappaz: On numerical approximation of solution branches of nonlinear equations, rapport, Département de mathématiques EPFL, 1981.
- [3] G. Iooss, D. Joseph: Elementary stability and bifurcation theory. Springer-Verlag, New-York, 1980.
- [4] E. Doedel: The numerical computation of branches of periodic solutions. Preprint, Computer Science Dept., Concordia University, Montreal, 1980.
- [5] C. Bernardi: Approximation of Hopf bifurcation. Numer. Math. 39, 15-37, 1982.
- [6] J. Descloux: Numerical approximation of Hopf bifurcation for a parabolic equation, rapport, Département de mathématiques EPFL, 1983.
- [7] F. Brezzi, J. Rappaz, P.A. Raviart: Finite dimensional approximation of nonlinear problems, part III: simple bifurcation points, Numer. Math. 38, 1-30, 1981.
- [8] W.J. Beyn: On discretizations of bifurcation problems. Bifurcation problems and their numerical solution, pp. 46-73. Editors Mittelmann and Weber. ISNM 54, Birkhäuser, Basel, 1980.
- [9] F. Brezzi, J. Descloux, J. Rappaz, B. Zwahlen: On the rotating beam: some theoretical and numerical results. Rapport Dept. Math. EPFL, 1983.
- [10] J. Descloux: Two remarks on continuation methods. In preparation.
- [11] H. Weber: Numerical solution of Hopf bifurcation problems. Math. Meth. Appl. Sciences, 2, 178-190, 1980.
- [12] H. Cartan: Calcul différentiel. Hermann, Paris, 1967.

Jean Descloux, DMA-Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, CH

QUADRATICALLY APPENDED LINEAR MODELS FOR
LOCATING GENERALIZED TURNING POINTS

A. Griewank
Southern Methodist University
Dallas, Texas 75275

Summary

In this paper we consider the numerical location of generalized turning points by forming local model functions which reflect the rank drop characteristics of the original problem. As it turns out the resulting iterative procedure is equivalent to solving a certain augmented nonlinear system recently introduced by A. Griewank and G. W. Reddien [6-8]. After discussing various ways of obtaining the required second derivative information the paper concludes with some numerical experiments in computing a simple bifurcation point of a two point boundary value problem.

Generalized Turning Points and Simple Bifurcation

For $m \leq n$ consider a vector function

$$F(x, \lambda) : D \subset \mathbb{R}^{n+1} \rightarrow \mathbb{R}^m ,$$

whose Lipschitz continuous Jacobian matrix

$$\nabla_{x, \lambda} F \equiv (\nabla_x F, \partial F / \partial \lambda)$$

has full row rank m in the region of interest D . Then the solution set $F^{-1}(0) \subset D$ forms a smooth $p = n+1-m$ dimensional manifold whose tangent plane is perpendicular to the λ -axis in \mathbb{R}^{n+1} if and only if

$$\text{rank}(\nabla_x F) = m-1 .$$

Any solution $(x^*, \lambda^*) \in F^{-1}(0)$ at which this rank drop occurs will be called a generalized turning point of F with respect to the distinguished variable λ . At such points the restriction of λ to $F^{-1}(0)$ attains a stationary value whose Hessian with respect to a suitable parameterization is given by the matrix

$$H = \{ u^T \nabla_x^2 F(x^*, \lambda^*) v_i v_j \}_{\substack{i=1..p \\ j=1..p}} . \quad (1)$$

Here $u \in \mathbb{R}^{n-1}$ is a nonzero left null vector of the Jacobian $\nabla_x F(x^*, \lambda^*)$ and the p vectors $v_i \in \mathbb{R}^n$ form a basis of its null space. The symmetric $p \times p$ matrix H is important because its inertia determines the structure of the slices $F^{-1}(0) \cap \{\lambda = \hat{\lambda}\}$ for fixed $\hat{\lambda} \approx \lambda^*$. As a general nondegeneracy condition we assume that H is nonsingular.

The concept of generalized turning points can be used to characterize constrained optima, singular solutions of square systems and in particular simple bifurcation points or isolas [6]. The last two possibilities will be of main interest throughout the remainder of this paper.

Given an underdetermined system

$$f(x) = 0, \quad f : D \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^{n-1}$$

with

$$\text{rank}(\nabla_x f(x)) \geq n - 2 \quad \text{at all } x \in D$$

we can always find a vector $r \in \mathbb{R}^{n-1}$ such that the unfolding

$$F(x, \lambda) \equiv f(x) + \lambda r : D \times \mathbb{R} \rightarrow \mathbb{R}^{n-1} \tag{2}$$

satisfies the assumptions above on a suitable restricted domain D . Then the generalized turning points (x^*, λ^*) with $\lambda^* = 0$ correspond exactly to the bifurcation points or isolas of f depending on whether the 2×2 matrix H has a negative or positive determinant. If H is singular x^* is likely to be a cusp point and if λ^* is nonzero but small we have some kind of perturbed bifurcation. It should be noted that λ is not a control parameter but an unfolding variable whose value at a generalized turning point of (2) measures the imperfection of the bifurcation. For simplicity we will assume in the remainder that $\partial F / \partial \lambda = r$ is constant as is the case in (2).

Forming Local Models with Generalized Turning Points

Since linear functions have constant Jacobians they cannot exhibit any of the rank drop phenomena mentioned above. Therefore we have to augment the first two terms in the Taylor series by a quadratic term in order to obtain a local model that can reflect the essential features of the given function F . Because of our assumption that the dimension of range $(\nabla_x F)$ varies only by 1 we can avoid the full complexity of the second derivative tensor $\nabla_x^2 F$. Instead we consider only a simple term of the form $\frac{1}{2}rs^T B s$ where

the symmetric $n \times n$ matrix B should be chosen such that at a fixed current point $(x, \lambda) \in \mathbb{R}^{n+1}$ with $F = F(x, \lambda)$ and $\nabla_x F = \nabla_x F(x)$

$$F(x+s, \lambda+\mu) \approx L(x, \mu) \equiv F + \nabla_x F \cdot s + r(\mu - \frac{1}{2}s^T B s). \quad (3)$$

The local model function $L(s, \mu) : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n-1}$ has the Jacobian

$$\nabla_s L(s, \mu) = \nabla_x F + rs^T B$$

which has the reduced rank $m - 1 = n - p$ if and only if for suitable matrices $T, V \in \mathbb{R}^{n \times p}$

$$A \begin{pmatrix} V \\ -g^T \end{pmatrix} \equiv \begin{pmatrix} \nabla_x F, & r \\ T^T, & 0 \end{pmatrix} \begin{pmatrix} V \\ -g^T \end{pmatrix} = \begin{pmatrix} 0 \\ I_p \end{pmatrix} \quad (4)$$

with

$$g^T \equiv s^T B V \in \mathbb{R}^p. \quad (5)$$

The columns of V span the null space of $\nabla_s H$ and are normalized by the columns of T which must be chosen linearly independent and nonorthogonal to the tangent space of $F^{-1}(0)$. Then the $(n+1) \times (n+1)$ matrix $A = A(x)$ is nonsingular so that the resulting matrix $V = V(x)$ and vector $g = g(x)$ are differentiable in $x \in D$. Combining (3) and (5) we find that $(s^*, \mu^*) \in \mathbb{R}^{n+1}$ is a generalized turning point of L if and only if

$$\begin{pmatrix} \nabla_x F, & r \\ V^T B, & 0 \end{pmatrix} \begin{pmatrix} s^* \\ \mu^* - \frac{1}{2}s^{*T} B s^* \end{pmatrix} = \begin{pmatrix} -F \\ -g \end{pmatrix}. \quad (6)$$

The $(n+1) \times (n+1)$ matrix on the left-hand side is nonsingular provided

$$\det(V^T B V) \neq 0 \quad (7)$$

in which case (s^*, μ^*) is uniquely defined by (6).

Now suppose we generate a sequence of iterates

$$x_{j+1} = x_j + s_j, \quad \lambda_{j+1} = \lambda_j + \mu_j$$

with $(s_j, \mu_j) = (s_j^*, \mu_j^*)$ obtained as generalized turning points of the model $L(s, \mu)$ formed at (x_j, λ_j) with some matrix B_j satisfying (7) for $V = V_j \equiv V(x_j)$.

If the sequence converges at all, the right-hand side of (6) must vanish at the limit point $(\mathbf{x}^*, \lambda^*)$ which is thus a solution of the nonlinear system

$$F(\mathbf{x}, \lambda) = 0, \quad g(\mathbf{x}, \lambda) = 0 \quad (8)$$

and hence by (4) a generalized turning point of F . Now it follows from a theorem of Dennis and More [3] that superlinear convergence

$$\lim_{j \rightarrow \infty} \|(\mathbf{x}_{j+1} - \mathbf{x}^*, \lambda_{j+1} - \lambda^*)\| / \|(\mathbf{x}_j - \mathbf{x}^*, \lambda_j - \lambda^*)\| = 0$$

to the root of a square nonsingular system occurs if and only if the relative error between the actual step taken and the Newton correction tends to zero. By comparison of (6) with the exact Jacobian of (8) we see that this is equivalent to the condition

$$\lim_{j \rightarrow \infty} \|\nabla_{\mathbf{x}}^T B_j - \nabla_{\mathbf{x}}^T g(\mathbf{x}_j) s_j\| / \|(\mathbf{s}_j, \mu_j)\| = 0.$$

By implicit differentiation of (4) one can show [6] that

$$\nabla_{\mathbf{x}}^T g(\mathbf{x}) = V(\mathbf{x})^T E(\mathbf{x}) \quad \text{with} \quad E(\mathbf{x}) \equiv \nabla_{\mathbf{x}}^2(u^T F(\mathbf{x}, \lambda))$$

where the approximating left null vector $u \in \mathbb{R}^{p-1}$ is defined by

$$(u^T, -g^T) A = (0^T, 1) \in \mathbb{R}^{n+1}$$

Consequently the optimal choice for B seems to be the symmetric matrix $E(\mathbf{x})$ which is the Hessian of the Lagrangian in the optimization case. For B sufficiently close to E the nondegeneracy condition (7) follows from the assumption that the matrix H as defined in (1) is nonsingular.

Obtaining Second Derivative Information

Unless the approximating left null vector $u = u(\mathbf{x})$ happens to be sparse $E(\mathbf{x})$ depends on the full second derivative tensor $\nabla_{\mathbf{x}}^2 F$ so that its analytical evaluation can be rather costly. Therefore one may prefer to update approximations $B_j \approx E(\mathbf{x}_j)$ on the basis of the secant condition

$$s_j^T B_{j+1} = y_j^T \equiv u_j^T [\nabla_{\mathbf{x}}^T F(\mathbf{x}_j + \mathbf{s}_j) - \nabla_{\mathbf{x}}^T F(\mathbf{x}_j)] = s_j^T E_{j+1} + o(\|\mathbf{s}_j\|). \quad (9)$$

In the bifurcation case the projected Hessian $H = V^T E V$ is necessarily indefinite so that $y_j^T s_j$ may be positive, negative or zero even if the step s_j

belongs to the tangent space range (V_j). Hence we cannot apply any of the positive definite updates used in optimization [4, 10] and apparently there is no other way to "take out the scaling", i.e., to make the updating procedure independent of Euclidean norms and other incidental features of the problem. Instead we could use the Broyden update [4]

$$B_{j+1} = B_j + a_j s_j^T / s_j^T s_j \quad \text{with } a_j = y_j - B_j s_j \quad (10)$$

or in order to maintain symmetry the PSB formula

$$B_{j+1} = B_j + (a_j s_j^T + s_j a_j^T) / s_j^T s_j - (a_j^T s_j) s_j s_j^T / (s_j^T s_j)^2. \quad (11)$$

However, only the projection $v_{j+1}^T B_j \in \mathbb{R}^{p \times n}$ enters into the definition of the step s_j so that the full storage of $B_j \in \mathbb{R}^{n \times n}$ may be rather wasteful especially when p is small as in the bifurcation case where $p = 2$.

Then it is preferable to update only an approximation $G_j \approx \nabla_x g(x_j)$ on the basis of the secant condition

$$G_{j+1} s_j = z_j \equiv g(x_{j+1}) - g(x_j) = v_{j+1}^T y_j \quad (12)$$

where the last equation follows from (4) and (9). For the optimization case Overton has recently proposed to update $G_j \in \mathbb{R}^{p \times n}$ by the Broyden formula

$$G_{j+1} = G_j + (z_j - G_j s_j) s_j^T / s_j^T s_j \quad (13)$$

which ensures local and Q-superlinear convergence by the standard theory of least change secant updates [4]. However, this update does not maintain the natural condition that $G_{j+1} V_{j+1} \approx H(x_{j+1})$ be symmetric. A simple way to symmetrize $G_j V_{j+1}$ is given by

$$G_{j+\frac{1}{2}} \equiv \frac{1}{2} [I + (G_j V_{j+1})^T (G_j V_{j+1})^{-1}] G_j. \quad (14)$$

In order to satisfy the secant condition (12) we now have to find an update from $G_{j+\frac{1}{2}}$ to G_{j+1} that maintains the symmetry with respect to V_{j+1} . Moreover, we want to exploit the fact that, when $\nabla_x F$ is evaluated analytically but $\nabla_x g$ only approximated, then all but the first few steps are more or less tangential to the manifold $F^{-1}(0)$. If any step is exactly of the form $s_j = V_{j+1} c_j$ for some coefficient vector

$$c_j = (G_{j+2} v_{j+1})^{-1} G_{j+2} s_j \quad (15)$$

then it follows from (9) that G_{j+1} should satisfy

$$c_j^T G_{j+1} = c_j^T v_{j+1}^T B_{j+1} = y_j^T = c_j^T \nabla_x g(x_{j+1}) + o(\|s_j\|). \quad (16)$$

This transposed secant condition represents n linear equations compared to merely p imposed by (12). It can be satisfied by the rank two update

$$G_{j+1} = G_{j+2} + v_{j+1}^T \left[\frac{b_j s_j^T + b_j s_j^T}{s_j^T s_j} - \frac{s_j s_j^T b_j s_j}{(s_j^T s_j)^2} \right] \quad (17)$$

where

$$b_j^T = y_j^T - c_j^T G_{j+2} \quad \text{with} \quad v_{j+1}^T b_j = a_j.$$

Since c_j as given by (15) and consequently b_j are always well defined, even s_j is not exactly tangential, the formula can be applied over arbitrary steps. It always satisfies the secant condition (12) while ensuring the symmetry of $G_{j+1} v_{j+1}$. Whenever $s_j \in \text{range}(v_{j+1})$ the formula represents a projected version of the PSB formula (11) and in the unlikely event that $s_j \in \text{range}(v_{j+1})^\perp$ it reduces to the Broyden update (10).

Alternatively one can utilize (16) by taking special steps along the columns of v_{j+1} which allows the row wise evaluation of $G_{j+1} = \nabla g(x_{j+1})$ at the expense of p additional evaluations but not factorizations of $\nabla_x F$. If even the Jacobian $\nabla_x F$ can only be obtained by differencing the projected Hessian H may still be computed on the basis of second divided differences in F as suggested in [9]. For simple turning points this approach requires an intermediate correction step towards $F^{-1}(0)$ and may be viewed as a variant of the method of Brent and Brown.

Numerical Experiments

The following results were obtained on the two point boundary value problem

$$x'' + \xi \phi(x, t, \xi) = 0, \quad x(0) = 0 = x(1)$$

where

$$\phi(x, t, \xi) = x - \xi\psi(t) + \frac{1}{2}(x - \xi\psi(t))^2 - \psi''(t)$$

with

$$\psi(t) = e^t \sin \pi t.$$

This boundary value problem has a simple bifurcation point at

$$x = \xi\psi \text{ with } \xi = (k\pi)^2$$

for some integer k.

Discretizing the problem by Numerov's method we obtain for both linear systems (4) and (6) a sparsity problem of the arrow-like form

$$\begin{array}{ccccccccc} x & \tilde{1} & & & x & x & & & \\ \tilde{1} & x & \tilde{1} & & x & x & & & \\ \tilde{1} & x & & & x & x & & & \\ & & & & \cdot & \cdot & & & \\ & & & & \cdot & \cdot & & & \\ & & & & x & \tilde{1} & x & x & \\ & & & & \tilde{1} & x & x & x & \\ x & x & x & . & . & . & x & x & x & 0 \\ x & x & x & . & . & . & x & x & x & 0 \end{array}$$

Since for a sufficiently small discretization width h the off diagonal elements marked by $\tilde{1}$ are nearly equal to 1 they can be used to eliminate all entries to the left of the box in the last two rows. Because the resulting 2×2 matrix in that box is nonsingular for suitable choices of T we may then back solve for two different values of the last variables. The resulting two vectors can be combined linearly to satisfy the first equation and hence the whole linear system. This kind of procedure for solving bordered sparse systems was originally proposed by Keller [5] and has recently been examined by Chan [2]. At each iterate we have to perform two eliminations, one to compute v_j, g_j and u_j with the last rows given by T, and one to compute the actual step with the last rows given by $G_j \approx \nabla_x g(x_j)$.

The calculations were started from the trivial function $x_0(t) = 0$ with $\xi_0 = 10$ and stopped when the stepsize fell below 10^{-7} . The computation

were performed in interpretive Basic on a Sharp PC 1500 with 10 digits accuracy and 10000 characters memory. The following table lists the number of iterations and the computing time in minutes for Newton's method and two quasi-Newton schemes using the Broyden formula (13) or the Ad Hoc combination of (14) and (17), respectively. The entries in the first two columns give the discrepancies $\xi_h - \pi^2$ and the imperfection parameters λ_h which both decine like h^{-4} , as is to be expected for Numerov's method.

h^{-1}	$\xi_h - \pi^2$	λ_h	Newton	Broyden	Ad Hoc
5	-7.3-3	1.8-2	7/4	14/7	12/6
10	-4.5-4	1.0-3	7/7.7	15/14	13/12.5
20	-2.8-5	6.4-5	7/15	15/27.5	13/24
40	-1.7-6	4.1-6	7/30	13/48	12/44.5
			Iter/Mins	Iter/Mins	Iter/Mins

The results in the table show clearly that secant updating of $G_j \approx \nabla_x g$ leads still to rather rapid convergence which is essentially unaffected with respect to grid refinements. This invariance principle has been established by Allgower et al [1] for Newton's method but is as yet unproven for secant updates. The imposition of symmetry helps a little bit but not as much as one would have hoped and further experimenting seems required. Due to the sparsity of $\nabla_x F$ the evaluation time and overall computing time grows linearly in h^{-1} . On the test problem, Newton's method does not need any more time per step, because all derivatives were obtained analytically at essentially no extra cost.

References

- [1] Allgower, E. L. and Bohmer, K. (1982), Mesh independence for operator equations and their discretizations, Bonner Preprints, 456.
- [2] Chan, T. F. (1983), Techniques for large sparse systems arising from continuation methods, these Proceedings.
- [3] Dennis, J. E. and More, J. J. (1974), A characterization of superlinear convergence and its application to quasi-Newton methods, Mathematics of Computation 28, 549-560.
- [4] Dennis, J. E. and More, J. J. (1988), Quasi-Newton methods, motivation and theory, SIAM Review 19, 46-89.

- [5] Keller, H. B. (1977), Numerical solution of bifurcation and nonlinear eigenvalue problems, in Applications of Bifurcation Theory, Edited by P. Rabinowitz, Academic Press, N.Y., 359-384.
- [6] A. Griewank and G. W. Reddien, Characterization and computation of generalized turning points, SIAM J. Numer. Anal., in press.
- [7] A. Griewank and G. W. Reddien, Computation of turning and bifurcation points for two-point boundary-value problems, in Proceedings, The Seventh Annual Lecture Series in the Mathematical Sciences, University of Arkansas, Fayetteville, Arkansas, March 24-26, 1983.
- [8] A. Griewank and G. W. Reddien (1983), Computation of generalized turning points and two point boundary value problems, these Proceedings.
- [9] Pönisch, G. and M. Schwetlich (1981), Computing turning points of curves implicitly defined by nonlinear equations depending on a parameter, Computing 26, 107-121.
- [10] Powell, M. H. D., (Ed.) (1982), Nonlinear Optimization 1981, Academic Press, New York.

Hysteresis in a Model for Parasitic Infection

K. P. Hadeler, Tübingen

In the classical epidemic model of Kermack and McKendrick 1927 and also in its various extensions only the prevalence of the disease in the host population is considered. The latter is subdivided into several classes such as susceptibles S , infectious I , and recovered R . The transition $S \rightarrow I \rightarrow R \rightarrow S$ is modeled by ordinary differential equations

$$\begin{aligned}\dot{S} &= -\beta SI + \gamma R \\ \dot{I} &= \beta SI - \alpha I \\ \dot{R} &= \alpha I - \gamma R\end{aligned}\tag{1}$$

The total population size $P = S + I + R$ is constant, thus the system is essentially two-dimensional,

$$\begin{aligned}\dot{S} &= -\beta SI + \gamma(P - S - I) \\ \dot{I} &= \beta SI - \alpha I\end{aligned}\tag{2}$$

The state space is the triangle $T = \{S \geq 0, I \geq 0, S + I \leq P\}$.

The parameter $\tau = \alpha/\beta$, i. e. the rate of recovery α divided by the contact rate β , plays an essential rôle, which is described in the threshold theorem:

In the endemic case $\gamma > 0$ there is a simple bifurcation. If $\tau > P$ then the disease cannot persist, all trajectories converge to $(P, 0)$. For $\tau < P$ there is a stable stationary state $(S_1, I_1) = (\tau, \gamma(P - \tau)/(\beta + \gamma))$ which attracts all trajectories in T except $(P, 0)$. Thus if the contact rate β runs from 0 to infinity, then at $\beta = \alpha/P$ the noninfected state loses its stability and gives rise to a branch of infected states.

In the epidemic case $\gamma = 0$ there is a line of stationary points $(S, 0)$. Again the number τ plays the rôle of a threshold: For initial data $(S(0), I(0))$ with $S(0) < \tau$ the function $I(t)$ is decreasing, while for $S(0) > \tau$ the number of infectious is first increasing, then levels off to a positive limit. In particular for $P < \tau$ the variable I decreases along all trajectories in T .

The prevalence models are well-suited to describe bacterial or viral diseases. On the other hand, in helminthic infections such as in onchocerciasis and other filariasis infections, the host acquires a small number of parasites at different times, and the degree of illness is related to the number of parasites, also there is no direct infection, the disease is transmitted by vectors. For such diseases a model has been proposed [2] which takes account of the individual ages of hosts, and also of the possibly nonlinear relation between average parasite load and parasite acquisition. An analogue of the Kermack-McKendrick threshold theorem occurs in the form of a bifurcation phenomenon.

Let t be chronological time and let a be the age of hosts. The following parameters are introduced: $\mu(a)$ is the age dependent mortality of hosts in the absence of parasites, $b(a)$ is the contribution of the age class a of hosts to the birth of new hosts.

The birth rate of parasites within the host is $\varphi \geq 0$. In typical helminthic infections one has $\varphi = 0$, since larvae can only develop after an intermediate stage in the vector. The death rate of parasites in the host is $\sigma > 0$, and $\varphi(t)$ is the acquisition rate of parasites. The parameter α is the differential death rate of hosts due to the presence of one parasite. Thus the mortality of a host of age a carrying r parasites is $\mu(a) + r\alpha$, and parasites act on the mortality of the host independently. This independence assumption is of great importance in the following mathematical treatment.

The acquisition rate φ of new parasites by the hosts is related to the average parasite load \bar{w} of the population by a function

$$\varphi = \beta f(\bar{w}) \quad (3)$$

Here f is function $f: [0, \infty) \rightarrow [0, \infty)$ with $f(0) = 0$, $f'(0) = 1$, $f(u) > 0$ for $u > 0$, and $\beta > 0$ is a parameter.

Let $n(t, a, r)da$ be the size of the cohort with age in $[a, a + da]$ at time t carrying r parasites. For the functions $n(t, a, r)$ one can derive an infinite system of equations (see [2], [4], [6]) which can be condensed into a single partial

differential equation for the generating function

$$u(t, a, z) = \sum_{r=0}^{\infty} n(t, a, r) z^r, \quad (4)$$

namely

$$u_t + u_a + g(z)u_z - [\varphi(t)(z-1) - \mu(a)]u = 0 \quad (5)$$

where

$$g(z) = (\alpha + \sigma + \rho)z - \sigma - \rho z^2. \quad (6)$$

The average parasite load is then

$$\bar{w}(t) = \frac{\int_0^\infty u_z(t, a, 1) da}{\int_0^\infty u(t, a, 1) da}. \quad (7)$$

The initial state of the population is given by a condition

$$u(0, a, z) = u_0(a, z). \quad (8)$$

The assumption of a prescribed host birth rate leads to the simple boundary condition

$$u(t, 0, z) = N(t). \quad (9)$$

In a biologically more realistic situation one can relate the birth rate of hosts to the total host population by a Lotka birth law [5]

$$u(t, 0, z) = \int_0^\infty b(a)u(t, a, \omega) da \quad (10)$$

where $\omega \in [0, 1]$ is a parameter: For $\omega = 1$ the infection does not influence reproduction, for $0 < \omega < 1$ the parasite load leads to a geometric decrease of fertility, for $\omega = 0$ only non-infected hosts reproduce.

In [2], [4], [6] the following approach has been chosen: One assumes that $\varphi(t)$ were known. Then, using the boundary conditions (8), (9), one can solve the linear partial differential equation by the method of characteristics. From this solution one can determine the average parasite load \bar{w} according to (7) and introduce it into equation (3). In this way a nonlinear Volterra integral equation for the parasite acquisition rate $\varphi(t)$ is obtained

$$\varphi(t) = \beta f(\bar{w}(t)) \quad (11)$$

$$\bar{w}(t) = \frac{\int_0^t e^{A_1\varphi - M(a)} N B_1 \varphi da + e^{C_1\varphi} \int_t^\infty e^{-M(a-t) + M(a)} [u_{0z} G_z + u_0 D_1 \varphi] da}{\int_0^t e^{A_1\varphi - M(a)} N da + e^{C_1\varphi} \int_t^\infty e^{-M(a-t) + M(a)} u_0 da}$$

Here G is the Riccati solution operator

$$G(t, z) = \frac{z_1(z-z_2) - z_2(z-z_1)e^{-xt}}{(z-z_2) - (z-z_1)e^{-xt}}$$

with $z_1 \geq 1 \geq z_2$

$$z_{1,2} = \frac{1}{2\varphi} [\alpha + \sigma + \varphi + \sqrt{(\alpha + \sigma + \varphi)^2 - 4\sigma\varphi}]^{1/2}$$

$$\alpha = \sqrt{(\alpha + \sigma + \varphi)^2 - 4\sigma\varphi}$$

In the special case $\varphi = 0$ one has with $\alpha = \alpha + \sigma$

$$G(t, z) = 1 - (1-z)e^{\alpha t} + \frac{\alpha}{\alpha} (1-e^{\alpha t})$$

Furthermore

$$(A_\omega \varphi)(t, a) = \int_{t-a}^a [G(s-t, \omega) - 1] \varphi(s) ds$$

$$(B_\omega \varphi)(t, a) = \int_{t-a}^a G_z(s-t, \omega) \varphi(s) ds$$

$$(C_\omega \varphi)(t) = \int_0^t [G(s-t, \omega) - 1] \varphi(s) ds$$

$$(D_\omega \varphi)(t) = \int_0^t G_z(s-t, \omega) \varphi(s) ds$$

$$M(a) = \int_0^a \mu(s) ds$$

The arguments of u_0 and N are of the form $u_0(a-t, G(-t, 1)), N(t-a)$.

In [2], [4], [6] it has been shown, that this equation has a unique global solution under suitable conditions on f, μ , and u_0 . From this solution $\varphi(t)$ the function $u(t, a, z)$ can be easily obtained via the solution of the initial value problem (5), (8), (9).

The stationary problem is of particular interest. First consider the boundary condition (9) with N constant. For a stationary solution the acquisition rate φ satisfies the equation

$$\varphi = \beta f(W(\varphi)\varphi) \quad (12)$$

where

$$W(\varphi) = I_1(\varphi)/I_0(\varphi)$$

$$I_0(\varphi) = \int_0^\infty \exp(-Q_1(a)\varphi - M(a))da$$

$$I_1(\varphi) = \int_0^\infty \exp(-Q_1(a)\varphi - M(a))q(a)da$$

$$Q_\omega(a) = -(z_1 - 1)a + \frac{1}{\varphi} \log \frac{(\omega - z_2) - (z_1 - \omega)e^{\chi a}}{z_1 - z_2}$$

$$q(a) = \frac{1}{\varphi} \frac{e^{\chi a} - 1}{z_2 + (z_1 - 1)e^{\chi a}}$$

For $\varphi = 0$ one has

$$Q_\omega(a) = \frac{\alpha}{\chi} \left(a - \frac{1}{\chi} (1 - e^{-\chi a}) \right) - (1 - \omega) \frac{1}{\chi} (1 - e^{-\chi a}) \quad (13)$$

$$q(a) = \frac{1}{\chi} (1 - e^{-\chi a})$$

Since $f(0) = 0$, the value $\varphi = 0$, corresponding to a non-infected host population, is always stationary. Furthermore equation (12) yields a solution branch

$$\beta = \varphi/f(W(\varphi)\varphi) \quad (14)$$

where φ runs from 0 to $+\infty$. In view of $f'(0) = 1$ this branch starts at

$$\beta_0 = \int_0^\infty e^{-M(a)}da / \int_0^\infty e^{-M(a)}q(a)da. \quad (15)$$

In [2], [3] the following analogue of the Threshold Theorem has been shown: For $\beta < \beta_0$ the stationary solution corresponding to $\varphi = 0$ is stable, for $\beta > \beta_0$ it loses its stability, and the population enters a state with endemic disease.

In [2] it has been shown that under a growth condition, $f(u)u^{-2} \rightarrow 0$ for $u \rightarrow \infty$, the branch exists for all $\beta > \beta_0$.

It has also been shown [2] that under a Krasnoselskij concavity condition

$$\frac{df(u)}{du} \geq 0, \quad \frac{d}{du} \left(\frac{f(u)}{u} \right) \leq 0 \quad (16)$$

the branch is monotone, i. e. for every $\beta > \beta_0$ there is exactly one stationary $\varphi > 0$.

The following example shows numerically that monotonicity of f is not sufficient to ensure the monotonicity of the branch of stationary solutions. The function

$$f(u) = \begin{cases} u, & u \leq 1 \\ ((u-2)^3 + 4)/3, & 1 \leq u \leq 4 \\ (8u - 28)/(u - 3), & u \geq 4 \end{cases} \quad (17)$$

is continuously differentiable, sublinear and monotone. The concavity condition is violated. For $\mu = 0.4, \sigma = 0.5, \alpha = 0.5, \rho = 0$ the branch of nontrivial stationary solutions has two turning points (Figure 1). Thus in this case the model exhibits a hysteresis phenomenon: Increase of the contact rate β leads to a sudden jump in parasite acquisition, subsequent decrease of β does not immediately lead to the earlier level.

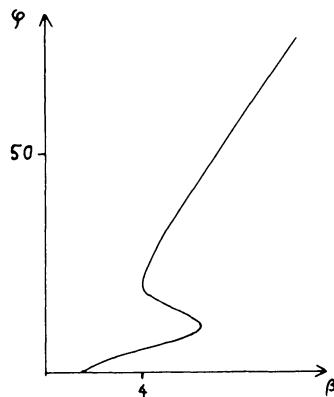


Figure 1

Finally we determine the direction of the bifurcating branch. In (14) we form the derivative with respect to φ and expand. In the limit for $\varphi \rightarrow 0$ one obtains

$$\left. \frac{d\beta}{d\varphi} \right|_{\varphi=0} = -\frac{1}{2} f''(0) + \left. \frac{d}{d\varphi} \left(\frac{1}{W} \right) \right|_{\varphi=0} \quad (18)$$

The second term is positive, since W is a decreasing function of φ (see [2]). Hence there will be a backward bifurcation only if $f''(0)$ is sufficiently large.

In the special case where parasites do not multiply in the host, $\varphi = 0$, and the death rate of hosts is independent of age, $\mu(a) \equiv \mu$, one can evaluate all integrals explicitly,

$$I_0(0) = \frac{1}{\mu}, \quad I'_0(0) = -\frac{\alpha}{\mu^2}$$

$$I_1(0) = \frac{1}{\mu(1+\chi)}, \quad I'_1(0) = -\frac{\alpha(3\mu+2\chi)}{\mu^2(\mu+\chi)^2(\mu+2\chi)}$$

and

$$\left. \left(\frac{1}{W} \right)' \right|_{\varphi=0} = \frac{2\alpha}{2\chi+\mu}. \quad (19)$$

The right hand side is less than 1. Thus for $f''(0) \geq 2$ backward bifurcation is unavoidable.

Figure 2 shows the example $\sigma = 0, 8$, $\alpha = 0, 1$, $\rho = 0$, $\mu \equiv 1$

$$f(u) = \frac{u + u^2}{1+u/2+u^2/10} = u + \frac{1}{2}u^2 + \dots$$

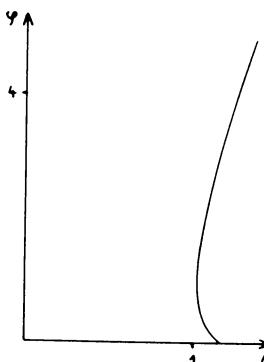


Figure 2

Again, in the case of backward bifurcation, one expects hysteresis between the non-infected and the endemic state.

Secondly consider the case of a Lotka birth law (10). Instead of constant stationary solutions one looks for stationary age distributions $u(a, z)$ with constant acquisition rate φ and exponential growth of the total population size with exponent λ . In [5] it has been shown that the pair φ, λ can be determined from the equations

$$\begin{aligned} \varphi &= \beta f(W(\varphi, \lambda))\varphi \\ \int_0^\infty b(a) \exp(-Q_w(a)\varphi - M(a) - \lambda a) da &= 1 \end{aligned} \quad (20)$$

where

$$\begin{aligned} W(\varphi, \lambda) &= I_1(\varphi, \lambda)/I_0(\varphi, \lambda) \\ I_0(\varphi, \lambda) &= \int_0^\infty \exp(-Q_1(a)\varphi - M(a) - \lambda a) da \\ I_1(\varphi, \lambda) &= \int_0^\infty \exp(-Q_1(a)\varphi - M(a) - \lambda a) q(a) da \end{aligned}$$

There is a branch of nontrivial solutions, which bifurcates from $\varphi = 0$ at

$$\beta_1 = \int_0^\infty e^{-M(a) - \lambda_0 a} da / \int_0^\infty e^{-M(a) - \lambda_0 a} q(a) da \quad (21)$$

where λ_0 is the unique root of the equation

$$\int_0^\infty b(a) e^{-M(a) - \lambda a} da = 1. \quad (22)$$

For the direction of the bifurcation at $\varphi = 0$ one finds

$$\left. \frac{d\beta}{d\varphi} \right|_{\varphi=0} = -\frac{1}{2} f'''(0) + \left[\frac{d}{d\varphi} \left(\frac{1}{W} \right) + \frac{d}{d\lambda} \left(\frac{1}{W} \right) \cdot \frac{d\lambda}{d\varphi} \right]_{\varphi=0} \quad (23)$$

where from (20)

$$\left. \frac{d\lambda}{d\varphi} \right|_{\varphi=0} = -\frac{\int_0^\infty b(a) e^{-M(a) - \lambda_0 a} Q_w(a) da}{\int_0^\infty b(a) e^{-M(a) - \lambda_0 a} a da} \quad (24)$$

Again we consider the special case $\varphi = 0$, $\mu(a) \equiv 1$.

Then

$$\frac{d}{d\varphi} \left(\frac{1}{W} \right) \Big|_{\varphi=0} = \frac{2\alpha}{2\alpha + \mu + \lambda_0} \quad (25)$$

and

$$\frac{d}{d\lambda} \left(\frac{1}{W} \right) \Big|_{\varphi=0} = 1 .$$

The special case $b(a) \equiv 1$ is very simple. Then

$$\lambda_0 = 1 - \mu , \quad \frac{d\lambda}{d\varphi} \Big|_{\varphi=0} = \frac{(1 - \omega) - \alpha}{1 + \alpha}$$

and thus

$$\frac{d\beta}{d\varphi} \Big|_{\varphi=0} = - \frac{1}{2} f''(0) + \frac{2\alpha}{1+2\alpha} + \frac{(1 - \omega) - \alpha}{1 + \alpha} \quad (26)$$

Finally we consider the stability problem. As we have mentioned earlier, in the case of the boundary condition (9), the trivial branch of stationary solutions $\varphi = 0$ is stable for $\beta < \beta_0$ ([2], [3]). Nothing was known till now about the stability of the nontrivial branch. In a forthcoming publication [8] I show that the stability of the nontrivial stationary solutions agrees with the slope of the bifurcation diagram: they are stable for $d\beta/d\varphi > 0$ and unstable for $d\beta/d\varphi < 0$.

References

- [1] Anderson, R.M., May, R.M., The transmission dynamics of human helminth infections, control by chemotherapy , Nature 297: 557-563 (1982).
- [2] Hadeler, K.P., Dietz, K., Nonlinear hyperbolic partial differential equations for the dynamics of parasite populations, Computers and Math. with Appl. Vol.9, Nr. 3, p.415-430 (1983).
- [3] Hadeler, K.P., An integral equation for helminthic infections: Stability of the non-infected population. Proc. Vth Int. Conf. on trends in Theory and Practice of Nonl. Diff. Equ. 1982 (Ed. V. Lakshmikantham).

- [4] Hadeler, K.P., Dietz, K., An integral equation for helminthic infections: Global existence of solutions, In: Conference Proceedings "Recent Trends in Mathematics", Reinhardtsbrunn 1982, Teubner Texte zur Mathematik 50, Teubner Verlag Leipzig 1982/83.
- [5] Hadeler, K.P., Integral equations for infections with discrete parasites: Hosts with Lotka birth law. In: Autumn course on Mathematical Ecology, Trieste 1982, (Ed. S. Levin, T. Hallam).
- [6] Hadeler, K.P., Dietz, K., A transmission model for multiplying parasites and killing. J. Math. Biol. mscr. submitted.
- [7] Karlin, S., Tavaré, S., Linear birth and death processes with killing. J. Appl. Prob. 19, 477-487 (1982).
- [8] Hadeler, K.P., A transmission model for multiplying parasites and killing: Stability of the endemic states. In preparation.

Lehrstuhl für Biomathematik
Universität Tübingen
Auf der Morgenstelle 28
7400 Tübingen

CONTINUATION OF PERIODIC SOLUTIONS IN ORDINARY DIFFERENTIAL
EQUATIONS - NUMERICAL ALGORITHM AND APPLICATION TO LORENZ MODEL

Martin Holodniok and Milan Kubíček

An algorithm for continuation of periodic solutions in O.D.E. in dependence on a parameter is presented and applied to the Lorenz model. The algorithm is based on the shooting method coupled with continuation along the arc of the solution locus. The stability of periodic solutions is determined by characteristic multipliers computed in the course of the continuation.

The algorithm crosses without difficulties limit points and in the case of bifurcation points, period doubling bifurcation points and points of tori bifurcation it proceeds along the original branch of solutions.

I. INTRODUCTION

Let us consider mathematical models in the form of systems of nonlinear ordinary differential equations depending on a chosen physical parameter. Steady state solutions of the models then result from a set of nonlinear (algebraic) equations. A number of methods for automatic computation of the dependence of steady state solutions on a parameter have been developed, e.g., the methods using the arc-length of solution locus as a parameter for continuation [8,7,14].

To obtain and continue periodic solutions of the system of ordinary differential equations is a more difficult task. Three different approaches can be designed for computation of periodic solutions. The easiest one consists in a dynamic simulation of the studied system leading to a stable periodic orbit. Finite difference methods are used in the

second approach. The third approach is based on the shooting method, cf.e.g. [3,11,20].

The sequential approach was used for the continuation of periodic solutions by Hassard [4], Rinzel and Miller[15], Seydel [20] and Chibnik [2].

An algorithm for the continuation of periodic solutions based on the shooting method and continuation along the arc-length of solution locus, DERPAR [8], is described in this paper. This algorithm was successfully applied to a number of practical problems. For example, the problem of two interconnected well mixed cells with the Brusselator chemical reaction schema [13] has been studied; results of the continuation of periodic solutions will be summarized in the paper [19]. Results for the Lorenz model [9] are presented in this paper. The complete picture of dependences of periodic solutions on the parameter with many examples of different types of periodic solutions was published in [6].

II. DEVELOPMENT OF THE ALGORITHM

Let us consider an autonomous system of ordinary differential equations

$$\frac{dy_i}{dt} = f_i(y_1, \dots, y_n, \alpha), \quad i = 1, 2, \dots, n, \quad (1)$$

with a parameter α . A periodic solution with the period T satisfies for all t

$$y_i(t + T) = y_i(t), \quad i = 1, 2, \dots, n. \quad (2)$$

On using transformation $t = T z$ we obtain the system

$$\frac{dy_i}{dz} = T f_i(y_1, \dots, y_n, \alpha), \quad i = 1, 2, \dots, n. \quad (3)$$

The mixed boundary conditions (2) appear now in the form

$$y_i(1) - y_i(0) = 0, \quad i = 1, 2, \dots, n. \quad (4)$$

We use the shooting method to solve the system (3) and (4) for fixed value of the parametr α . We choose n values

$$y_i(0) = x_i, \quad i = 1, 2, \dots, n, \quad (5)$$

and integrate the system (3) on the interval $z \in [0, 1]$, starting from $z = 0$. The value of the period T must be also chosen. As a result of integration we obtain the values of the solution at $z = 1$

$$y_i(1) = g_i(x_1, \dots, x_n, T, \alpha), \quad i = 1, \dots, n, \quad (6)$$

depending on the choice of x_1, \dots, x_n, T (for fixed α). Inserting (6) into boundary conditions (4) we get a system of n equations

$$F_i(x_1, \dots, x_n, T, \alpha) = g_i(x_1, \dots, x_n, T, \alpha) - x_i = 0, \quad i = 1, \dots, n, \quad (7)$$

with $n+1$ unknowns x_1, \dots, x_n, T . To solve this system we need to hold one of the unknowns fixed. It cannot be T , because the solution of (7) exists only for discrete and apriori unknown values of T . Hence, let us fix the value of $x_k = y_k(0)$ for some k . The solution of (7) can be obtained on applying Newton's method for unknowns $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n, T$ (where x_k and α remain fixed in the course of the iteration). We do not obtain the solution, when the fixed value x_k does not lie on the k -th component profile $y_k(z)$ of the wanted periodic solution. See Fig. 1 for an incorrect choice of value x_k (in Fig. 1 denoted x_k^{bad}).

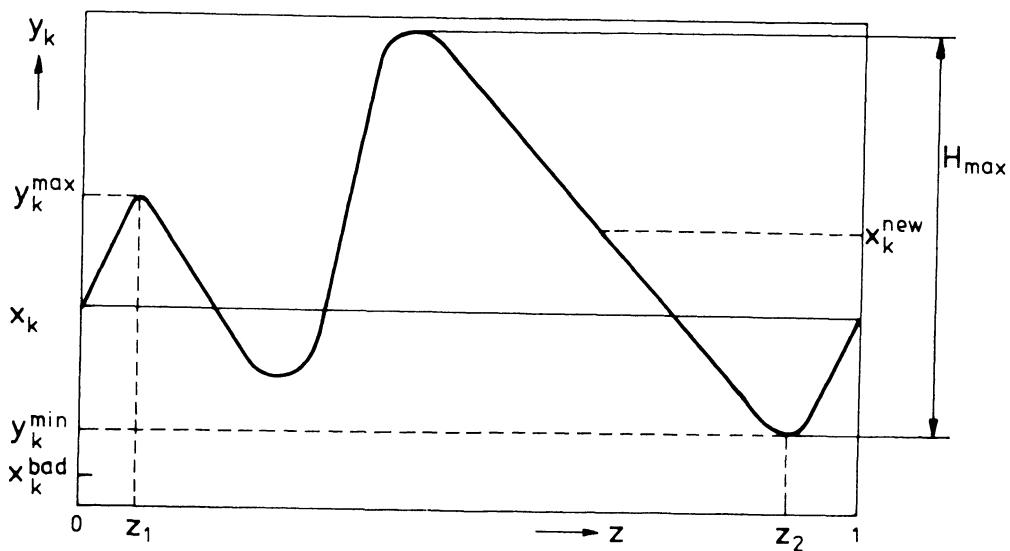


Fig. 1: The choice of fixed value x_k on the periodic orbit - - schematically.

If we now want to continue the periodic solutions depending on a parameter α , we shall continue the solution of the system (7), i.e. the values of unknowns

$x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n, T$ in dependence on the parameter α (x_k remains fixed).

Any algorithm for the computation of the dependence of a solution on a parameter can be used for this purpose [8, 7, 14]. We have used standard routine DERPAR [8] to continue a branch of solutions of n nonlinear equations for $n+1$ variables (n unknowns and one parameter). This algorithm is based on the continuation along the arc-length of the solution locus. It consists of two steps, predictor and corrector; the Newton's method is used as a corrector. A more detailed description is given in [8]. The continuation algorithm requires evaluation of the functions F_i in (7) (by integrating equations (3)) and of the Jacobi matrix

$$\frac{\partial F_i}{\partial x_j}, \quad \frac{\partial F_i}{\partial T}, \quad \frac{\partial F_i}{\partial \alpha}, \quad i = 1, \dots, n, \\ j = 1, \dots, n.$$

These elements can be evaluated by integrating variational differential equations for variational variables

$$p_{ij}(z) = \frac{\partial y_i}{\partial x_j}, \quad q_i(z) = \frac{\partial y_i}{\partial \alpha}. \quad (8)$$

Then

$$\frac{\partial F_i}{\partial x_j} = p_{ij}(1) - \sigma_{ij}, \quad (9a)$$

$$\frac{\partial F_i}{\partial T} = f_i(y(1), \alpha), \quad (9b)$$

$$\frac{\partial F_i}{\partial \alpha} = q_i(1). \quad (9c)$$

The knowledge of the column $\frac{\partial F_i}{\partial x_k}$ is redundant for the continuation algorithm [8]. However, to estimate the stability of a particular periodic solution, it is necessary to have a full matrix $\{\frac{\partial g_i}{\partial x_j}\}$, the monodromy matrix.

Continuation of the solution of the system (7) for variables $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n, T, \alpha$ will take place so long as the chosen value of x_k lies on the profile $y_k(z)$. If the value leaves the profile the algorithm fails. To prevent the failure, we have to change x_k adaptively in the course of the continuation. We must test two conditions for the value of x_k (see Fig. 1):

1. x_k lies "far enough" both from the local maximum y_k^{\max} and the local minimum y_k^{\min} , where y_k^{\max} and y_k^{\min} are maximum and minimum of $y_k(z)$ on the interval $z \in [0, z_1] \cup [z_2, 1]$, where $y_k(z)$ is monotonous.
2. If we consider H_{\max} as a maximum difference between maximum and minimum of the monotonous part of $y_k(z)$ on the entire

interval $z \in [0,1]$ then the interval $[y_k^{\min}, y_k^{\max}]$ must be "large enough" as compared with H_{\max} . A more detailed formulation of these conditions was published in [5].

If both conditions are satisfied, then the value of x_k is not changed. If conditions are not satisfied, then x_k is changed to x_k^{new} (see Fig. 1) in the middle between the absolute maximum and the minimum of $y_k(z)$. The remaining components $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n$ must be reevaluated, too. Some control parameters in the procedure DERNPAR must be reorganized (e.g., direction parameters used to keep direction along the solution locus curve). Schematic flow diagram of the algorithm is presented in Fig. 2.

The stability of the computed periodic solutions can be determined on the basis of characteristic multipliers, e.g. [1], i.e., eigenvalues λ of the monodromy matrix

$$B = \{\frac{\partial g_i}{\partial x_j}\} = \{p_{ij}(1)\}, \quad i = 1, \dots, n, \quad (10) \\ j = 1, \dots, n.$$

The eigenvalues can be computed by means of standard algorithms, see, e.g. [23]. One of the multipliers is always equal to unity. If the others are located inside the unit circle, the periodic solution is stable. If at least one multiplier is located outside the unit circle, the periodic solution is unstable.

The problems with continuation could arise when some multiplier passes through the unit circle, i.e., at the so called bifurcation points. The algorithm continues on the original branch of periodic solutions at the period doubling bifurcation point, at the bifurcation to invariant torus and at the bifurcation point in the solution diagram, e.g., symmetry breaking bifurcation point, or continues on the second branch in the case of limit point in the solution diagram. All bifurcation points can be detected by computing multipliers of periodic solutions during the continuation. The disappearance of periodic solutions at the point of Hopf bifurcation or

at the so called homoclinic point can also be detected by the algorithm.

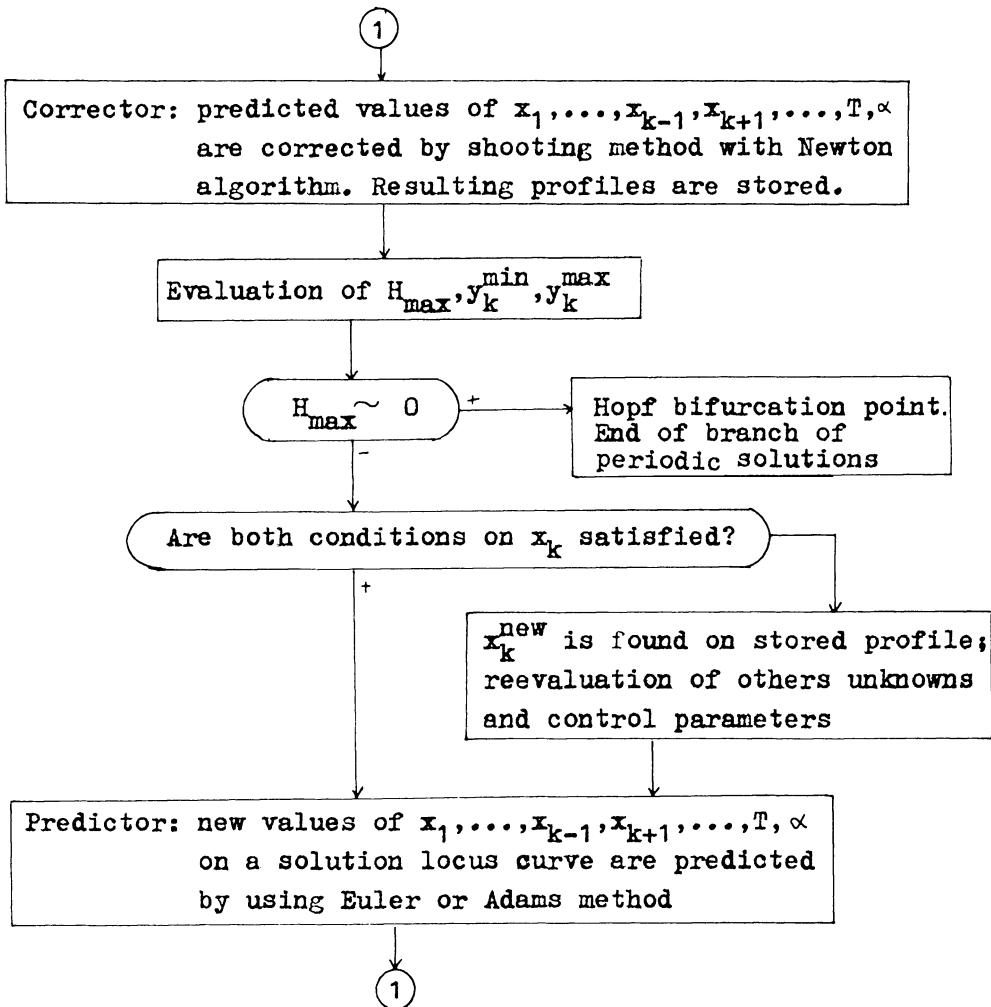


Fig. 2: Schematic flow diagram of the algorithm for continuation of periodic solutions

III. APPLICATION TO THE LORENZ MODEL

The Lorenz model [9] of the flow in the layer of

liquid heated from below is formed by the system of three first-order differential equations ($\cdot = d/dt$)

$$\begin{aligned}\dot{x} &= -\sigma x + \sigma y \\ \dot{y} &= -xz + rx - y \\ \dot{z} &= xy - bz\end{aligned}\tag{11}$$

The system is obtained by a reduction of the system of Navier-Stokes equations and of the equation describing the heat transfer. The dimensionless parameters correspond to:

σ - Prandtl number, $r = R/R_c$ reduced Rayleigh number and b is related to a wavenumber of the convective structure. The model was originally used by Lorenz in the early sixties to demonstrate problems in weather prediction. Particularly in the last ten years it became a subject of number of both computational and analytical studies as the most popular example of existence of a very rich structure of periodic and chaotic solutions.

McLaughlin and Martin [10] summarized derivation of the model equations and studied properties of the solution analytically. Shimizu and Morioka [21,12] have found four regions of existence of periodic solutions, Rössler [16,17,18] studied different types of chaotic solutions in this model. In 1982 Sparrow [22] reviewed the results in Lorenz model simulation. We have used above described continuation algorithm to study the dependence of periodic solutions on the parameter r in [6].

The Lorenz model is symmetric in the following sense: if periodic solution $P=(x,y,z)$ exists, then also solution $\bar{P}=(-x,-y,z)$ exists. Hence P and \bar{P} are either identical (we have single solution which contains symmetry, "S - solution") or P and \bar{P} are different (we have two solutions, "A - solution"). The values of the parameters σ and b were fixed, $\sigma = 16$ and $b = 4$, and r was chosen as a continuation parameter.

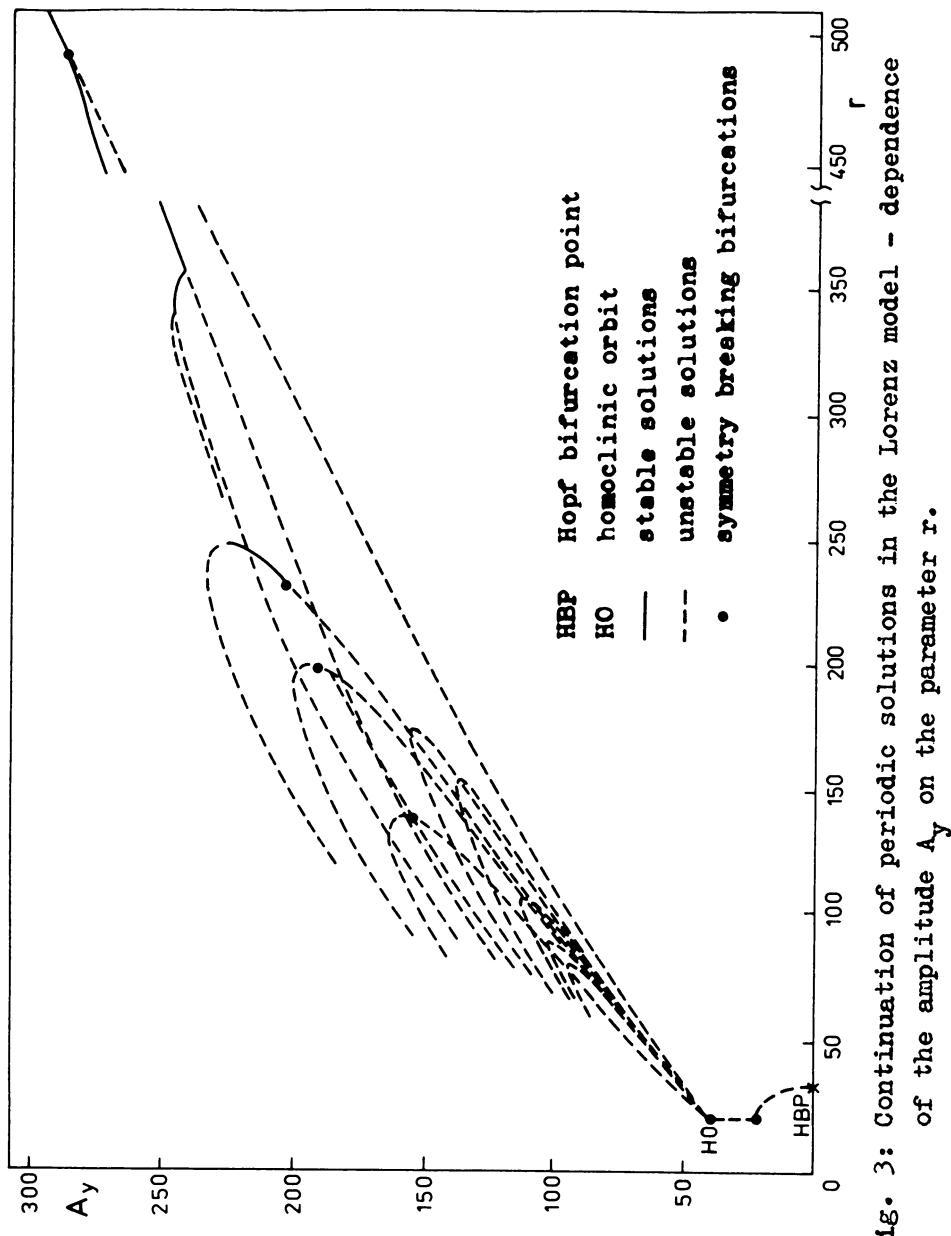


Fig. 3: Continuation of periodic solutions in the Lorenz model – dependence of the amplitude A_y on the parameter r .

The global solution diagram shown in Fig. 3, has been obtained by a successive use of the continuation algorithm described above. It is for the first time that such a full picture of both stable and unstable periodic solutions has been obtained. A_y is the solution amplitude, i.e.

$$A_y = \max_{t \in [0, T]} y(t) - \min_{t \in [0, T]} y(t) \quad (12)$$

We describe some interesting details of this solution diagram. The more complete discussion of the behaviour of periodic solutions with many examples of periodic orbits and also with a "databank" of periodic solutions was published in [6].

At the point of Hopf's bifurcation (HBP) two branches of unstable periodic solutions branch off subcritically. With decreasing value of r both asymmetric orbits touch asymptotically the stationary point ($x=0, y=0, z=0$) and form a homoclinic orbit (denoted HO in Fig. 3). After the contact of both orbits a complex homoclinic orbit is formed and A_y is increased. Many other branches end at this homoclinic orbit, too.

There are many limit points (where two branches of periodic solutions coincide) in the Fig. 3 for $r \in (50, 200)$. A small "window" of stable periodic solutions always exists near the limit point, i.e., one branch starting from the limit point is stable. The stable branch becomes unstable either at the so called symmetry breaking bifurcation point followed by a cascade of period doubling bifurcation points, cf. Fig. 4a, or at the period doubling bifurcation point followed by a cascade of such points, cf. Fig. 4b.

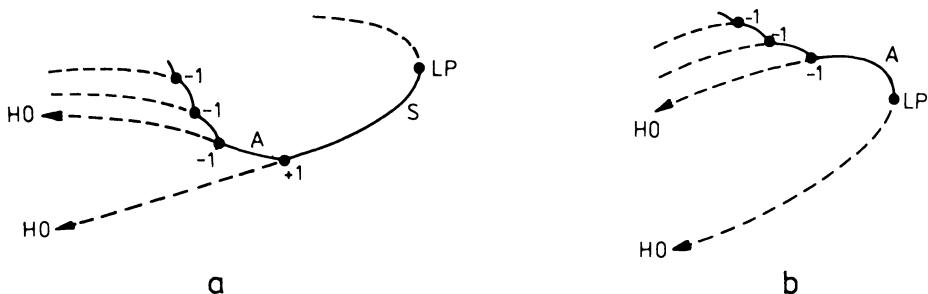


Fig. 4: Schematic picture of symmetry breaking bifurcation (+1) and period doubling bifurcations (-1) from:
 a) stable symmetric solution (S)
 b) stable asymmetric solution (A)
 (LP limit point, for notation see also Fig. 3)

The cascade of branches produced by period doubling is plotted in Fig. 3 for basic branch of periodic solutions ($r \rightarrow \infty$), only. The basic branch is stable and symmetric for $r > 500$; at $r \sim 470$ it loses its stability through symmetry breaking bifurcation. The branches with period T , $\sim 2T$ and $\sim 4T$ are plotted for $r \sim 350$. The situation is similar in the neighbourhood of other limit points (cf. Fig. 4a,b). The bifurcating branches are not included in Fig. 3. The lower branches originating at limit points approach common homoclinic orbit denoted HO in Fig. 3. The upper branches end at different other homoclinic orbits.

IV. CONCLUSION

The algorithm described above was fully tested on Lorenz model and proved its effectiveness as a tool for continuation of the dependence of periodic solutions on a parameter. It was also used for other examples (see, e.g. [5]). From the

starting point (which can be found, e.g., by trial and error technique or by means of dynamic simulation in the case of stable periodic solutions) the algorithm continues the branch of periodic solutions and controls the choice of x_k . The structure of both stable and unstable periodic solutions can be thus followed with low amount of the man-machine interactions.

ACKNOWLEDGEMENTS

Authors would like to thank to the Leibniz Rechenzentrum der Bayerischen Akademie der Wissenschaften for enabling computations on the CYBER 175 and to the donors of the Alexander von Humboldt Stiftung for financial support to one of them (Martin Holodniok) during the computations. The authors would like to express their thanks also to Deutscher Akademischer Austauschdienst (DAAD) for financial support during the conference. We would like to thank to prof. Miloš Marek for very useful discussion and help with the preparation of the manuscript.

LITERATURE

- [1] E.A.Coddington, N. Levinson: Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955
- [2] A.I. Chibnik: Periodic solutions of the system of n differential equations, Preprint of the Institute of biological investigations of the Academy of Sciences, Puschino (USSR), 1979 (in Russian)
- [3] L.O. Chua, P.M. Lin: Computer Aided Analysis of Electronic Circuits: Algorithms and Computational Techniques, Prentice-Hall, Englewood Cliffs, 1975
- [4] B. Hassard: Bifurcation of periodic solutions of the Hodgkin-Huxley model for the squid giant axon, J. Theor.Biol. 71 (1978), 401
- [5] M. Holodniok, M. Kubíček: DERPER - An algorithm for continuation of periodic solutions in ordinary differential equations, J. Comput. Physics, in press
- [6] M. Holodniok, M. Kubíček, M. Marek: Stable and unstable periodic solutions in the Lorenz model. Preprints of the Math. Inst., Technical University München, G.F.R., 1982

- [7] H.B. Keller: Numerical solution of bifurcation and nonlinear eigenvalue problems, in "Applications of Bifurcation Theory", P. Rabinowitz, ed., Academic Press, New York 1977, p. 359
- [8] M. Kubíček: Algorithm 502: Dependence of solution of nonlinear systems on a parameter, ACM Trans. on Math. Software 2 (1976), 98
- [9] E.N. Lorenz: Deterministic nonperiodic flow, J. Atmos. Sci. 20 (1963), 130
- [10] J.B. McLaughlin, P.C. Martin: Transition to turbulence in a statically stressed fluid system, Phys. Rev A 12 (1975), 186
- [11] R.M. Mehra, J.V. Carroll: Bifurcation analysis of aircraft high angle-of-attack flight dynamics, Preprints of the AIAA Atmospheric Flight Mechanics Conference, AIAA, New York, 1980, p. 358
- [12] N. Morioka, T. Shimizu: Transition between turbulent and periodic states in the Lorenz model, Physics Letters, 66A (1978), 447
- [13] G. Nicolis, I. Prigogine: Self-Organization in Non-Equilibrium Systems, J. Wiley, New York, 1977
- [14] W.C. Rheinboldt: Solution fields of nonlinear equations and continuation methods, SIAM J. Num. Anal. 17 (1980), 221
W.C. Rheinboldt: Numerical analysis of continuation methods for nonlinear structural problems, Computers and Structures 13 (1981), 103
W.C. Rheinboldt, J.V. Burkardt: A program for a locally-parametrized continuation process, Tech. Rept. ICMA-81-30, Institute for Computational Mathematics and Applications, University of Pittsburgh, 1981
- [15] J. Rinzel, R.N. Miller: Numerical calculation of stable and unstable periodic solutions to the Hodgkin-Huxley equations, Math. Biosciences 49 (1980), 27
- [16] O.E. Rössler: Different types of chaos in two simple differential equations, Z.f. Naturforsch. 31 (1976), 1664
- [17] O.E. Rössler: An equation for continuous chaos, Physics Letters 57 A (1976), 397
- [18] O.E. Rössler: Horseshoe-map chaos in the Lorenz equation, Physics Letters, 60A (1977), 392
- [19] I. Schreiber, M. Holodniok, M. Kubíček, M. Marek: Periodic phenomena in coupled cells, in preparation
- [20] R. Seydel: Numerical computation of periodic orbits that bifurcate from stationary solutions of ordinary differential equations, Appl. Math. Comp. 9 (1981), 257
- [21] T. Shimizu, N. Morioka: Chaos and limit cycles in the Lorenz model, Physics Letters, 66A (1978), 182

- [22] C. Sparrow : The Lorenz Equations: Bifurcations, Chaos, and Strange Attractors, Applied Mathematical Sciences 41, Springer Verlag, Berlin, 1982
- [23] J.H. Wilkinson, C. Reinsch: Handbook for Automatic Computation II, Linear Algebra, Springer Verlag, Berlin, 1971

Computer Center and Department of Chemical Engineering,
Prague Institute of Chemical Technology,
166 28 Praha 6, Czechoslovakia

Singular Points and their Computation

A. D. Jepson* and A. Spence

1. Introduction. The equilibria of many physical systems can be modelled by nonlinear multi-parameter equations of the form

$$f(\mathbf{x}, \lambda, \alpha) = 0, \quad f: \mathbb{R}^n \times \mathbb{R} \times \mathbb{R}^p \rightarrow \mathbb{R}^n \quad (1.1)$$

(see [12]). Here f is a smooth function, $\mathbf{x} \in \mathbb{R}^n$ is the state variable, $\lambda \in \mathbb{R}$ is the bifurcation parameter, and $\alpha \in \mathbb{R}^p$ is a vector of control parameters. The distinction of the bifurcation parameter λ from the control parameters α is made for two reasons. Firstly, it is often the case in experiments that one parameter is varied quasi-statically while the other parameters are held fixed. Secondly, the distinction of λ is numerically convenient and once the results are obtained they can be readily reinterpreted in terms of other choices of the bifurcation parameter.

The bifurcation diagram for (1.1) at $\alpha = \alpha_0$ is the graph of solutions for

$$f(\mathbf{x}, \lambda, \alpha_0) = 0. \quad (1.2)$$

Efficient continuation techniques have recently been developed to compute bifurcation diagrams for problems of the form (1.2) with α_0 given (see [9], [11]). Two limitations of these techniques are: (i) they require a starting point on each connected component of the bifurcation diagram; and (ii) they are inefficient when applied to multi-parameter problems of the form (1.1). Here and in [6] we consider an approach which is designed to alleviate these two difficulties. Our approach is based on combining the above mentioned continuation methods with the singularity theory of Golubitsky and Schaeffer [5].

The following example, which was first given in [4], helps to illustrate several important concepts. We consider

$$f(\mathbf{x}, \lambda, \alpha) := \mathbf{x}^3 + \lambda^2 + \alpha_1 + \alpha_2 \mathbf{x} + \alpha_3 \mathbf{x} \lambda = 0, \quad (1.3)$$

with $\mathbf{x} \in \mathbb{R}$ (i.e. $n=1, p=3$ in (1.1)). In Figure 1 the varieties B and H are shown to divide the control parameter space for $\alpha_3 > 0$ into five open regions, labelled A through E. These varieties are given by

$$B = \{\alpha \in \mathbb{R}^3 : f(\mathbf{x}, \lambda, \alpha) = f_{\mathbf{x}} = f_{\lambda} = 0\}, \quad (1.4a)$$

$$H = \{\alpha \in \mathbb{R}^3 : f(\mathbf{x}, \lambda, \alpha) = f_{\mathbf{x}} = f_{\mathbf{x}\mathbf{x}} = 0\}, \quad (1.4b)$$

*Research supported by the University of Toronto and NSERC Canada Grant 3-643-126-70, with travel funding from the British Council Academic Links and Interchange Scheme.

and they are even about $\alpha_3 = 0$ (see [4]). The qualitative shape of the bifurcation diagrams of (1.3) is the same for all values of the control parameters α in any one region A, B, C, D, or E. These shapes are sketched in Figure 1, for example for any $\alpha \in C$ the bifurcation diagram for (1.3) is an S-shaped curve with precisely two quadratic turning points (see [12] for the definition). The varieties B and H are themselves divided into open regions by the two paths from 0 which pass through the points 2 and 6 respectively. Moreover B is divided further by the path passing through both 0 and 3. The bifurcation diagrams for (1.3) are qualitatively similar in each of the open regions of B and H , and on each of the above mentioned paths (after the point 0 is removed).

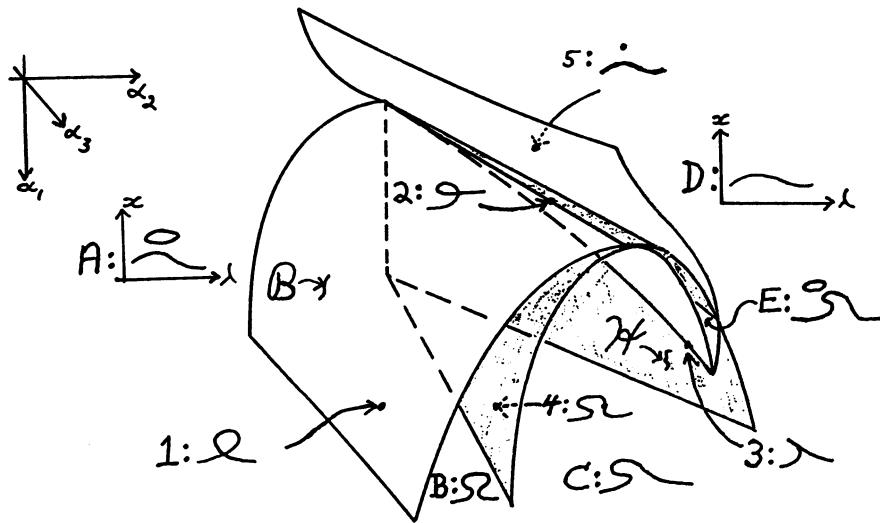


Figure 1.

The intuitive notion of qualitatively similar bifurcation diagrams is made precise in [6]. Roughly speaking, two diagrams are said to be qualitatively similar if they have the same number, type, and arrangement of the singularities defined in [5]. For our purposes here it is sufficient to note that the diagrams given in Figure 1 are all qualitatively different.

The two major objectives of our approach are: (i) to completely decompose the control parameter space \mathbb{R}^p into regions for which the bifurcation diagrams of (1.1) are qualitatively similar (see Fig. 1), and (ii) provide initial points on each connected component of the bifurcation diagram at any given $\alpha \in \mathbb{R}^p$.

For basic definitions of some of the terminology used in this paper we refer the reader to [12]. Here, in Sections 2 and 3, we briefly discuss the results obtained in [6] for the scalar case of (1.1), that is $n = 1$, $x \in \mathbb{R}$. In Section 4 we show how the methods for scalar problems can be generalized so that

the resulting techniques apply to (1.1) for $n \geq 1$. Finally, in Section 5 we consider a specific case of our general approach, and compare our results with those given in [12].

2. Singularity Theory. In order to attain the first objective given above it is clear from Figure 1 that a computational method is needed both for the calculation of the varieties B and H , and for distinguishing any exceptional points, like 2 or 3, on these varieties. In this section, and in Section 3, we summarize the technique suggested in [6] for the following scalar case of (1.1):

$$f(\mathbf{x}, \lambda, \alpha) = 0; f : \mathbf{R} \times \mathbf{R} \times \mathbf{R}^p \rightarrow \mathbf{R}. \quad (2.1)$$

Then in Sections 4 and 5 we discuss generalizations of these results for the vector case (1.1).

In [5] the notion of contact equivalence is used to classify the possible singularities of (1.1) into equivalence classes. Here we only consider scalar problems, although the ideas generalize in a straightforward manner to the vector case. Let $h, g : \mathbf{R} \times \mathbf{R} \rightarrow \mathbf{R}$ be two smooth functions and suppose that $h(\mathbf{x}_0, \lambda_0) = g(0, 0) = 0$. The two bifurcation problems

$$h(\mathbf{x}, \lambda) = 0, \quad (2.2a)$$

$$g(\mathbf{x}, \lambda) = 0, \quad (2.2b)$$

are said to be *contact equivalent* at $(\mathbf{x}_0, \lambda_0)$ and 0, respectively, if

$$g(\mathbf{x}, \lambda) = T(\mathbf{x}, \lambda) h(X(\mathbf{x}, \lambda), \Lambda(\lambda)) \quad (2.3a)$$

for (\mathbf{x}, λ) near 0. Here T, X, Λ are smooth functions with $X(0) = \mathbf{x}_0$, $\Lambda(0) = \lambda_0$, and

$$T(0) \neq 0, \frac{\partial X}{\partial \mathbf{x}}(0) \neq 0, \frac{\partial \Lambda}{\partial \lambda}(0) \neq 0. \quad (2.3b)$$

(It is convenient to allow X and Λ to be reflections, unlike the definition given in [5].) The intuitive notion of qualitatively similar *local* behaviour of the two bifurcation problems (2.2a,b) is made precise through this definition of contact equivalence. The resulting equivalence classes are referred to as "singularities".

The concept of the *stability* of a singularity is given in

Definition 2.4. A singularity $(\mathbf{x}_0, \lambda_0, \alpha_0)$ of (2.1) (with $\alpha = \alpha_0$ fixed) is said to be *structurally stable* if for any smooth g , for any neighborhood V of $(\mathbf{x}_0, \lambda_0, \alpha_0)$, and for any ε with $|\varepsilon|$ sufficiently small, the perturbed problem

$$h(\mathbf{x}, \lambda, \alpha; \varepsilon) := f(\mathbf{x}, \lambda, \alpha) + \varepsilon g(\mathbf{x}, \lambda, \alpha) = 0 \quad (2.4)$$

has a singularity of the same type at some point $(\mathbf{x}(\varepsilon), \lambda(\varepsilon), \alpha(\varepsilon)) \in V$ (i.e. $h(\mathbf{x}, \lambda, \alpha; \varepsilon)$ is contact equivalent to f for ε sufficiently small). If in addition

$$(x(\varepsilon), \lambda(\varepsilon), \alpha(\varepsilon)) = (x_0, \lambda_0, \alpha_0) + O(\varepsilon)$$

for each g , then the singularity is said to be (*structurally*) *linearly stable*. It is well established in the literature that only structurally stable behaviour should be considered in physical models.

The concept of a *universal unfolding* is given in

Definition 2.5. Let h be as in (2.2a) with a singular point (x_0, λ_0) . Suppose $g : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p \rightarrow \mathbb{R}$ is a smooth function such that

$$h(x, \lambda) = g(x, \lambda, \beta_0)$$

for (x, λ) near (x_0, λ_0) . Then $g(x, \lambda, \beta)$ is called a p -parameter *unfolding* of h . Let $f(x, \lambda, \alpha)$, for α near α_0 , be a second unfolding of h . Then g is said to *factor through* f if

$$g(x, \lambda, \beta) = T(x, \lambda, \beta) f(X(x, \lambda, \beta), \Lambda(\lambda, \beta), \alpha(\beta)), \quad (2.5)$$

with T, X, Λ, α smooth functions, $\alpha(\beta_0) = \alpha_0$, $T(x, \lambda, \beta_0) = \pm 1$, $X(x, \lambda, \beta_0) = \pm x$, $\Lambda(\lambda, \beta_0) = \pm \lambda$, and for fixed β equation (2.5) is a contact equivalence relation (see (2.4)). Finally, f is a *universal unfolding* of h if every unfolding of h factors through f . It is important to notice (by 2.5) that if f is a universal unfolding of h then *any* small perturbation of h can be written in terms of f . Intuitively, f includes *all* possible small perturbations of h .

In [5] a number, called the *codimension* of the singularity, is associated with each equivalence class. The codimension has the important property that it is the *minimum* number of control parameters, α , needed to make a singularity stable. In Figure 2 all possible singularities of (2.1) having codimension $q \leq 3$ are arranged in a hierarchy. (Here we ignore composite singular points such as the double limit points considered in [5].) The (q, j) -singularity has codimension q , and is defined to be the singularity which the polynomial in the (q, j) -node of the hierarchy has at $(x, \lambda) = 0$. The $(0, 1)$ -node is called the *root* and represents the quadratic turning point singularity (see [12]). We use the word *branch* to denote the link between two nodes.

This hierarchy was introduced in [6]; it allows a straightforward, unified explanation of our technique for solving (2.1). In particular, appropriate defining conditions for the (q, j) -singularity can be obtained using the hierarchy by consideration of a path from the root to the (q, j) -node. Proceeding down from the root we keep j constant whenever there is a choice (this avoids unnecessary use of polynomials of derivatives of f such as D_2 , which would make the following Theorem 2.8 fail). Once this path has been determined we require that the label on each branch in the path vanishes at (x, λ, α) . In addition we require that (2.1) is satisfied and $f_x = 0$. The resulting

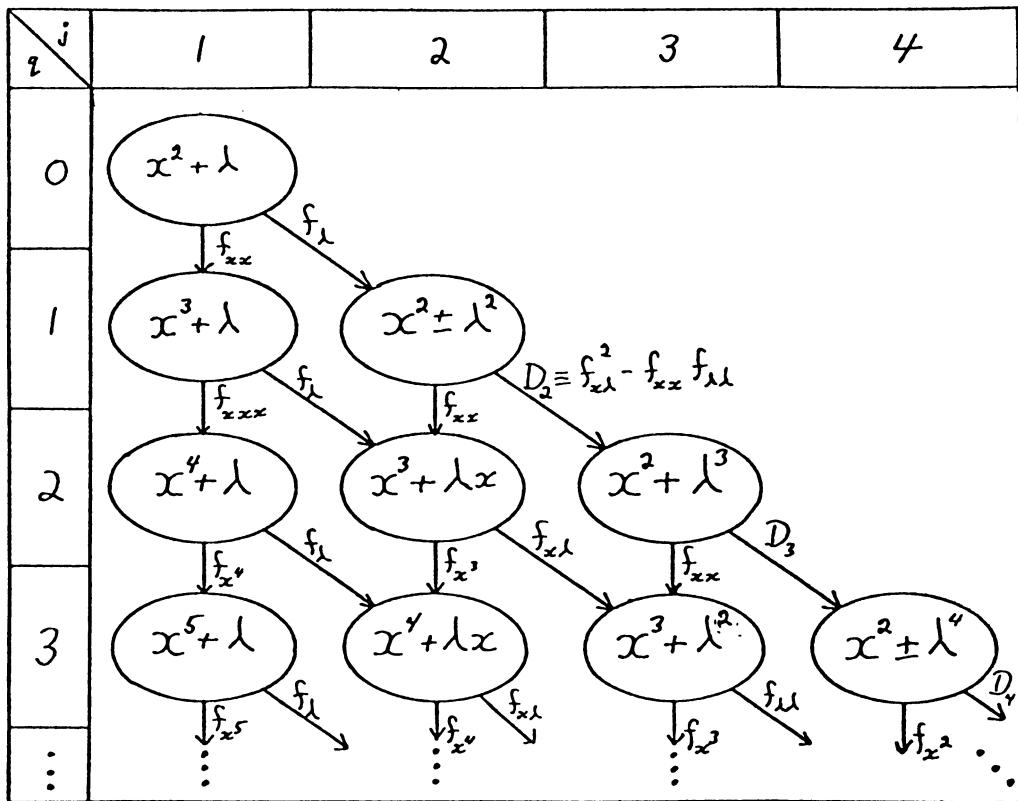


Figure 2.

equations form the (q,j) -extended system. For example, the extended system for a pitchfork bifurcation singularity, $(q,j) = (2,2)$, is

$$F_{2,2}(x, \lambda, \alpha) := (f, f_x, f_{xx}, f_\lambda) = 0 \quad (2.6a)$$

(see also [5]). Furthermore, nondegeneracy conditions called *side-constraints* are obtained for the (q,j) -singularity by requiring that the labels on the branches leaving the (q,j) -node are nonzero. Thus for the $(2,2)$ -node we have

$$C_{2,2}^1 := f_{xxx} \neq 0; \quad C_{2,2}^2 := f_{x\lambda} \neq 0. \quad (2.6b)$$

Similar defining conditions for several types of singularities are also given in [1], [2], [4], [5], [8], and [10].

The extended systems obtained from the hierarchy have three desirable properties, as described in the following three theorems.

Theorem 2.7. Let $0 \leq q \leq 3$, and suppose that

$$F_{q,j}(\mathbf{x}, \lambda, \alpha) = 0 \in \mathbf{R}^{q+2}, \quad C_{q,j}^k(\mathbf{x}, \lambda, \alpha) \neq 0, \quad k = 1, 2, \quad (2.7)$$

are the (q,j) -extended system and side constraints defined above. Then $(\mathbf{x}_0, \lambda_0, \alpha_0)$ satisfies (2.7) if and only if $(\mathbf{x}_0, \lambda_0, \alpha_0)$ is a (q,j) -singular point of (2.1). (That is the extended systems and side constraints are specific for particular singularities.)

Theorem 2.8. Let $(\mathbf{x}_0, \lambda_0, \alpha_0)$ satisfy (2.7). Then the following three statements are equivalent:

- i) $\text{rank } \frac{\partial F_{q,j}}{\partial (\mathbf{x}, \lambda, \alpha)}(\mathbf{x}_0, \lambda_0, \alpha_0) = q + 2$ (i.e. full);
- ii) $(\mathbf{x}_0, \lambda_0, \alpha_0)$ is a (structurally) linearly stable singularity of (2.1);
- iii) $f(\mathbf{x}, \lambda, \alpha)$ is a universal unfolding of the singularity at $(\mathbf{x}_0, \lambda_0, \alpha_0)$.

(Hence the extended systems have isolated roots whenever the singularity is linearly stable and $p = q$.)

Theorem 2.9. Let $1 \leq q \leq 3$ and suppose $(\mathbf{x}_0, \lambda_0, \alpha_0)$ is a linearly stable (q,j) -singularity of (2.1). Let $j' = j$ or $j-1$ be chosen such that $(q-1, j')$ is a node in the hierarchy directly above the (q,j) -node. Then $(\mathbf{x}_0, \lambda_0, \alpha_0)$ is a solution of

$$F_{q-1,j'}(\mathbf{x}, \lambda, \alpha) = 0 \in \mathbf{R}^{q+1} \quad (2.10a)$$

with

$$\text{rank} \left[\frac{\partial F_{q-1,j'}}{\partial (\mathbf{x}, \lambda, \alpha)}(\mathbf{x}_0, \lambda_0, \alpha_0) \right] = q + 1 \quad (\text{i.e. full}). \quad (2.10b)$$

The significance of Theorem 2.9 is explained below. Theorems 2.7 and 2.9 are proved in [6], while Theorem 2.8 is an easy consequence of Theorems 3.5 and 3.10 in [6].

3. Numerical Approach. Recall the two objectives listed in Section 1, namely, to completely decompose the control parameter space, and to obtain initial points for particular bifurcation diagrams. In [6] a three step approach is introduced and applied to a nontrivial scalar model of a stirred tank chemical reactor. The model involves four control parameters ($p=4$ in (2.1)), yet both objectives are attained. Here we present a brief summary of this three step approach; we use (1.3) to illustrate the procedure.

Step 1. Descending the Hierarchy. We begin by attempting to locate the points of highest codimension in a given problem (2.1). To do this we iterate down the hierarchy, as described below.

Suppose that, at some stage in the descent, we know a solution $(\mathbf{x}_1, \lambda_1, \alpha_1)$ of the (q,j) -extended system

$$F_{q,j}(\mathbf{z}, \boldsymbol{\beta}; \gamma) = 0 \in R^{q+2}. \quad (3.1)$$

Here $\mathbf{z} = (\mathbf{x}, \lambda, \alpha) \in R^{q+2}$, $\alpha = (\alpha_1, \alpha_2, \gamma)$ (say), with $\alpha \in R^q$ called the *unfolding* parameters, $\beta \in R$, and $\gamma \in R^{p-q-1}$ called the *slice* parameters. For the slice parameters held constant, problem (1.3) can be viewed as a one parameter bifurcation problem with the state variable \mathbf{z} and the bifurcation parameter β . From Theorem 2.8 we see that $(\mathbf{z}_1, \lambda_1, \alpha_1)$ corresponds to a *regular* point of (3.1) if it is a (q,j) -singular point that is linearly stable with (β, γ) held constant. Paths of such points can be computed using a standard numerical continuation method (see [9],[11]). It is important to note that Theorem 2.9 shows that the extended systems are also well behaved at codimension $q+1$ singularities. In particular, generically (3.1) has only simple turning points (see [12]). This aspect is discussed further in Step 2, below.

Singular points of higher codimension, on this computed path, can be found by locating points at which a side constraint, $C_{q,j}^k$, vanishes. Given a $(q+1,j')$ -singularity located in this way, a slice parameter γ_i can be allowed to vary and the $(q+1,j')$ -extended system used to compute paths of $(q+1,j')$ -singular points. In this manner we attempt to descend the hierarchy.

For example, suppose a $(0,1)$ -singularity was known for problem (1.3) (say with α at C in Figure 1). Then we could compute a path of $(0,1)$ -singularities by holding α_1 and α_2 constant in the $(0,1)$ -extended system. The side constraint $C_{0,1}^1$ changes sign as the H -variety is crossed, and therefore it is a simple matter to locate a point such as 4, in Figure 1, on this path. This point provides an initial (regular) solution of the $(1,1)$ -extended system. By freeing α_1 (say) a path in the H -variety can be computed. Similarly, points in the path 0-2 of Figure 1 can be found by detecting a sign change in $C_{1,1}^2$, and then the $(2,2)$ -extended system can be used to compute the entire path. Finally, the $(3,3)$ -singularity at 0 on this path is detected via a sign change in $C_{3,3}^2$ (see Figure 2). A very similar descent sequence was used for the chemical reactor model in [6].

Step 2. Ascending the Hierarchy. Given a (q,j) -singular point $(\mathbf{z}_0, \beta_0, \gamma_0)$ that is a regular point of (3.1) (the generic case), this point is necessarily a solution of

$$F_{q-1,j}(\mathbf{z}', \boldsymbol{\beta}'; \gamma') = 0 \in R^{q+1}, \quad (3.2)$$

for $j' = j$ or $j-1$ (see Figure 2). Here $(\mathbf{z}', \boldsymbol{\beta}') = \mathbf{z}$, and $\gamma' = (\beta, \gamma)$ (i.e. we include β in the frozen slice parameters). Then Theorem 2.9 shows that, for γ' fixed, problem (3.2) has at worst a simple turning point (see [12]) at the (q,j) -singular point. Therefore standard arclength-like continuation techniques can be used to compute the solution branch of (3.2) near $(\mathbf{z}_0, \beta_0, \gamma_0)$ (see [9], or [11]).

For example, having found the point 0 in Figure 1, the branch 0-3

consisting of (2,3)-singularities could be computed. Furthermore, by fixing α_3 , paths on the B - or H -variety could be obtained from the (1,2)- or (1,1)-extended system, respectively. In this manner the control parameter space can be completely decomposed; the signs of the side constraints can be used to reconstruct the qualitative form of the bifurcation diagrams in each region.

Step 3. Obtaining Initial Points. Finally by unfolding the codimension 1 singularities computed in Step 2, we can obtain all the quadratic turning points of (1.3) at any given α_0 . This provides initial points on each component of the bifurcation diagram of (1.3) at α_0 which have at least one singular point. Points on any remaining component could be obtained by following paths of regular points of (1.3) from B or H .

4. Extended Systems for Vector Problems. In [12] appropriate extended systems for the calculation of singular points of the problem

$$f(x, \lambda, \alpha) = 0, f: R^n \times R^p \times R^p \rightarrow R^n, \quad (4.1)$$

are constructed. There it is indicated how the approach of Crandall and Rabinowitz [3] can be used to obtain defining conditions for various types of singularity. These conditions are in turn used to construct the extended systems. It is easily shown that these extended systems simplify to the corresponding systems obtained in Section 2 when n , the dimension of f , is one.

Here we show how the results of Section 2 can be used to construct extended systems for singular points of the vector problem (4.1). The systems given in [12] can be recovered, but more importantly this approach produces both some interesting alternative systems and extended systems for higher order singularities. Furthermore the important results on the isolatedness of roots follow in a natural manner from the analysis of the scalar case given in Section 2.

The key step in our approach is the development of a numerically convenient form of the Liapunov-Schmidt decomposition. We begin our analysis with a slight generalization of the standard L-S procedure. Suppose $(x_0, \lambda_0, \alpha_0)$ is a singular point of (4.1) with

$$\text{Null}(f_x^0) = \text{Span}\{\varphi_1\} \neq \{0\}, \quad (4.2a)$$

$$\text{Null}([f_x^0]^T) = \text{Span}\{\psi_1\} \neq \{0\}. \quad (4.2b)$$

Let $v_1, w_1 \in R^n$ satisfy

$$w_1^T \psi_1 \neq 0 \neq v_1^T \varphi_1, \quad (4.3a)$$

and let $\{v_1, v_2, \dots, v_n\}, \{w_1, w_2, \dots, w_n\}$ be bases for R^n with

$$\varphi_1 \notin \text{Span}\{v_2, \dots, v_n\}, \psi_1 \notin \text{Span}\{w_2, \dots, w_n\}. \quad (4.3b)$$

Define the rectangular matrices V_2 , W_2 by

$$V_2 := (v_2, \dots, v_n), \quad W_2 := (w_2, \dots, w_n)^T. \quad (4.4)$$

Now we can write the Jacobian f_x^0 as

$$f_x^0 = \begin{bmatrix} W_2 \\ w_1^T \end{bmatrix}^{-1} \begin{bmatrix} A & b \\ c^T & d \end{bmatrix} \begin{bmatrix} V_2 & v_1 \end{bmatrix}^{-1} \quad (4.5)$$

Here A is a $(n-1) \times (n-1)$ matrix, $b, c \in R^{n-1}$, and $d \in R$. Furthermore we have

Lemma 4.6. A is nonsingular.

Proof. The result follows easily from (4.2, 3, 4, 5).

The reduction of (4.1) to a scalar equation now proceeds as follows. First we rewrite (4.1) as

$$\begin{bmatrix} \hat{f}(v, \varepsilon, \lambda, \alpha) \\ g(v, \varepsilon, \lambda, \alpha) \end{bmatrix} = \begin{bmatrix} W_2 \\ w_1^T \end{bmatrix} f(x(v, \varepsilon), \lambda, \alpha) = 0, \quad (4.7a)$$

with

$$x(v, \varepsilon) = x_0 + \begin{bmatrix} V_2 & v_1 \end{bmatrix} \begin{bmatrix} v \\ \varepsilon \end{bmatrix}, \quad v \in R^{n-1}, \quad \varepsilon \in R. \quad (4.7b)$$

The chain rule implies

$$\hat{f}_v^0 := \hat{f}_v(0, 0, \lambda_0, \alpha_0) = A,$$

i.e. \hat{f}_v^0 is nonsingular. Therefore the Implicit Function Theorem implies that

$$\hat{f}(v, \varepsilon, \lambda, \alpha) = 0 \in R^{n-1} \quad (4.8)$$

has a unique (smooth) solution $v(\varepsilon, \lambda, \alpha)$ for $(\varepsilon, \lambda, \alpha)$ near $(0, \lambda_0, \alpha_0)$, with $v(0, \lambda_0, \alpha_0) = 0$. Finally we see that (4.7a) is equivalent to the scalar equation

$$h_0(\varepsilon, \lambda, \alpha) := g(v(\varepsilon, \lambda, \alpha), \varepsilon, \lambda, \alpha) = 0, \quad (4.9)$$

for $(\varepsilon, \lambda, \alpha)$ near $(0, \lambda_0, \alpha_0)$.

In the standard Liapunov-Schmidt reduction the two bases used in (4.7) are chosen to satisfy

$$v_1 = \varphi_1; \quad w_1 = \psi_1, \quad \psi_1^T w_i = 0 \quad \text{for } i = 2, \dots, n \quad (4.10)$$

(see [13]). The fact that the local behaviour of $h_0(\varepsilon, \lambda, \alpha)$ does not depend on the particular choice of bases is the content of

Theorem 4.11. Let $h_0(\varepsilon, \lambda, \alpha)$ and $\hat{h}_0(\varepsilon, \lambda, \alpha)$ be two functions obtained from (4.1) by using the above reduction procedure with bases $\{v_i\}_{i=1}^n$, $\{w_i\}$ and $\{\hat{w}_i\}$, $\{\hat{w}_i\}$ respectively. Then the bifurcation problems

$$h_0(\varepsilon, \lambda, \alpha) = 0 \text{ and } \hat{h}_0(\varepsilon, \lambda, \alpha) = 0 \quad (4.11)$$

are contact equivalent at $(0, \lambda_0, \alpha_0)$.

The proof of Theorem 4.11 is a straightforward but lengthy calculation based on the application of Lemma 3.8 on p.41 of [4]. We discuss this in more detail in [7]. As a consequence of this theorem the definition of a (q,j) -singularity of the vector problem (4.1) can be given in terms of the singular behaviour of the *reduced problem*. In this manner we carry over the definitions of singularity types, stability, and universal unfoldings from Section 2.

A numerical method for vector problems could be based directly on this reduction process. That is, the extended systems given in Section 2 could be used on the scalar problem (4.9). To evaluate h_0 at a given $(\varepsilon, \lambda, \alpha)$ the non-linear equation (4.8) would have to be solved for v . This leads to two nested iterations which might prove inefficient, and hence we do not discuss such techniques here. Instead we discuss a technique which, in some sense, solves (4.8) and the extended system for $h_0(\varepsilon, \lambda, \alpha)$ simultaneously. To do this we introduce $h(\varepsilon, \lambda, \alpha, c)$, an extension of $h_0(\varepsilon, \lambda, \alpha)$, defined in a neighbourhood of $(x_0, \lambda_0, \alpha_0)$.

The Implicit Function Theorem provides the existence of a smooth function $u(\varepsilon, \lambda, \alpha, c)$ which satisfies

$$\hat{f}(u, \varepsilon, \lambda, \alpha) = c \in R^{n-1}, \quad (4.12)$$

for $(\varepsilon, \lambda, \alpha, c)$ near $(0, \lambda_0, \alpha_0, 0)$. From (4.8) we see that

$$u(\varepsilon, \lambda, \alpha, 0) = v(\varepsilon, \lambda, \alpha), \quad (4.13)$$

and therefore u is an extension of $v(\varepsilon, \lambda, \alpha)$. Moreover, we define

$$h(\varepsilon, \lambda, \alpha, c) := g(u(\varepsilon, \lambda, \alpha, c), \varepsilon, \lambda, \alpha), \quad (4.14)$$

with g given by (4.7), and find

$$h(\varepsilon, \lambda, \alpha, 0) = h_0(\varepsilon, \lambda, \alpha) \quad (4.15)$$

for $(\varepsilon, \lambda, \alpha)$ near $(0, \lambda_0, \alpha_0)$.

It is important to note that h can be evaluated without solving (4.12) for u . In particular, given $(x_1, \lambda_1, \alpha_1)$ near $(x_0, \lambda_0, \alpha_0)$, (4.7a) can be used to calculate \hat{f} and g with $x(v, \varepsilon) = x_1$. Then equations (4.12–14) provide c and h respectively. This is in contrast to the evaluation of h_0 , which requires the solution of (4.8). Furthermore the extended systems defined below can be evaluated without the need for any iterations.

An extended system for a (q, j) -singularity of (4.1), with $0 \leq q \leq 3$, is defined by

$$H_{q,j}(x, \lambda, \alpha) := \begin{bmatrix} \hat{f}(v, \varepsilon, \lambda, \alpha) \\ F_{q,j}(\varepsilon, \lambda, \alpha, c) \end{bmatrix} = 0. \quad (4.16)$$

Here

$$x = x_0 + \begin{bmatrix} V_2 & v_1 \end{bmatrix} \begin{bmatrix} v \\ \varepsilon \end{bmatrix}; \quad c = \hat{f}(v, \varepsilon, \lambda, \alpha). \quad (4.17)$$

Also, $F_{q,j}$ is the (q, j) -extended system defined in Section 2 applied to the scalar problem

$$h(\varepsilon, \lambda, \alpha, c) = 0, \quad (4.18)$$

with ε treated as the state variable (i.e. "x" in Section 2), and c as additional control parameters. Similarly the side constraints,

$$C_{q,j}^k(\varepsilon, \lambda, \alpha, c) \neq 0, \quad k = 1, 2, \quad (4.19)$$

are the constraints given in Section 2 applied to (4.18). For example,

i) $H_{0,1} := (\hat{f}^T, h, h_\varepsilon) = 0, \quad C_{0,1}^1 := h_{\varepsilon\varepsilon} \neq 0 \neq h_{\varepsilon\lambda} := C_{0,1}^2.$

ii) $H_{2,2} := (\hat{f}^T, h, h_\varepsilon, h_{\varepsilon\varepsilon}, h_\lambda) = 0, \quad C_{2,2}^1 := h_{\varepsilon\varepsilon\varepsilon\varepsilon} \neq 0 \neq h_{\varepsilon\lambda} := C_{2,2}^2.$

In Section 5 we discuss the numerical solution of $H_{0,1}$ for a specific choice of v_1 and w_1 .

We end this section by showing that the extended systems defined above have the three desirable properties discussed in Sections 2 and 3. In particular, we begin with

Theorem 4.20. Let $0 \leq q \leq 3$. The point $(x_0, \lambda_0, \alpha_0)$ is a (q, j) -singularity of (4.1) if and only if (4.2, 16, and 19) are satisfied.

Theorem 4.21. Let $p = q \leq 3$, and suppose $(x_0, \lambda_0, \alpha_0)$ is a (q, j) -singularity of (4.1). Then

$$\text{Rank} \left(\frac{\partial H_{q,j}^0}{\partial z} \right) = n + q + 1 \quad (\text{i.e. full}), \quad (4.21)$$

with $z := (x, \lambda, \alpha)$, if and only if f is a universal unfolding.

Theorem 4.22. Let $p = q + 1 \leq 3$, and suppose $(x_0, \lambda_0, \alpha_0)$ is a universally unfolded $(q+1, j')$ -singularity of (4.1). Let $j \geq 1$, with $j = j'$ or $j = j' - 1$. Then $(x_0, \lambda_0, \alpha_0)$ is either a regular point or a simple turning point (with respect to α_i) of (4.16).

Proof of Theorem 4.20. By the remark after Theorem 4.11, $(x_0, \lambda_0, \alpha_0)$ is a (q, j) -singularity of (4.1) if and only if $h_0(\epsilon, \lambda, \alpha)$ has a (q, j) -singularity at $(0, \lambda_0, \alpha_0)$. The result now follows from (4.15–16), and Theorem 2.7. ■

Proof of Theorem 4.21. Theorem 4.20 guarantees that $H_{q,j}^0 = 0$. From (4.16, 17) we have

$$\frac{\partial H_{q,j}^0}{\partial z} = \begin{bmatrix} \hat{f}_v & \hat{f}_\epsilon & \hat{f}_{(\lambda, \alpha)} \\ F_c c_v & F_c c_\epsilon + F_\epsilon & F_c c_{(\lambda, \alpha)} + F_{(\lambda, \alpha)} \end{bmatrix} \begin{bmatrix} [V_2 \ v_1]^{-1} & 0 \\ 0 & I \end{bmatrix}$$

Here we have dropped the subscripts q, j and the superscript 0. We see that H_z^0 is nonsingular if and only if the following matrix is nonsingular:

$$\begin{bmatrix} \hat{f}_v & \hat{f}_{(\epsilon, \lambda, \alpha)} \\ F_c c_v & F_c c_{(\epsilon, \lambda, \alpha)} + F_{(\epsilon, \lambda, \alpha)} \end{bmatrix} = \begin{bmatrix} \hat{f}_v & 0 \\ F_c c_v & I_{q+2} \end{bmatrix} \begin{bmatrix} I_{n-1} & \hat{f}_v^{-1}[\hat{f}_{(\epsilon, \lambda, \alpha)}] \\ 0 & F_{(\epsilon, \lambda, \alpha)} \end{bmatrix} \quad (4.23)$$

Here we have differentiated (4.17) to find $c_v = \hat{f}_v$, etc. By Lemma 4.6 $A = \hat{f}_v$ is nonsingular. Furthermore, Theorem 2.8 shows that $F_{(\epsilon, \lambda, \alpha)}^0$ is nonsingular precisely when f is a universal unfolding, and so the result follows from (4.23).

Proof of Theorem 4.22. By Theorem 4.21, $\frac{\partial H_{q+1,j'}^0}{\partial z}$ is nonsingular. Therefore, from (4.23) and the definitions of $F_{q,j}$, $F_{q+1,j'}$, it can be shown that

$$\text{Rank} \left(\frac{\partial H_{q,j}^0}{\partial z} \right) = n + q + 1, \quad (\text{i.e. full}).$$

Hence the result follows. ■

The extended systems given in [12] can be recovered by choosing appropriate vectors v_1, w_1 , etc. In particular, we should take $v_1 = \varphi_1(x, \lambda, \alpha)$ with φ_1 being the singular vector defined in [12]. We do not pursue this here. Instead we consider

$$v_1 = e_r, \quad w_1 = e_l, \quad (4.24)$$

and V_2 (W_2) equal to I_n with the r^{th} column (l^{th} row) removed. Here r and l should be chosen such that (4.3) is satisfied. For example, we could take r and

l to correspond to large components in the vectors φ_1 and ψ_1 , i.e.

$$|\varphi_{1,r}| \approx \max_{1 \leq i \leq n} |\varphi_{1,i}|, \quad |\psi_{1,r}| \approx \max_{1 \leq i \leq n} |\psi_{1,i}|.$$

(During a continuation process the values of r and l could be chosen by considering the φ_1 and ψ_1 at the previous step.) With this choice of v_1 , w_1 we find from (4.7) that \hat{f} is just f with the l^{th} row crossed out, which we write as $\hat{f} = (f_l)'$. Similarly $v = (x_r)'$, $\varepsilon = x_r$, and $g = f_l$. We can rewrite (4.12, 14) as

$$\hat{f} := (f_l)'(u, x_r, \lambda, \alpha) = c, \quad (4.25a)$$

$$h(x_r, \lambda, \alpha, c) := f_l(u(x_r, \lambda, \alpha, c), x_r, \lambda, \alpha) = 0. \quad (4.25b)$$

In particular, the reduced equation (4.25b) can be written in terms of a *single component* of f . We discuss the $(0,1)$ -extended system of this form in the next section.

5. Numerical Details. In this section we discuss in detail the implementation of the approach discussed in Section 4 for the choice (4.24) and for the $(0,1)$ -extended system. Thus the material in this section can be regarded as another technique for the computation of a simple quadratic turning point. However we emphasise that a similar implementation can be used for other extended systems.

For convenience we take $r = l = n$ in (4.24), so $\hat{f} = (f_1, \dots, f_{n-1})^T$, $h = f_n$, and $x = (w, \varepsilon)^T$. The extended system is

$$H_{0,1} := H = \begin{bmatrix} \hat{f}(w, \varepsilon, \lambda, \alpha) \\ h(\varepsilon, \lambda, \alpha, c) \\ h_\varepsilon(\varepsilon, \lambda, \alpha, c) \end{bmatrix} = 0, \quad (5.1a)$$

with $c := \hat{f}(w, \varepsilon, \lambda, \alpha)$, and the side constraints are

$$h_{\varepsilon\varepsilon} \neq 0 \neq h_\lambda(\varepsilon, \lambda, \alpha, c). \quad (5.1b)$$

Here, for $c \in R^n$, c near 0, $u(\varepsilon, \lambda, \alpha, c)$ and $h(\varepsilon, \lambda, \alpha, c)$ are defined by

$$c = \hat{f}(u, \varepsilon, \lambda, \alpha), \quad h(\varepsilon, \lambda, \alpha, c) := f_n(u(\varepsilon, \lambda, \alpha, c), \varepsilon, \lambda, \alpha). \quad (5.2)$$

To evaluate H at a given point $(\mathbf{x}, \lambda, \alpha) = (\mathbf{w}, \varepsilon, \lambda, \alpha)$ we proceed as follows. We evaluate $\mathbf{c} = \hat{\mathbf{f}}(\mathbf{w}, \varepsilon, \lambda, \alpha)$, and note from (5.2) that we can take $\mathbf{u}(\varepsilon, \lambda, \alpha, \mathbf{c}) = \mathbf{w}$. Therefore, by (4.25),

$$\begin{bmatrix} \hat{\mathbf{f}}(\mathbf{u}, \varepsilon, \lambda, \alpha) \\ \mathbf{h}(\mathbf{u}, \varepsilon, \lambda, \alpha) \end{bmatrix} = \mathbf{f}(\mathbf{x}, \lambda, \alpha). \quad (5.3)$$

To compute \mathbf{h}_ε we differentiate (5.2) to find

$$\hat{\mathbf{f}}_{\mathbf{w}} \mathbf{u}_\varepsilon + \hat{\mathbf{f}}_\varepsilon = 0, \quad \mathbf{h}_\varepsilon = \mathbf{f}_{n\varepsilon} + \mathbf{f}_{n\mathbf{w}} \mathbf{u}_\varepsilon. \quad (5.4a)$$

We assume that (4.3) is satisfied, so $\hat{\mathbf{f}}_{\mathbf{w}}$ is nonsingular. Therefore, by (5.4a),

$$\mathbf{u}_\varepsilon = -\hat{\mathbf{f}}_{\mathbf{w}}^{-1} \hat{\mathbf{f}}_\varepsilon. \quad (5.4b)$$

Finally $H(\mathbf{w}, \varepsilon, \lambda, \alpha)$ is given by (5.1a, 3, and 4).

To evaluate $\frac{\partial H}{\partial(\mathbf{w}, \varepsilon, \lambda)}$ we first consider (4.23), which provides

$$\frac{\partial H}{\partial(\mathbf{w}, \varepsilon, \lambda)} = \begin{bmatrix} \hat{\mathbf{f}}_{\mathbf{w}} & 0 \\ F_c \mathbf{c}_{\mathbf{w}} & I_2 \end{bmatrix} \begin{bmatrix} I_{n-1} & \hat{\mathbf{f}}_{\mathbf{w}}^{-1} \mathbf{f}_\varepsilon & \hat{\mathbf{f}}_{\mathbf{w}}^{-1} \mathbf{f}_\lambda \\ 0 & h_\varepsilon & h_\lambda \\ 0 & h_{\varepsilon\varepsilon} & h_{\varepsilon\lambda} \end{bmatrix} \quad (5.5)$$

with

$$F_c = \frac{\partial}{\partial c} \begin{bmatrix} \mathbf{h} \\ \mathbf{h}_\varepsilon \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{n\mathbf{w}} \\ \mathbf{h}_{\varepsilon c} \end{bmatrix}, \quad \mathbf{c}_{\mathbf{w}} = \hat{\mathbf{f}}_{\mathbf{w}}.$$

A straightforward calculation shows that for

$$\varphi_1 := \begin{bmatrix} \mathbf{u}_\varepsilon \\ 1 \end{bmatrix} = \begin{bmatrix} -\hat{\mathbf{f}}_{\mathbf{w}}^{-1} \hat{\mathbf{f}}_\varepsilon \\ 1 \end{bmatrix}, \quad \psi_1^T = (-\mathbf{f}_{n\mathbf{w}} \hat{\mathbf{f}}_{\mathbf{w}}^{-1}, 1) \quad (5.6a)$$

$$z_1 := \begin{bmatrix} \mathbf{u}_\lambda \\ 0 \end{bmatrix} = \begin{bmatrix} -\hat{\mathbf{f}}_{\mathbf{w}}^{-1} \hat{\mathbf{f}}_\lambda \\ 0 \end{bmatrix}, \quad (5.6b)$$

the factorization given in (5.5) can be written as

$$\frac{\partial H}{\partial(w, \varepsilon, \lambda)} = \begin{bmatrix} \hat{f}_w & 0 & 0 \\ f_{nw} & 1 & 0 \\ \psi_1^T f_{zw} \varphi_1 & 0 & 1 \end{bmatrix} \begin{bmatrix} I_{n-1} & -u_\varepsilon & -u_\lambda \\ 0^T & f_{nx} \varphi_1 & \psi_1^T f_\lambda \\ 0^T & \psi_1^T f_{zz} \varphi_1 \varphi_1 & \psi_1^T [f_{zx} + f_{z\lambda}] \varphi_1 \end{bmatrix} \quad (5.7)$$

Recall \hat{f}_w was nonsingular (at the root), so given an LU-factorization of \hat{f}_w we can compute $\frac{\partial H}{\partial(w, \varepsilon, \lambda)}$ in the factored form (5.7), with φ_i, ψ_1 given by (5.6). The factorization is easily completed by LU-factoring the 2×2 matrix

$$E := \begin{bmatrix} f_{nx} \varphi_1 & \psi_1^T f_\lambda \\ \psi_1^T f_{zz} \varphi_1 \varphi_1 & \psi_1^T [f_{zx} + f_{z\lambda}] \varphi_1 \end{bmatrix}. \quad (5.8)$$

The operation count for this factorization is essentially the same as for the factorization discussed in [12].

References.

- [1] W.J. Bejn, Numerical analysis of singularities in a diffusion reaction model, to appear in Proceedings of the EQUADIFF, Wurzburg (1982), Springer Lecture Notes, Berlin.
- [2] W.J. Bejn, Defining equations for singular solutions and numerical applications, this proceedings.
- [3] M.G. Crandall and P.H. Rabinowitz, Bifurcation, perturbation of simple eigenvalues and linearised stability, Arch. Rat. Mech. Anal., 52 (1973), pp. 161-180.
- [4] M. Golubitsky and B. Keyfitz, A qualitative study of the steady-state solutions for a continuous flow stirred tank chemical reactor, SIAM J. Math. Anal., 11 (1980), pp. 316-339.
- [5] M. Golubitsky and D. Schaeffer, A theory for imperfect bifurcation via singularity theory, Commun. Pure and Appl. Math., 32 (1979), pp. 21-98.
- [6] A.D. Jepson and A. Spence, The numerical solution of nonlinear equations having several parameters. Part I: Scalar equations, submitted to SIAM J. Numer. Anal..
- [7] A.D. Jepson and A. Spence, The numerical solution of nonlinear equations having several parameters. Part II: Vector equations, in preparation.
- [8] J.P. Keener and H.B. Keller, Perturbed bifurcation theory, Arch. Rat. Mech. Anal., 50 (1973), pp. 159-175.
- [9] H.B. Keller, Numerical solution of bifurcation and nonlinear eigenvalue problems, in Applications of Bifurcation Theory, P.H. Rabinowitz (Ed.), Academic Press, N.Y. (1973), pp. 359-384.
- [10] G. Moore, The numerical treatment of non-trivial bifurcation points, Numer. Func. Anal. and Optimiz., 6 (1980), pp. 441-472.
- [11] W.C. Rheinboldt and J.V. Burkardt, A locally parameterized continuation process, ACM TOMS, 9 (1983), pp. 215-235.
- [12] A. Spence and A.D. Jepson, Numerical computation of cusps, bifurcation points and isola formation points in two-parameter problems, this proceedings.
- [13] I. Stackgold, Branching of solutions of nonlinear equations, SIAM Rev., 13 (1971), pp. 289-332.

ON A GENERAL TECHNIQUE FOR FINDING DIRECTIONS
PROCEEDING FROM BIFURCATION POINTS

Ralph Baker Kearfott

Various quite satisfactory analytical and numerical techniques are available for analysing bifurcation points when something about the structure is known *a priori*. The author previously introduced a method applicable when such information is not present, or when the arcs intersect tangentially. That method is discussed here, with particular emphasis on avenues to improvement in efficiency and reliability.

1. Introduction and Background

We consider the solution sets of the parametrized nonlinear system

$$(1.1) \quad H(y) = H(x, \lambda) = 0$$

where $H : \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$. Of interest are the bifurcation points $y^* = (x^*, \lambda^*)$ where $\text{rank}(DH(y^*)) < n$, and where two or more arcs in the solution set of (1.1) intersect.

Due to the ubiquitous occurrence of bifurcation in physical processes, specific techniques for finding y^* and following paths away from y^* are well developed. These techniques are applicable when the bifurcation point is simple (only two arcs intersect), when $y^* = (0, \lambda^*)$ (bifurcation from the "trivial" solution), when symmetry can be used, when unfoldings ([5]) are easily computed, etc. These, described in other papers in this volume and elsewhere, are usually quite efficient, and are to be preferred for the specific problems they are meant to solve.

Other methods are meant to be more general. Perhaps the two salient ones are (1) use of simplicial discretizations, possibly in conjunction with "artificial bifurcation" (cf. [16]), and (2) solution of a system of polynomial equations such that the degree of each equation equals the order

$p > 1$ of the first non-zero higher-order derivative tensor (see [12] and also [11]).

Simplicial methods are mathematically equivalent to replacing H by a perturbation which has no bifurcation points, for which the paths can be followed. Typically, paths intersecting $\lambda = 0$ are followed until they diverge or intersect $\lambda = 1$, but not all paths corresponding to a bifurcation point may correspond to simplicial paths intersecting at $t = 0$; artificial bifurcation and other ad hoc techniques are used to connect these paths. Such techniques would be harder to apply when large numbers of arcs intersect.

In the Keller/Langford method ([12]), solutions of the system of polynomial equations correspond to normalized tangent vectors to arcs emanating from y^* . The solutions are in a one-to-one correspondence with the arcs, provided there are no multiple solutions (arcs do not intersect tangentially) and provided the solutions are isolated on the intersection of any sufficiently small sphere in \mathbb{R}^n about the bifurcation point. (This follows from arguments in [12] and [11].) These equations have not been extensively employed since the coefficients depend on the p -th order partial derivatives of the components of H . However, if H is polynomial, these derivatives can be computed analytically at compile time ([17]). In such cases, the numerical techniques described below might be successfully applied to the Keller/Langford equations (though below a different polynomial system is used).

2. The General Method

An implementation of the basic method is discussed in [8]. The technique depends on the fact that arcs bifurcating from y^* can be approximated by arcs in the k -dimensional affine space $\Pi(y^*)$ through y^* and with directions defined by the null space $\mathcal{N}(D(H)(y^*))$ of $D(H)(y^*)$. In particular, for sufficiently small δ solutions to $H(y) = 0$ on a sphere $\mathcal{S}_\delta(y^*)$ of radius δ about y^* correspond to minima of $\|H\|^2$ in $\mathcal{S}_\delta(y^*) \cap \Pi(y^*)$ (throughout, $\|\cdot\|$ means Euclidean norm); a precise

one-to-one correspondence can be shown under rather general transversality conditions on the tangent and normal vectors at y^* ([10]).

In [8] minima of $\|H\|$ in $\mathcal{S}_\delta(y^*) \cap \Pi(y^*)$ were found directly by using the simplex method of Nelder and Mead. Since an unspecified number of starting points was required to insure at least one such point occurred in the domain of attraction of each minimum of $\|H\|$, the procedure involved a heuristic. Also, the efficiency of the overall implementation left something to be desired.

Here, we discuss potentially more efficient procedures which in addition are less heuristic. The first of these, presented below, depends on the fact that all solutions to a polynomial system of equations can be approximated by homotopy methods. The second, explained briefly at the end of the paper, is based on a deterministically driven search similar, but not identical, to that used for computing the Brouwer degree of maps ([7]).

Let H be represented by a column vector, and define $\varphi(y) = H^T(y)H(y) = \|H(y)\|^2$. Then the minimization problem is:

$$(2.1) \quad \min_{y \in \Pi(y^*)} \varphi(y) \quad \text{subject to } y \in \mathcal{S}_\delta(y^*).$$

Let $\mathcal{N}(DH(y^*)) = \text{sp} \{v_1, \dots, v_k\}$ so that $y \in \Pi(y^*) \Rightarrow y = y(a) = y(a_1, \dots, a_k) = y^* + \sum_{\ell=1}^k a_\ell v_\ell$; let $\nabla_a \varphi$ represent the gradient of φ with respect to a , and set $J = DH(y)$. Then:

$$(2.2) \quad \nabla_a \varphi = \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix} J^T H(y).$$

Applying the Lagrange multiplier technique to (2.1), we thus conclude that all solutions of (2.1) are solutions to the system:

$$(2.3) \quad \begin{bmatrix} v_1^T \\ \vdots \\ v_k^T \end{bmatrix} J^T H(y) + \Lambda \begin{bmatrix} 2a_1 \\ \vdots \\ 2a_k \end{bmatrix} = 0, \quad \sum_{\ell=1}^k a_\ell^2 - \delta^2 = 0$$

of $k+1$ polynomial equations in the $k+1$ unknowns a_1, \dots, a_k, Λ . Let $Y = (a_1, \dots, a_k, \Lambda)$ and define $F : \mathbb{R}^{k+1} \rightarrow \mathbb{R}^{k+1}$ to equal the left-hand sides of (2.3). Then finding paths bifurcating from y^* reduces to finding all solutions to the polynomial system $F(Y) = 0$.

Various homotopy algorithms will give, in theory with probability one, all solutions to $F(Y) = 0$ (cf. eg. [2], [3], [4], [14], [15]). We discuss these briefly in the next section; here, we use an example to observe properties of (2.3).

Consider:

$$(2.4) \quad H_1(x, \lambda) = x[x^4 - (2\lambda - 1)^2],$$

the solution set of which occurs in Fig. (2.1). Corresponding to a degenerate case exhibiting symmetry (see [6]),

the bifurcation point at $(0, 1/2)$ has two branches intersecting tangentially.

The arc directions are of the form (a_1, a_2) ; the Keller/Langford equations are: $-24a_1a_2^2 = 0$; $a_1^2 + a_2^2 = \delta^2$, with simple roots $(0, \delta)$ and $(0, -\delta)$ and with double roots $(\delta, 0)$ and $(-\delta, 0)$.

The system (2.3) consists of a polyno-

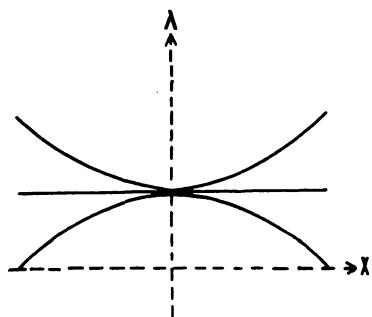


Fig. (2.1)

mial of degree 9, a polynomial of degree 7, and a polynomial of degree 2, which in general will have 126 complex solutions, counting multiplicities. Among these solutions are distinct roots corresponding to the tangentially intersecting branches, but maxima of $\|H\|$ and non-real solutions not on $\mathcal{S}_\delta(y^*)$ also occur. Such solutions must be ignored or rejected early in the computation for (2.3) to be practical. Note, however, that in all cases only first derivatives enter into (2.3), whereas at least second order derivatives are required for the Keller/Langford equations.

4. A Tentative Method

Here, we consider solution algorithms for (2.3) (or for the Keller/Langford equations) involving imbeddings $\tilde{H} : \mathbb{C}^{k+1} \times [0, 1] \rightarrow \mathbb{C}^{k+1}$

of the complex extension of F , such that roots of $\tilde{H}(z, 0)$ are known and $\tilde{H}(z, 1) = F(z)$; roots of F are found by following paths from $t = 0$ to $t = 1$, $([2], [3], [4], [14], [15], \text{ etc.})$. In general, if the degree of the ℓ -th component of F is d_ℓ , there are $ns = \prod_{k=1}^{k+1} d_\ell$ solutions, and ns corresponding paths to be followed. Good general continuation method software (see [13], [14], [18], [19], [20], and also [1]) is available, but, due to the size of ns , naive implementations will not be reliable.

For example, the minimum distance between desired roots of F is on the order of $\sqrt[p]{\delta}$ for some p , depending on the degree of contact at y^* (for (2.4), $p = 2$); the continuation method has been observed to jump from one path to another when the total stepsize has been allowed to exceed this value. Also, scaling problems and path-jumping will occur if the roots of $H(z, 0)$ have norms appreciably different from δ .

Additional problems occur when paths $H(y) = 0$ intersect $\mathcal{O}_\delta(y^*) \cap \Pi(y^*)$ and the tangents at the points of intersection are perpendicular to $\mathcal{O}_\delta(y^*)$; this would happen, e.g., if $k = n + 1$ and the arcs were linear rays emanating from y^* . In such cases, DF is singular at the solutions, and the roots are found only with reduced accuracy.

At real solutions of $F(Y) = 0$, $\|a\| = \delta$ and $O(\Lambda) = \sqrt[q]{\delta}$ where q is related to the degree of F . Also, for appropriate $\tilde{H}(z, 0)$, paths leading to real solutions of F will be within $\{Y : \|Y\| \leq M\sqrt[q]{\delta}\}$ for some small M . This allows early rejection of irrelevant solutions.

The method was tested using (2.4), $\delta = .515388$, and using PITCON ([18], [19]). The Garcia/Zangwill homotopy ([2], [14]) was used since we had initial difficulty implementing the potentially better methods in [15]. The maximum predictor step was $.2\delta$, and path following stopped whenever $\|z\| > 5$; roots of $\tilde{H}(z, 0)$ had norms approximately 1. Computations were done on a Honeywell 68/80 using 27 bit mantissas, and $D\tilde{H}$ was computed using central differences. In Table 4.1 paths were oriented in the direction of increasing t , and the tangent-normal corrector parametrization was used. (Rheinboldt's local parametrization gave similar results with less function evaluations, also jumping paths.) The fifth

column gives the determinant of the tangent system at $F = 0$, and the sixth gives the number of evaluations of \tilde{H} for that particular path. Note that root no. 2 and irrelevant roots 1 and 6 were found twice, indicating the algorithm left the path it was following. (Note also that the conjugate of irrelevant solution 3 should have been found but was not.)

Solutions					
no	x	$\lambda - \frac{1}{2}$	Λ	det	nf
1	.5	.125	0	1×2^{-2}	80
2	.5	-.125	0	3×2^{-3}	88(83)
3	-.5	.125	0	4×2^{-3}	83
4	-.5	-.125	0	3×2^{-3}	126
5	0	.515388	0	3×2^0	95
6	0	-.515388	0	3×2^0	88

Irrelevant Solutions					
no	x	$\lambda - \frac{1}{2}$	Λ	det	nf
1	.515388	0	-.0124	1×2^{-2}	96(91)
2	$2.062i$	+2.125	0	1×2^{36}	208
3	$2.062i$	-2.125	0	1×2^{38}	220
4	-.2924	-.4244	-.2439	2×2^1	87
5	-.2924	.4244	-.2439	1×2^1	117
6	.2924	-.4244	-.2439	2×2^1	104(106)
7	-.515388	0	-.0124	2×2^{-3}	87
8	$-2.062i$	2.125	0	3×2^{37}	253

Table 4.1

5. A Possible Alternative

Adjustments to the above scheme may make it more practical; it also may be practical when applied to the Keller/Langford equations instead of (2.3). However, the problem of finding all solutions to (2.3) has several characteristics which make an alternative method practical; these

are (1) the fact that k is small or can be made small by use of symmetry; (2) existence of large numbers of solutions, distributed more or less uniformly over $[\Pi(y_*) \cap \mathcal{O}_\delta^{(y^*)}] \times [-K, K]$. In this case, use of a generalized method of bisection as described in [7] and [9] becomes more attractive.

In such a generalized bisection method, simplices would not be further subdivided according to sign changes or whether a topological degree were nonzero. Instead, an analysis of the values of the components of F and bounds on the norm of D^2F (or more generally, Lipschitz constants for the derivatives of the components of F) would be used to indicate whenever a component of F : (i) could not vanish on a simplex, or (ii) could vanish at most once on any line in the simplex; if (i) held for one component, if (ii) held for all components, or if the diameter of the given simplex were smaller than a given tolerance, then triangulation of that simplex would stop.

Generalized bisection has been considered non-competitive in general. However, the large number of solutions and need to obtain them all makes generalized bisection appropriate for their isolation. Details and convergence proofs will be given in a later work.

6. References

- [1] E. L. Allgower and K. Georg: Simplicial and continuation methods for approximating fixed points and solutions to systems of equations, SIAM Review 22 no. 1 (1980), pp. 28-85.
- [2] C. B. Garcia and W. I. Zangwill: Finding all solutions to polynomial systems and other systems of equations, Mathematical Programming 16 (1979), pp. 159-176.
- [3] C. B. Garcia and W. I. Zangwill: Pathways to Solutions, Fixed Points, and Equilibria, Prentice-Hall, Englewood Cliffs, 1981.
- [4] C. B. Garcia and T. Y. Li: On the number of solutions to polynomial systems of equations, SIAM J. Numer. Anal. 17 no. 4 (1980), pp. 540-546.
- [5] M. Golubitsky and D. Schaeffer: A theory for imperfect bifurcations via singularity theory, Comm. Pure and Applied Math. 32 (1979), pp. 21-98.

- [6] M. Golubitsky and W.F. Langford: Classification and unfoldings of degenerate Hopf bifurcations, *J. Diff. Eq.* 41 no. 3 (1981), pp. 375-415.
- [7] R. B. Kearfott: An efficient degree computation method for a generalized method of bisection, *Numer. Math.* 32 (1979), pp. 109-127.
- [8] R. B. Kearfott: Some general bifurcation techniques, *SIAM J. Sci. Stat. Comput.* 4 no. 1 (1983), pp. 52-68.
- [9] R. B. Kearfott: An improved program for generalized bisection, to appear.
- [10] R. B. Kearfott: Analysis of a general bifurcation technique, to appear.
- [11] H. B. Keller: Numerical solution of bifurcation and nonlinear eigenvalue problems, in *Applications of Bifurcation Theory*, ed. P.H. Rabinowitz., Academic Press, New York, 1977.
- [12] H. B. Keller and W.F. Langford: Iterations, perturbations, and multiplicities for nonlinear bifurcation problems, *Arch. Rat. Mech. Anal.* 48 (1972), pp. 83-108.
- [13] M. Kubicek: Dependence of solution of nonlinear systems on a parameter, *ACM TOMS* 2 no. 1 (1976), pp. 98-107.
- [14] A.P. Morgan: A method for computing all solutions to systems of polynomial equations, *ACM TOMS* 9 no. 1 (1983), pp. 1-17.
- [15] A.P. Morgan: Computing all solutions to polynomial systems using homogeneous coordinates in Euclidean space, in *Numerical Analysis of Parametrized Nonlinear Equations*, University of Arkansas seventh lecture series, 1983.
- [16] H.-O. Peitgen and M. Prüfer: The Leray-Schauder continuation method is a constructive element in the numerical study of nonlinear eigenvalue and bifurcation problems, in *Functional Differential Equations and Approximation of Fixed Points*, Springer Lecture Notes no. 730, New York, 1979.
- [17] L.B. Rall: Differentiation in PASCAL-SC: type gradient, Mathematics Research Center technical report no. 2400, University of Wisconsin, Madison, 1982.
- [18] W.C. Rheinboldt and J.V. Burkardt: A program for a locally-parametrized continuation process, *ACM TOMS* 9 no. 2 (1983), pp. 236-241.
- [19] W.C. Rheinboldt and J.V. Burkardt: A locally parametrized continuation process, *ACM TOMS* 9 no. 2 (1983), pp. 215-235.

- [20] L. T. Watson and D. Fenner: The Chow-Yorke algorithm for fixed points or zeros of C^2 maps, ACM TOMS 6 (1980), pp. 252-260.

Ralph Baker Kearfott

Department of Mathematics and Statistics

University of Southwestern Louisiana

USL Box 4-1010

Lafayette, Louisiana 70504

USA

Steady State and Periodic Solution
Paths: their Bifurcations and Computations
by

A.D. Jepson¹ and H.B. Keller²

1. Introduction. In this work we present a brief account of the theory and numerical methods for the analysis and solution of nonlinear autonomous differential equations of the form

$$(1.1) \quad \frac{d}{d\tau} w = f(w, \lambda, \alpha); \quad f: \mathcal{B}_1 \times \mathbb{R}^2 \rightarrow \mathcal{B}_2.$$

Here λ, α are real parameters, f is a smooth function, and $\mathcal{B}_1, \mathcal{B}_2$ are Banach spaces. A thorough analysis of (1.1) involves, at least, the following stages.

(i) A study of steady state solutions of (1.1); that is solutions (w, λ, α) of

$$(1.2) \quad 0 = f(w, \lambda, \alpha);$$

(ii) The determination of particular steady solutions from which solutions periodic in τ bifurcate;

(iii) An analysis of periodic solutions of (1.1), including their dependence of (λ, α) ; and

(iv) A study of non-periodic or transient solutions.

The computational methods sketched here are based on continuation or path following techniques and provide powerful tools for carrying out the first three stages of the above analysis. Most of the discussion is based on the work of H.B. Keller [15] and A.D. Jepson[10].

In the first stage we seek manifolds,

$$(1.3a) \quad M_S : \{(w(\lambda, \alpha), \lambda, \alpha) \mid (\lambda, \alpha) \in D \subset \mathbb{R}^2\},$$

or smooth branches,

$$(1.3b) \quad \Gamma_S : \{(w(s), \lambda(s), \alpha(s)) \mid s \in (a, b)\},$$

of steady solutions of (1.1). At present our methods compute only one parameter families or "paths" such as Γ_s (see Section

¹Computer Science, University of Toronto, Toronto, Ontario, Canada.

²Applied Mathematics, California Institute of Technology, Pasadena, California.

2); two dimensional manifolds such as M_S must be determined by computing families of paths. Of particular importance are points on the "boundaries" of a solution manifold M_S , and also points for which every neighborhood contains nonunique solutions. In Figure 1 these two types of points are represented by the fold point A and by the bifurcation point B, respectively. Computational methods for the direct calculation of families or paths of such special points are indicated in Sections 3 and 4. They provide an effective means of analyzing (1.2) over wide ranges of the parameters (λ, α) .

The second stage of analysis involves the computation of points, such as C in Figure 1, at which branches of periodic solutions (Γ_2) bifurcate from branches of steady solutions (Γ_1) . The generic bifurcation of this type is called a Hopf bifurcation, and in Section 5 we discuss a method for the computation of families or paths of Hopf bifurcation points.

In the third stage of analysis we seek manifolds,

$$(1.4a) \quad M_p : \{(w(\tau; \lambda, \alpha), T(\lambda, \alpha), \lambda, \alpha) | \lambda, \alpha \in D\},$$

or one-parameter families,

$$(1.4b) \quad \Gamma_s : \{(w(\tau; s), T(s), \lambda(s), \alpha(s)) | s \in (a, b)\},$$

of periodic solutions $(w(\tau; \lambda, \alpha), \lambda, \alpha)$ of (1.1). Here the unknown period is denoted by $T(\lambda, \alpha)$ or $T(s)$. Methods for the computation of paths of periodic solutions, such as Γ_2 and Γ_3 in Figure 1, are considered in Section 2. Furthermore, paths of fold points and bifurcation points, such as E and D respectively, on the boundary of a manifold M_p , of periodic solutions, can be computed in much the same way as the corresponding paths for steady solutions (see Sections 3 and 4). Finally in Section 5 we discuss the computation of a periodic solution branch near a Hopf bifurcation point and consider the problem of switching from a steady state branch such as Γ_1 to the bifurcating periodic solution branch represented by Γ_2 in Figure 1.

We emphasize that all of our methods are based on well known continuation and branch switching techniques initiated in [15]. Most of the extensions to periodic solutions were

first done independently in [10] and [7]. Here we illustrate how these bifurcation techniques can be applied to many different facets of problem (1.1). The combination of the resulting methods provides a powerful set of tools for the analysis of nonlinear autonomous systems. Indeed several powerful computer programs for doing this have been developed. The most advanced is Doedel's AUTO [7]. In closing these introductory remarks we stress that our analysis and techniques are by no means restricted to ordinary differential equations. Rather they apply to O.D.E.'s in a Banach space setting. Thus for example the unsteady Navier-Stokes equations are included.

2. Regular Points. Let \mathcal{B}_3 , \mathcal{B}_4 be Banach spaces and $F: \mathcal{B}_3 \times \mathbb{R} \rightarrow \mathcal{B}_4$ be a smooth function. A solution (z_0, β_0) of (2.1) $F(z, \beta) = 0$

is called a regular point of (2.1) if the Frechet derivative

$$(2.2) \quad F_z^0 \equiv \frac{\partial F}{\partial z}(z_0, \beta_0)$$

is nonsingular. Otherwise the solution (z_0, β_0) is called a singular point. For a regular point, (z_0, β_0) , the Implicit Function Theorem insures the existence and local uniqueness of a smooth branch of solutions, $(z(\beta), \beta)$, for β near β_0 and satisfying $z(\beta_0) = z_0$. Predictor-solver continuation [19], [15] and simplicial [23] methods have been used to approximate solution branches near regular points. Here we do not consider simplicial methods since they are efficient only when \mathcal{B}_3 and \mathcal{B}_4 have respectively small (finite) dimensions. A brief summary of predictor-solver methods is given below.

Observe that $F(z(\beta), \beta) = 0$ is an identity in β and so formal differentiation with respect to β yields, with $\dot{z}(\beta) \equiv dz/d\beta$:

$$(2.3) \quad F_z(z(\beta), \beta) \dot{z}(\beta) = -F_\beta(z(\beta), \beta) .$$

If $(z_0, \beta_0) = (z(\beta_0), \beta_0)$ is a regular point then (2.3) can be solved for $\dot{z}(\beta_0) \equiv \dot{z}_0$. A good approximation to $z(\beta)$ for $|\beta - \beta_0|$ small can then be obtained by using the tangent approximation

$$(2.4) \quad z^0(\beta) = z_0 + (\beta - \beta_0) \dot{z}_0 .$$

This can be viewed as one step of Euler's method for integrating

(2.3). The predictor (2.4) is then used to supply the initial guess, $z^0 \equiv z^0(\beta_1)$ for a solver such as the chord method:

$$(2.5) \quad F_z(z^0(\beta_1), \beta_1)(z^{v+1} - z^v) = -F(z^v, \beta_1); \quad v = 0, 1, 2, \dots .$$

Conditions ensuring the convergence of (2.5) are easily given by requiring that the step size $|\beta_1 - \beta_0|$ is not too large. In fact, for small enough step sizes, (2.5) converges to the solution $z(\beta_1)$ on the branch passing through (z_0, β_0) . If this new solution, $(z(\beta_1), \beta_1)$, is also a regular point of (2.1) then we can apply a second predictor-solver step to find $(z(\beta_2), \beta_2)$. In this way global solution branches may be obtained; the procedure can fail only if F_z becomes singular or the branch departs from the region in which F is smooth.

Clearly the predictor (2.4) and the solver (2.5) can be replaced by many other schemes. In particular Newton's method and its variants are most effective as solvers. We do not pursue these aspects of the procedures here; instead we prefer to concentrate on the construction of appropriate functions $F(z, \beta)$ relevant to solving different stages of the general problems outlined in §1. For example, to compute a path of steady states for (1.1) we could use

$$(2.6) \quad F(z, \beta) \equiv f(w, \lambda(\beta), \alpha(\beta)), \quad w \equiv z ;$$

where $\lambda(\beta)$, $\alpha(\beta)$ are given (smooth) functions, and $z \in \mathcal{B}_1$. In this case we might expect our continuation procedure to fail or experience difficulties near points at which f_w (and hence F_z) is singular. An alternate choice of F , for which F_z can be nonsingular when f_w is singular, removes some of these difficulties and is considered in Section 3. In fact, a major theme in the remainder of this paper is the construction of functions F for which the desired solutions are regular points.

As a final example for this section we determine appropriate functions $F(z, \beta)$ for the calculation of periodic solutions. For this purpose we use the standard approach of introducing the new time variable

$$(2.7a) \quad t = \tau/T ,$$

and writing the solution as

$$(2.7b) \quad u(t) = w(\tau(t)) \in \mathcal{B}_1[0,1] .$$

Then we seek periodic solutions of (1.1) having period T by considering the boundary value problem

$$(2.8) \quad \begin{aligned} a) \quad & \frac{d}{dt} u = Tf(u, \lambda, \alpha) , \\ b) \quad & u(1) - u(0) = 0 . \end{aligned}$$

Clearly a solution $(u(t), T, \lambda, \alpha)$ of (2.8) is not unique since $(u(t+\theta), T, \lambda, \alpha)$ is also a solution for any $\theta \in \mathbb{R}$. In an attempt to make solutions unique we can impose an additional scalar constraint called a "phase condition." For the present we denote this condition simply as

$$(2.8c) \quad p(u, T, \lambda, \alpha) = 0$$

where $p: \mathcal{B}_1[0,1] \times \mathbb{R}^3 \rightarrow \mathbb{R}$. Then (2.8a,b,c) can be written as

$$(2.9) \quad F_B(z, \beta) \equiv \begin{pmatrix} \frac{d}{dt} u - Tf(u, \lambda(\beta), \alpha(\beta)) \\ u(0) - u(1) \\ p(u, T, \lambda(\beta), \alpha(\beta)) \end{pmatrix} = 0 ,$$

where $z \in (u, T)$, $F_B: (\mathcal{B}_1[0,1] \times \mathbb{R}) \times \mathbb{R} \rightarrow \mathcal{B}_2[0,1] \times \mathcal{B}_1 \times \mathbb{R}$, and $\lambda(\beta)$, $\alpha(\beta)$ are smooth functions of $\beta \in \mathbb{R}$.

Another reformulation of (1.1) proceeds by using (2.8a) and then denoting the solution of the initial value problem

$$(2.10) \quad a) \quad \frac{d}{dt} y = Tf(y, \lambda, \alpha), \quad b) \quad y(0) = \zeta \in \mathcal{B}_1$$

by

$$(2.10) \quad c) \quad y = y(t; \zeta, T, \lambda, \alpha) .$$

Now conditions (2.8b,c) become

$$(2.11) \quad F_S(z, \beta) \equiv \begin{pmatrix} \zeta - y(1; \zeta, T, \lambda, \alpha) \\ p(y(t; \zeta, T, \lambda, \alpha), T, \lambda, \alpha) \end{pmatrix} = 0 .$$

Here $z = (\zeta, T)$, $\lambda = \lambda(\beta)$, $\alpha = \alpha(\beta)$, and $F_S: (\mathcal{B}_1 \times \mathbb{R}) \times \mathbb{R} \rightarrow \mathcal{B}_1 \times \mathbb{R}$. We call F_S and F_B the shooting method formulation and the boundary problem formulation, respectively. The theoretical discussion of formulation (2.9) is hardly distinct from that for (2.11). However there are very significant numerical differences between the two formulations, see [13] for a discussion of these differences.

Several possible choices for the phase condition (2.8c) are given below.

Phase Condition I. Let $u_0(t)$ be a solution of (2.8a,b) with $(T, \lambda, \alpha) = (T_0, \lambda_0, \alpha_0)$. Pick any $\xi^* \in \mathcal{B}_1^*$ (\equiv dual space of \mathcal{B}_1) such that

$$(2.12a) \quad \xi^* \frac{du_0}{dt}(0) \neq 0 .$$

Then we use the phase condition (see (2.8c))

$$(2.12b) \quad p_I(u, T, \lambda, \alpha) \equiv \xi^* \cdot (u(0) - u_0(0)) = 0 .$$

This condition (essentially due to Poincaré) ensures that the plane of possible initial values in \mathcal{B}_1 is transversal to the trajectory $u_0(t)$ at $t = 0$ (see Figure 2).

Phase Conditions II and III. Let $(u_0(t), T_0, \lambda_0, \alpha_0)$ be as above. During a continuation procedure it is convenient to match the phase of a nearby solution $(u(t), T, \lambda, \alpha)$ as closely as possible to that of $u_0(t)$. If $\mathcal{B}_1[0,1]$ is a Hilbert space, then one way to do this is to seek θ , the phase, such that

$$(2.13) \quad d(\theta) \equiv \int_0^1 \|u(t + \theta) - u_0(t)\|_2^2 dt = \min_{\phi \in \mathbb{R}} d(\phi) .$$

A necessary condition for this minimization to occur at $\theta = 0$ is

$$(2.14) \quad p_{II}(u) \equiv \int_0^1 [u(t) - u_0(t)]^* \frac{du}{dt}(t) dt = 0 .$$

This phase condition was first introduced by E.J. Doedel [7]. Obviously, by changing the roles of $u(t)$ and $u_0(t)$ we obtain

$$(2.15) \quad p_{III}(u) \equiv \int_0^1 [u(t) - u_0(t)]^* \frac{du_0}{dt}(t) dt = 0 ,$$

which is linear in the unknown $u(t)$.

From the above choices we see that many phase conditions could be introduced. The choice (2.12) essentially determines the phase by matching to a given orbit at one point, while (2.14) and (2.15) match along the entire orbits in a mean square sense. Generalizations are easily obtained by considering a weighted mean square in (2.13). Then, in fact, (2.12) would be a special case of (2.15) with a δ -function weight at $t = 0$. Computations

in a variety of cases indicate the practical superiority of p_{II} and p_{III} over p_I .

We now seek conditions to ensure that a solution of (2.9) or (2.11), with $p = p_I$, p_{II} or p_{III} , is regular.

In order to keep the technical difficulties to a minimum we take $\mathcal{B}_1 = \mathcal{B}_2 = \mathbb{R}^n$ and

$$u(t) \in C_n^2[0,1] \equiv \mathcal{B}_1[0,1],$$

that is $u(t) : [0,1] \rightarrow \mathbb{R}^n$ has two continuous derivatives.

Similarly we take $\mathcal{B}_2[0,1] \equiv C_n^1[0,1]$. The norm on $\mathcal{B}_1[0,1]$ is taken to be: $\|\cdot\|_\infty + \left\| \frac{d}{dt} \cdot \right\|_\infty + \left\| \frac{d^2}{dt^2} \cdot \right\|_\infty$, and on $\mathcal{B}_2[0,1]$ we use

$\|\cdot\|_\infty + \left\| \frac{d}{dt} \cdot \right\|_\infty$. Extensions of the following results to possibly infinite dimensional spaces \mathcal{B}_1 and \mathcal{B}_2 can be made using semi-group theory for differential equations (see [1]). We do not pursue this here.

The Frechet derivative of F_B in (2.9) is

$$(2.16a) \quad \frac{\partial F_B}{\partial z} \equiv \begin{pmatrix} \left[\frac{d}{dt} \cdot - Tf_u(u, \lambda, \alpha) \right] & -f(u, \lambda, \alpha) \\ [E_0 - E_1] & 0 \\ p_u(u, T, \lambda, \alpha) & p_T(u, T, \lambda, \alpha) \end{pmatrix}$$

Here E_0 and E_1 are the evaluation operators given by

$$(2.16b) \quad E_t v(\cdot) = v(t).$$

It is convenient to define the two point boundary value operator

$$(2.17) \quad A \equiv \begin{pmatrix} \frac{d}{dt} \cdot - Tf_u(u, \lambda, \alpha) \cdot \\ E_0 - E_1 \end{pmatrix},$$

with $A : C_n^2[0,1] \rightarrow C_n^1[0,1] \times \mathbb{R}^n$. For the shooting formulation

(2.11) we find that

$$(2.18) \quad \frac{\partial F_S}{\partial z} \equiv \begin{pmatrix} I - V(1,0) & Y_T(1; \zeta, T, \lambda, \alpha) \\ p_u V(\tau, 0) & p_T \end{pmatrix}$$

Here we have introduced the fundamental solution matrix

$V(t, \tau; \zeta, T, \lambda, \alpha) \equiv V(t, \tau)$, defined as the solution of

$$(2.19) \quad \begin{aligned} a) \quad \frac{d}{dt} V(t, \tau) &= Tf_u(Y(t; \zeta, T, \lambda, \alpha), \lambda, \alpha) V(t, \tau) \\ b) \quad V(\tau, \tau) &= I \text{ for } \tau \in [0,1]. \end{aligned}$$

By differentiation of (2.10) with respect to the initial data ζ it follows from (2.19) that

$$(2.19c) \quad y_\zeta(t; \zeta, T, \lambda, \alpha) = V(t, 0; \zeta, T, \lambda, \alpha) .$$

The function $y(1; \zeta, T, \lambda, \alpha)$ from \mathbb{R}^{n+3} into \mathbb{R}^n is called the Poincaré return map. From (2.19c) we see that the circuit matrix $V(1, 0)$ is the linearization of the Poincaré return map. The eigenvalues of $V(1, 0)$ for a fixed point, $(\zeta_0, T_0, \lambda_0, \alpha_0)$, of (2.11) are called the Floquet multipliers for the orbit $y(t; \zeta_0, T_0, \lambda_0, \alpha_0)$ [3]. By differentiating (2.10) with respect to t and using the periodicity of y it follows that

$$(2.20) \quad V(1, 0) \frac{dy}{dt}(0; \zeta_0, T_0, \lambda_0, \alpha_0) = \frac{dy}{dt}(0) .$$

In particular, if $y(t; \zeta_0, T_0, \lambda_0, \alpha_0)$ is a nontrivial periodic orbit (i.e. $\frac{dy}{dt}(0) \neq 0$), then this orbit has a Floquet multiplier equal to one. As we show below, the other Floquet multipliers play an important role in the bifurcation analysis of periodic solution branches.

We can now state the main result of this section as:

Theorem 2.21. The following four statements are equivalent for any nontrivial orbit $(y_0(t), T_0, \lambda_0, \alpha_0)$:

A) $\frac{\partial F_B^0}{\partial (y, T)}$ is nonsingular;

B) $\frac{\partial F_S^0}{\partial (\zeta, T)}$ is nonsingular;

C) $\kappa=1$ is a simple eigenvalue of $V^0(1, 0)$ with eigen-

vector $\zeta_1 = \frac{dy_0}{dt}$,

and

$p_u^0 V^0(t, 0) \zeta_1 \neq 0$;

D) the boundary problem A^0 in (2.17) satisfies

i) $\text{Null}(A^0) = \text{span}\{u_1(t)\}$, $u_1 = \frac{dy_0}{dt}$,

ii) $f(y_0(t), \lambda_0, \alpha_0) \notin \text{Range}(A^0)$,

and

iii) $p_u^0 u_1(t) \neq 0$.

Here the superscript 0 denotes evaluation at the orbit $(y_0, T_0, \lambda_0, \alpha_0)$.

The proof of Theorem 2.21 is an easy extension of a basic result by Poincaré (see Theorem 2.1 on p.352 of [3]).

A useful result we employ in this proof and throughout our study is [15], [4]:

Lemma 2.22. Let \mathcal{B}_1 and \mathcal{B}_2 be Banach spaces and consider the linear operator $A: \mathcal{B}_1 \times \mathbb{R}^k \rightarrow \mathcal{B}_2 \times \mathbb{R}^k$ of the form

$$(2.22a) \quad A = \begin{pmatrix} A & B \\ C^* & D \end{pmatrix}$$

where

$$(2.22b) \quad A: \mathcal{B}_1 \rightarrow \mathcal{B}_2, B: \mathbb{R}^k \rightarrow \mathcal{B}_2, C^*: \mathcal{B}_1 \rightarrow \mathbb{R}^k, D: \mathbb{R}^k \rightarrow \mathbb{R}^k .$$

i) If A is nonsingular then A is nonsingular if and only if $(D - C^*A^{-1}B)$ is nonsingular.

ii) If A is singular with

$$(2.23) \quad \dim \text{Null}(A) = \text{codim Range}(A) = k$$

then A is nonsingular if and only if:

$$(2.24) \quad \text{a) } \dim \text{Range}(B) = k, \quad \text{b) } \text{Range}(B) \cap \text{Range}(A) = \{0\} ,$$

$$\text{c) } \dim \text{Range}(C^*) = k, \quad \text{d) } \text{Null}(A) \cap \text{Null}(C^*) = \{0\} .$$

iii) If A is singular with $\dim \text{Null}(A) > k$ then A is singular.

It is easily checked that the phase constraints p_I , p_{II} , and p_{III} all satisfy the conditions stated in C and D of Theorem 2.21. Since $V^0(1,0)$ must have one as an eigenvalue, we see that statements C and D represent "generic" or nondegeneracy conditions on the periodic orbit. Therefore we obtain our desired result that, at least generically, a nontrivial periodic orbit corresponds to a regular point of (2.9) or (2.11).

3. Fold Points and Paths. A singular point (z_0, β_0) of (2.1) is called a simple fold point if F_z^0 is a Fredholm operator with

$$(3.1) \quad \begin{aligned} \text{a) } \text{Null}(F_z^0) &= \text{span}\{\phi_1\} \neq \{0\} , \\ \text{b) } \text{Null}([F_z^0]^*) &= \text{span}\{\psi_1^*\} \neq \{0\} , \\ \text{c) } \psi_1^* F_\beta^0 &\neq 0 \text{ (i.e. } F_\beta^0 \notin \text{Range}(F_z^0)) . \end{aligned}$$

Such a fold point is said to be quadratic if

$$(3.1) \quad d) \quad \psi_1^* F_{zz}^0 \phi_1 \phi_1 \equiv a_F \neq 0 .$$

If we assume that $(z(s), \beta(s))$ is a smooth solution branch passing through (z_0, β_0) at $s = s_0$, then by formal differentiation of (2.1) we find that

$$(3.2) \quad F_z^0 \dot{z}(s_0) + F_\beta^0 \dot{\beta}(s_0) = 0 .$$

Here \cdot denotes $\frac{d}{ds}$. By applying ψ_1^* to (3.2), and using (3.1a,b,c) we obtain

$$(3.3a) \quad \dot{\beta}_0 = 0, \quad \dot{z}_0 = \kappa \phi_1, \quad \kappa \in \mathbb{R} .$$

It is convenient to normalize s so that $\kappa = 1$. Now from the second formal derivative of (2.1) we find

$$(3.3b) \quad \ddot{\beta}_0 = a_F / \psi_1^* F_\beta^0 \neq 0 .$$

A sketch of the solution curve $(z(s), \beta(s))$ near a quadratic fold point is given in Figure 3.

Figure 3 strongly suggests that the singularity at (z_0, β_0) may be effectively removed if we use a more appropriate continuation parameter. A general way to introduce a new parameter is to include a defining condition of the form

$$(3.4) \quad N(z, \beta, s) = 0, \quad N: \mathcal{B}_3 \times \mathbb{R}^2 \times \mathbb{R} .$$

Then we consider solving

$$(3.5) \quad G(z, \beta, s) \equiv \begin{pmatrix} F(z, \beta) \\ N(z, \beta, s) \end{pmatrix} = 0$$

for $(z(s), \beta(s))$. A numerically convenient form for (3.4) is the pseudo-arc length normalization introduced by Keller [15]. For a known starting solution (z_1, β_1) , along with a normalized tangent vector $(\dot{z}_1, \dot{\beta}_1)$, the pseudo-arc length normalization is (see Figure 3)

$$(3.6a) \quad N(z, \beta, s) \equiv \{ \dot{z}_1^* (z - z_1) + \dot{\beta}_1^* (\beta - \beta_1) \} - (s - s_1) = 0$$

Here \dot{z}_1^* is a dual vector for \dot{z}_1 , in particular $\dot{z}_1^* \dot{z}_1 > 0$. Alternatively, in cases for which $z \in \mathbb{R}^n$, the normalization

$$(3.6b) \quad N(z, \beta, s) = \{e_k^T \cdot (z - z_1, \beta - \beta_1)\} - (s - s_1) = 0$$

is often used [19]. Here $e_k \in \mathbb{R}^{n+1}$ is a unit vector chosen such that the k^{th} component of $(\dot{z}_1, \dot{\beta}_1)$ is relatively large. Normalization (3.6b) corresponds to "parameter switching", the new continuation parameter being the k^{th} component of $(z - z_1, \beta - \beta_1)$.

The fact that the singularity at a fold point (quadratic or not) is removed by the use of (3.5) is the content of Theorem 3.7. Let (z_1, β_1) be a regular or a fold point of (2.1). Suppose $(\dot{z}_1, \dot{\beta}_1)$ satisfies (3.2) and the normalization condition (3.4) satisfies

$$(3.8) \quad \begin{aligned} & a) \quad N(z_1, \beta_1, s_1) = 0 , \\ & b) \quad N_{(z, \beta)}^1(\dot{z}_1, \dot{\beta}_1) \neq 0, \quad N_{(z, \beta)}^1 = \frac{\partial N}{\partial (z, \beta)}(z_1, \beta_1, s_1) . \end{aligned}$$

Then (z_1, β_1, s_1) is a regular point of (3.5).

This result is easily proved using Lemma 2.22 (see [15]). Notice that (3.8) is always satisfied by the normalizations (3.6a,b). In practical calculations a continuation procedure is used to solve (3.5) for (z, β) at $s=s_2$. The constants $\dot{z}_1^*, \dot{\beta}_1^*(e_k^T)$ in 3.6a (3.6b) are re-evaluated every few continuation steps, thereby ensuring that regular and fold points of (2.1) are regular points of (3.5).

Fold points can also occur in branches of periodic solutions. In fact they can occur in two different ways, as shown in

Theorem 3.9. Let (z_0, β_0) be a solution of (2.11), $z_0 = (\zeta_0, T_0)$ say, and $\zeta_1 = \frac{d}{dt} y(0; \zeta_0, T_0) \lambda(\beta_0), \alpha(\beta_0) \neq 0$ (see (2.10,11)). Suppose that the phase condition is nondegenerate at (z_0, β_0) , that is

$$p_u^0 v^0(t, 0) \zeta_1 \neq 0 .$$

Then (z_0, β_0) is a fold point of (2.11) if and only if either

$$(3.10) \quad \begin{aligned} & i) \quad \dim \text{Null}(V^0(1, 0) - I) = 2 \text{ and} \\ & a) \quad \zeta_1 \notin \text{Range}(V^0(1, 0) - I) , \\ & b) \quad \frac{\partial y}{\partial \beta}^0(1) \notin \text{Range}((V^0(1, 0) - I) \mid \zeta_1) ; \end{aligned}$$

or ii) $\dim \text{Null}(V^0(1,0)-I) = 1$ and
 a) $\zeta_1 \in \text{Range}(V^0(1,0)-I)$,

(3.11)

$$\text{Here } \frac{\partial y^0}{\partial \beta} (1) = \frac{\partial y}{\partial \beta}(1; \zeta_0, T_0, \lambda(\beta_0), \alpha(\beta_0)).$$

b) $\frac{\partial y^0}{\partial \beta} (1) \notin \text{Range}(V^0(1,0)-I)$.

This result can be stated in terms of the boundary value formulation (2.9), in which case the two types of fold points have $\dim \text{Null}(A^0) = 2$ and $\dim \text{Null}(A^0) = 1$ respectively (see (2.17)). The proof of Theorem 3.9 follows from Lemma 2.22 and the proof of Theorem 2.21 (see [10]).

We turn next to paths of fold points, called simple "folds", which may form pieces of the boundaries of steady state manifolds M_S (see 1.3a) or periodic solution manifolds M_p (see 1.4a). Folds are also directly related to several important physical phenomena including hysteresis loops and sudden transitions such as the ignition or detonation of a chemical reactor. Furthermore a (dynamically) stable steady or periodic solution of (1.1) becomes unstable as a quadratic fold is traversed. In analyzing problems of the form (1.1) it is therefore very useful to be able to compute fold curves directly (see the discussion in Section 1). These computations can be done by applying a predictor-solver continuation method to a system of equations which characterizes fold points.

We first consider folds for steady solutions of (1.1). Several "defining" systems (see [2]) have been proposed, one of which is

$$(3.12) \quad \begin{aligned} \text{a) } F(z, \beta) &\equiv \begin{pmatrix} f(w, \lambda, \alpha) \\ f_w \phi \\ l^* \phi - 1 \end{pmatrix} = 0, \\ \text{b) } z &= (w, \phi, \lambda), \quad \beta = \alpha, \quad l^* \in \mathcal{B}_3^*, \quad \text{a constant}. \end{aligned}$$

This system was first proposed by Keener and Keller in [16] where it was used analytically. Several authors have considered its use in numerical calculations ([17], [21], [11], and [18]). For numerical purposes the key property of (3.12) is that:

$(z_0, \beta_0) = (w_0, \phi_0, \lambda_0, \alpha_0)$ is a regular point of (3.12a) if and only if $(w_0, \lambda_0, \alpha_0)$ is a quadratic fold point of (1.2) with $\alpha = \alpha_0$ fixed (and ℓ^* is chosen appropriately) [21]. This ensures that a standard continuation procedure can be used to follow a path of quadratic fold points. A modification of (3.12) that appears to be more efficient is discussed in [20].

An alternative defining system, due to Fier and Keller [8], [9] is:

$$(3.13) \quad \begin{aligned} a) \quad F(z, \beta) &\equiv \begin{pmatrix} f(w, \lambda, \alpha) \\ f_w^* \psi^* \\ \psi^* f_\eta - 1 \end{pmatrix} = 0, \\ b) \quad z &\equiv (w, \psi^*, \eta), \\ c) \quad \lambda &= \lambda(\eta, \beta), \quad \alpha = \alpha(\eta, \beta) \end{aligned}$$

Here (3.13c) is a linear change of coordinates such that lines of constant η and β are approximately tangent and normal (respectively) to the projection of the fold curve onto the (λ, α) plane (see Figure 4). In [9] it is shown that (z_0, β_0) is a regular point of (3.13a) if and only if, for fixed $\beta = \beta_0$, the corresponding point $(w_0, \lambda(\eta_0, \beta_0), \alpha(\eta_0, \beta_0))$ is a quadratic fold point of

$$(3.14) \quad f(w, \lambda(\eta, \beta_0), \alpha(\eta, \beta_0)) = 0.$$

This result justifies the use of predictor-solver continuation methods on (3.13) for the calculation of folds.

The power of the approach is clearly exhibited by the results of Fier and Keller [8], [9] who used (3.13) on the problem of the flow of a viscous incompressible fluid between two rotating disks at distance d apart. This yields the system:

$$(3.14) \quad \begin{aligned} a) \quad f''' &= R[ff'' + 4gg'] \quad 0 < z < 1; \\ b) \quad g'' &= R[fg' - gf'] \\ c) \quad f(0) &= f'(0) = 0, \quad g(0) = 1; \\ d) \quad f(1) &= f'(1) = 0, \quad g(1) = \gamma. \end{aligned}$$

Here the velocity field (u, v, w) in cylindrical coordinates (r, θ, z) is given by

$$(3.14) \quad u = -rf'(z)/2, \quad v = rg(z), \quad w = f(z).$$

The Reynold's number, R , and disk speed ratio, γ , are:

$$(3.16) \quad R \equiv \Omega_0 d^2 / v, \quad \gamma \equiv \Omega_1 / \Omega_0 .$$

It is easily shown that all such flows are obtained by considering only the parameter set: $|\gamma| \leq 1$, $R \geq 0$. Extensive calculations have been reported in [22]. We show in Figure 5 the folds in the (γ, R) plane obtained by using (3.13) on a finite difference approximation to (3.14). Many new features not previously known (i.e. "butterfly" configurations) become apparent. Further the new procedure is many orders of magnitude faster than the old way of finding the fold families. Discussions of the solutions to this problem will appear in [8], [9].

The computation of folds for periodic solutions can be done by using either F_B or F_S (eqn (2.9) or (2.11) respectively) in place of f in the defining systems (3.12) and (3.13). The regular points of the resulting systems are quadratic fold points in paths of periodic solutions, and again standard continuation techniques can be used to follow the folds.

In considering paths of solutions for (3.12a) or (3.13a) we are naturally lead to wonder whether or not these folds can themselves have fold points, which are perhaps more appropriately named "meta-fold" points. Indeed such meta-fold points do occur, the cusps in Figure 5 result from the projection onto the (γ, R) -plane of a fold with a quadratic meta-fold point (the meta-fold point projects onto the tip of the cusp). In [20] four types of meta-fold points are classified for a system similar to (3.12). Given more parameters in (1.1) paths of meta-fold points could be considered, and the process continued. A systematic analysis of this sort of iteration process is given in [11] for the special case of (1.2) with $\mathcal{B}_1 = \mathcal{B}_2 = \mathbb{R}$, and $\alpha \in \mathbb{R}^4$.

4. Bifurcation Points. A solution (z_0, β_0) of (2.1) is called a potential bifurcation point if for some $k \geq 1$

$$(4.1) \quad \begin{aligned} a) \quad & \dim \text{Null}(F_{z, \beta}^0) = k + 1 , \\ b) \quad & \text{codim Range}(F_{z, \beta}^0) = k . \end{aligned}$$

Here $F_{z,\beta}^0 \equiv \frac{\partial F}{\partial z, \beta}(z_0, \beta_0)$. If $k = 1$ ($k > 1$) then (z_0, β_0) is called a potential simple (multiple) bifurcation point. Notice that conditions (3.1a,b,c) for a simple fold point imply (4.1) with $k = 0$. It is convenient to define $\phi_i \in \mathcal{B}_3 \times \mathbb{R}$, $\psi_i^* \in \mathcal{B}_4^*$ such that

$$(4.2) \quad \begin{aligned} a) \quad & \text{span}\{\phi_0, \dots, \phi_k\} = \text{Null}(F_{z,\beta}^0), \\ b) \quad & \text{span}\{\psi_1^*, \dots, \psi_k^*\} = \text{Null}([F_{z,\beta}^0]^*). \end{aligned}$$

We consider conditions under which (2.1) has at least one smooth solution branch passing through (z_0, β_0) .

Suppose $(z(s), \beta(s))$ is a smooth branch with $(z(s_0), \beta(s_0)) = (z_0, \beta_0)$. Then from (3.2) we see that

$$(4.3) \quad (\dot{z}(s_0), \dot{\beta}(s_0)) = \phi(c) \equiv \sum_{i=0}^k c_i \phi_i \in \text{Null}(F_{z,\beta}^0),$$

for some $c \in \mathbb{R}^{k+1}$. From the second formal derivative of (2.1) with respect to s we find

$$(4.4) \quad Q(c) \equiv \begin{pmatrix} \psi_1^* F_{z,\beta}^0(z, \beta)(z, \beta) \phi(c) \phi(c) \\ \vdots \\ \psi_k^* F_{z,\beta}^0(z, \beta)(z, \beta) \phi(c) \phi(c) \\ c^T c - 1 \end{pmatrix} = 0 \in \mathbb{R}^{k+1},$$

where the last row in (4.4) is used to normalize s . Here $Q: \mathbb{R}^{k+1} \rightarrow \mathbb{R}^{k+1}$ is quadratic in c . Equation (4.4) is called the algebraic bifurcation equation (ABE) for (2.1) at (z_0, β_0) , and we have shown that it is a necessary condition for the existence of a smooth solution branch through the potential bifurcation point. Note that the tangent vector to this branch is $(\dot{z}(s_0), \dot{\beta}(s_0))$ and it is nonzero. A sufficient condition for the existence of such a branch is the isolated root condition:

$$(4.5) \quad \begin{aligned} a) \quad & Q(c^0) = 0, \\ b) \quad & \frac{\partial Q}{\partial c}(c^0) \text{ is a nonsingular } (k+1) \times (k+1) \text{ matrix.} \end{aligned}$$

Essentially this latter result is proven in [14].

For our purposes here we are interested in the behaviour of predictor-solver continuation methods as (z_0, β_0) is approached and, perhaps, crossed. We consider only continuation in an auxiliary parameter s , that is when equation (3.5) is used. (Here the use of (3.5) allows the following discussion to be given in a uniform manner independent of the precise direction of the particular tangent (4.3). However the arguments can be extended to β -continuation in a straightforward manner.) In particular we assume that c^0 satisfies (4.5), and the normalization function (3.4) satisfies

$$(4.6) \quad \begin{aligned} a) \quad & N(z_0, \beta_0, s_0) = 0 \\ b) \quad & N_z^0 \phi(c^0) \neq 0. \end{aligned}$$

So s is a suitable parameter for the straight line passing through (z_0, β_0) with tangent $\phi(c^0)$. Under these conditions on c^0 and N we have

Theorem 4.7. There exists a unique branch, $(z(s), \beta(s), s)$, such that for some $\delta > 0$

$$(4.8) \quad \begin{aligned} a) \quad & G(z(s), \beta(s), s) = 0 \text{ for } |s - s_0| < \delta, \\ b) \quad & (z(s_0), \beta(s_0)) = (z_0, \beta_0), \\ c) \quad & (\dot{z}(s_0), \dot{\beta}(s_0)) = \phi(c^0). \end{aligned}$$

Furthermore $\frac{\partial G}{\partial(z, \beta)}(z(s), \beta(s), s) \equiv G_{z\beta}(s)$ is singular for $s = s_0$ and for some constants $K_1, K_2 > 0$

$$(4.9) \quad \frac{K_1}{|s - s_0|} \leq \|G_{z, \beta}^{-1}(s)\| \leq \frac{K_2}{|s - s_0|} \text{ for } 0 < |s - s_0| \leq \delta.$$

For a proof of Theorem 4.7 see [6], or [5]. It is important to note that, unlike fold points, potential bifurcation points are necessarily singular points of (3.5). This fact forces us to question the convergence of the solver stage in a continuation step taken near a potential bifurcation point. One important result that gives at least a partial answer to this question is (see [4] for $k=1$, [5] for $k \geq 1$).

Theorem 4.10. Suppose (4.5,6) are satisfied, and let $(z(s), \beta(s), s)$ be the branch defined in Theorem 4.7. Let \mathbb{K}_0 be the

cone about $(z(s), \beta(s), s)$ given by

$$(4.10) \quad \mathbb{K}_0 = \{(\hat{z}, \hat{\beta}, s) \mid N(\hat{z}, \hat{\beta}, s) = 0, \|\hat{z} - z(s)\| + |\hat{\beta} - \beta(s)| \leq K |s - s_0|\}$$

Then there exist $\delta, K > 0$ such that, for $0 < |s - s_0| \leq \delta$, Newton's method applied to (3.5), for fixed s , converges quadratically to $(z(s), \beta(s), s)$ for each initial guess $(z^0, \beta^0, s) \in \mathbb{K}_0$. Similarly, there exist $\delta^1, K^1 > 0$ such that the Chord method applied to (3.5) converges linearly to $(z(s), \beta(s), s)$ for each initial guess $(z^0, \beta^0, s) \in \mathbb{K}_0$. (See Figure 6.)

The convergence cones described in Theorem 4.10 indicate that a predictor-solver continuation method can "jump" over singular points. To do this, the predictor based on a known point, such as $(z(s_1), \beta(s_1))$ in Figure 6, should be sufficiently accurate so that the predicted value, $(z^0(s_2), \beta^0(s_2))$, lies in the convergence cone. Under mild smoothness conditions on the solution branch, a linear (Euler) predictor is sufficient.

It is possible that two or more solution branches intersect at a bifurcation point (see below). For reasons discussed in Section 1 it is often important to compute these bifurcating branches. As the first step in switching branches, we consider techniques for the detection and accurate location of potential bifurcation points. In actual computations the spaces \mathcal{B}_3 and \mathcal{B}_4 are finite dimensional, say $\mathcal{B}_3 = \mathcal{B}_4 = \mathbb{R}^n$. From Theorem 4.7 we see that $G_{z, \beta}(s)$ is singular at a potential bifurcation point $(z(s_0), \beta(s_0), s_0)$ and so

$$(4.11) \quad \det[G_{z, \beta}(s)] = 0$$

at $s=s_0$. During a continuation process the sign of the $\det[G_{z, \beta}(s)]$ can be monitored. If a sign change is detected, say in going from $(z(s_1), \beta(s_1), s_1)$ to $(z(s_2), \beta(s_2), s_2)$ in Figure 6, then either bisection or false position could be used to locate a root of (4.11) (see [15]). Of course the left hand side of (4.11) may not change sign across a root, and therefore to detect and locate such points a more complicated algorithm must be used. The behaviour of $\det[G_{z, \beta}(s)]$ near $s=s_0$ is discussed in [6].

Lemma 4.12. Suppose (z_0, β_0) is a potential bifurcation point, and suppose (4.5, 6, 8) are satisfied. Then for some $k \neq 0$

$$(4.12) \quad \det[G_{z, \beta}(s)] = (s - s_0)^k \kappa + O((s - s_0)^{k+1}).$$

In particular, $\det[G_{z, \beta}(s)]$ changes sign when k is odd.

Once a potential bifurcation point (z_0, β_0, s_0) has been located, we attempt to determine the number of branches crossing (z_0, β_0) . If two or more distinct branches cross at (z_0, β_0) , we call this solution a bifurcation point. One simple result that guarantees the existence of another branch is

Theorem 4.13. Assume the hypothesis of Lemma 4.12 is satisfied. If k is odd then there is at least one other branch passing through (z_0, β_0) .

The proof of Theorem 4.13 is an easy application of degree theory, and is omitted.

In order to compute a bifurcating branch we first attempt to find an approximation for its tangent at the bifurcation point. Recall that the tangent to a smooth branch passing through (z_0, β_0) must satisfy (4.3, 4). If $c = c^1$ is a second isolated root of (4.4) then Theorem 4.10 shows that a predictor-solver continuation method can be used to compute the bifurcating branch having the tangent

$$(4.14) \quad (\dot{z}_0, \dot{\beta}_0) = \phi(c^1).$$

In fact, the continuation method needs only a sufficiently accurate approximation to the tangent. Based on this observation, four methods for switching branches at simple bifurcation points (i.e. $k=1$) are discussed in [15]. Similar methods can be used at multiple bifurcation points.

We turn now to the determination of paths or families of bifurcation points. For this purpose it is important to note that bifurcation points are not generic ("structurally stable") in one parameter problems (see [11]). Therefore, unless special symmetry properties are present, we cannot expect paths of bifurcation points to exist in two parameter problems. In fact in [20] it is shown that a simple bifurcation point is generically an isolated (quadratic) meta-fold point in a two parameter

problem. Below we present two common problems that possess sufficient symmetries for paths of bifurcation points to be generic, and consider the calculation of these paths.

We first consider bifurcation of steady solutions from a trivial solution. In particular, we set

$$(4.15) \quad \begin{aligned} a) \quad F(z, \beta) &\equiv f(w, \lambda, \alpha_0) = 0 \\ b) \quad z &\equiv w, \quad \beta \equiv \lambda. \end{aligned}$$

Equation 1.2 is assumed to have the trivial solution

$$(4.16) \quad f(0, \lambda, \alpha) = 0$$

for all $(\lambda, \alpha) \in \mathbb{R}^2$. Furthermore we assume that $f_w^0 \equiv f_w(0, \lambda_0, \alpha_0)$ satisfies

$$(4.17) \quad \begin{aligned} a) \quad \text{Null}(f_w^0) &= \text{span}\{\phi_1\} \neq \{0\}, \\ b) \quad \text{Null}([f_w^0]^*) &= \text{span}\{\psi_1^*\} \neq \{0\}. \end{aligned}$$

From (4.16) it follows that

$$(4.17c) \quad f_\lambda(0, \lambda, \alpha) = 0, \quad (\lambda, \alpha) \in \mathbb{R}^2,$$

and a short argument shows that the $F(z, \beta)$ defined in (4.15) has a potential simple bifurcation point $(z_0, \beta_0) = (0, \lambda_0)$. Without loss of generality we can take the vectors ϕ_0, ϕ_1, ψ_1^* in (4.2) to be

$$(4.18) \quad \begin{aligned} a) \quad \phi_0 &= (0, 1), \quad \phi_1 = (\phi_1, 0), \quad \phi_i \in \mathcal{B}_3 \times \mathbb{R}, \\ b) \quad \psi_1^* &= \psi_1^*. \end{aligned}$$

A straightforward calculation provides the algebraic bifurcation equation (see (4.4)):

$$(4.19a) \quad Q(c) \equiv \begin{pmatrix} a_f c_1^2 + 2b_f c_1 c_0 \\ c_0^2 + c_1^2 - 1 \end{pmatrix} = 0$$

where

$$(4.19b) \quad a_f = \psi_1^* f_w^0 \phi_1 \phi_1, \quad b_f = \psi_1^* f_w^0 \lambda \phi_1.$$

A tangent to the trivial solution branch of (4.15a) is

$$(\dot{z}_0, \dot{\beta}_0) = (0, 1) = \Phi(c^0)$$

for $c^0 = (1, 0)$. It is easily checked that c^0 satisfies the

isolated root condition (4.5) if and only if

$$(4.20) \quad b_f \neq 0.$$

Since $k=1$ in (4.1a) we see that Theorem 4.13 ensures the existence of a bifurcating branch of solutions whenever (4.20) is satisfied. For this reason (4.20), or equivalently the isolated root condition (4.5), is often called the bifurcation condition for simple bifurcation from a trivial solution [15]. A second isolated root $c=c^1$ of (4.19a) is easily found in terms of a_f , b_f . The tangent to the bifurcating branch given by (4.14) can then be used in the branch switching techniques discussed above, to generate another solution path (which contains nontrivial solutions).

Branches or families of such bifurcation points can be followed by varying α_0 in (4.15a). In particular we consider

$$(4.21) \quad F(z, \beta) \equiv \begin{pmatrix} f_w(0, \lambda, \alpha) \phi \\ \ell^* \phi - 1 \end{pmatrix} = 0$$

with $z \equiv (\phi, \lambda)$, $\beta \equiv \alpha$, and $\ell^* \in \mathcal{B}_3^*$ an appropriately chosen functional. The regular points of (4.21) are shown to correspond to simple bifurcation points of (1.2) in:

Theorem 4.22. Let $\lambda_0, \alpha_0, \phi_1, \psi_1^*$ be as in (4.17), and assume ℓ^* has been chosen such that $\ell^* \phi_1 = 1$. Then $(z_0, \beta_0) \equiv ((\phi_1, \lambda_0), \alpha_0)$ is a regular point of (4.21) if and only if the bifurcation condition (4.20) is satisfied.

This theorem can be proved using Lemma 2.22, we omit the details. We note that Theorem 4.22 justifies the use of continuation methods applied to (4.21) for the computation of paths of bifurcation points.

For a second example we consider bifurcation in branches of periodic solutions. For brevity we consider only the shooting formulation (2.11), though we note that all of our results can be phrased in terms of the alternate formulation (2.9). We begin with a general discussion of simple bifurcation in periodic solution branches, and then we consider the special case of period doubling bifurcations.

Potential simple bifurcation points of (2.11) occur in two forms, as discussed in the following corollary to Theorem 3.9.

Corollary 4.23. Suppose that the hypothesis of Theorem 3.9 is satisfied. Then (z_0, β_0) is a potential simple bifurcation point of (2.11) if and only if either:

- (4.24) i) $\dim \text{Null}(V^0(1,0)-I) = 2$ and
 a) $\zeta_1 \notin \text{Range}(V^0(1,0)-I)$,
 b) $\frac{\partial y^0}{\partial \beta}(1) \in \text{Range}((V^0(1,0)-I), \zeta_1)$;

or:

- (4.25) ii) $\dim \text{Null}(V^0(1,0)-I) = 1$ and
 a) $\zeta_1 \in \text{Range}(V^0(1,0)-I)$,
 b) $\frac{\partial y^0}{\partial \beta}(1) \in \text{Range}(V^0(1,0)-I)$.

For a proof of Corollary 4.23 see [10].

The theory for the existence of bifurcating periodic solution branches, and for the convergence of continuation methods near potential bifurcation points, can be obtained by applying the general results of the first part of this section to either (2.9) or (2.11). As an example we consider period doubling bifurcations. In particular, suppose $(z(s), \beta(s))$ is a branch of regular solutions of (2.11). Furthermore suppose that the circuit matrix $V(1,0;s)$ associated with this branch has a simple eigenvalue $\mu(s)$ such that

$$(4.26) \quad \mu(s_0) = -1, \quad \frac{d\mu}{ds}(s_0) \neq 0.$$

Finally we assume that $\kappa=1$ is an eigenvalue of $[V(1,0;s_0)]^2$ with both geometric and algebraic multiplicities equal to two. Then $(z(s_0), \beta(s_0))$ is called a period doubling bifurcation point. The form of the bifurcation equation is given in Theorem 4.27. Suppose (z_0, β_0) is a period doubling bifurcation point of 1.1, with $z_0 = (\zeta_0, T_0)$ say.

Then (\hat{z}_0, β_0) , with $\hat{z}_0 = (\zeta_0, 2T_0)$ is a bifurcation point of (2.11) with an algebraic bifurcation equation of the form

$$(4.27a) \quad Q(c) \equiv \begin{pmatrix} 2b_F c_1 c_0 + c_F c_0^2 \\ c_0^2 + c_1^2 - 1 \end{pmatrix} = 0 ,$$

with

$$(4.27b) \quad b_F \neq 0 .$$

The branch switching techniques discussed above can be used to calculate both branches near (\hat{z}_0, β_0) (since (4.27b) implies the roots of (4.27a) are isolated). Furthermore this bifurcation is generic in one parameter problems, and paths of doubling bifurcation points can exist in two parameter problems. One choice of $F(z, \beta)$, for which these period doubling bifurcation points are necessarily regular points of (2.1), is (see (2.11)):

$$(4.28a) \quad F(z, \beta) \equiv \begin{pmatrix} \zeta - y(1; \zeta, T, \lambda, \alpha) \\ p(y(t), T, \lambda, \alpha) \\ V(1, 0; \zeta, T, \lambda, \alpha) \eta + \eta \\ \ell^* \eta - 1 \end{pmatrix}$$

Here $\ell^* \in \mathbb{R}^n$ must be chosen appropriately, and:

$$(4.28b) \quad z = (\zeta, \eta, T, \lambda) \in \mathbb{R}^{2n+2}, \beta = \alpha .$$

5. Hopf Bifurcation. In this section we assume that $\mathcal{B}_1 = \mathcal{B}_2 = \mathbb{R}^n$ in problem (1.1), although the results stated below can be extended to infinite dimensional problems. Suppose that

$(w(\lambda), \lambda, \alpha_0)$ is a smooth branch of solutions of (1.2) for λ near λ_0 . Then we call $(w(\lambda_0), \lambda_0, \alpha_0)$ a potential Hopf bifurcation point if $f_w(\lambda_0) \equiv f_w(w(\lambda_0), \lambda_0, \alpha_0)$ has two simple purely imaginary eigenvalues, $\pm i\omega(\lambda_0) \neq 0$. Furthermore $(w(\lambda_0), \lambda_0, \alpha_0)$ is a Hopf bifurcation point if $f_w(\lambda)$ has two simple eigenvalues $\mu(\lambda) \pm i\omega(\lambda)$ with

$$(5.1) \quad a) \quad \mu(\lambda_0) = 0, \frac{d}{d\lambda} \mu(\lambda_0) \neq 0 ,$$

$$b) \quad \omega(\lambda_0) \neq 0, ik\omega(\lambda_0) \notin \sigma(f_w(\lambda_0)) \equiv \text{spectrum of } f_w(\lambda_0)$$

for $k = 0, \pm 2, \pm 3, \dots$.

Let $a_0, b_0, c_0, d_0 \in \mathbb{R}^n$ satisfy

$$(5.2) \quad \begin{aligned} a) \quad & f_w^0(a_0 + ib_0) = i\omega(\lambda_0)(a_0 + ib_0), \\ b) \quad & (c_0 - id_0)^T f_w^0 = i\omega(\lambda_0)(c_0 - id_0)^T, \\ c) \quad & \begin{pmatrix} c_0^T \\ d_0^T \end{pmatrix} (a_0 \ b_0) = I. \end{aligned}$$

Then we define the phase function $p(u, T)$ by

$$(5.3) \quad p(u, T) = \int_0^1 \{c_0^T \sin(2\pi\tau) + d_0^T \cos(2\pi\tau)\} u(\tau) d\tau + (t - T_0)v.$$

for $T_0 = 2\pi/\omega(\lambda_0)$ and some $v \in \mathbb{R}$. Using (5.2, 3) we will show that $(w(\lambda_0), T_0, \lambda_0, \alpha_0)$ is a multiple bifurcation point for (2.11). Similar results can be shown for (2.9), and for more general phase conditions (2.8c).

We consider (see (2.11)):

$$(5.4) \quad F(z, \beta) \equiv F_S(z, \beta) = 0$$

where $z = (\zeta, T)$, p is given by (5.3), and $\lambda = \beta$, $\alpha = \alpha_0$ in (2.11).

It is shown that $(z_0, \beta_0) \equiv ((w(\lambda_0), T_0), \lambda_0)$ is a potential multiple bifurcation point of (5.4) in

Lemma 5.5. Let $(w(\lambda_0), \lambda_0, \alpha_0)$ be a potential Hopf point. Then for F as in (5.4):

$$(5.6) \quad \begin{aligned} a) \quad & \text{Null}\{F_{z, \beta}^0\} = \text{span}\{\phi_0, \phi_1, \phi_2\}, \\ b) \quad & \text{Null}\{[F_z^0]^*\} = \text{span}\{\psi_1^*, \psi_2^*\}. \end{aligned}$$

Here we can take

$$(5.7) \quad \begin{aligned} a) \quad & \phi_0 = \begin{pmatrix} \phi_0 \\ 0 \\ 1 \end{pmatrix} \text{ for some } \phi_0 \in \mathbb{R}^n, \\ b) \quad & \phi_1 = \begin{pmatrix} a_0 \\ 0 \\ 0 \end{pmatrix}, \quad \phi_2 = \begin{pmatrix} -v b_0 \\ 1 \\ 0 \end{pmatrix} \in \mathbb{R}^{n+2}, \\ c) \quad & \psi_1^* = (c_0^T, 0), \quad \psi_2^* = (d_0^T, 0). \end{aligned}$$

The algebraic bifurcation equation for (5.4) at a Hopf point is given in

Theorem 5.8. In the notation of Lemma 5.5, the ABE for (5.4) can be written as

$$(5.8) \quad Q(c) \equiv \begin{pmatrix} 2\omega^0 c_1 c_2 + (T_0 \omega_\lambda^0) c_1 c_0 + (T_0 \mu_\lambda^0 v) c_2 c_0 \\ 2\omega^0 v c_2^2 - (T_0 \mu_\lambda^0) c_1 c_0 + (T_0 \omega_\lambda^0 v) c_2 c_0 \\ c_0^2 + c_1^2 + c_2^2 - 1 \end{pmatrix} = 0.$$

Here $\omega^0 = \omega(\lambda_0)$, $\omega_\lambda^0 = \omega_\lambda(\lambda_0)$, and $\mu_\lambda^0 = \mu_\lambda(\lambda_0)$.

Furthermore if $(w(\lambda_0), \lambda_0, \alpha_0)$ is a Hopf bifurcation point then

$$(5.9) \quad c^1 = (0, 1, 0)$$

is an isolated root, and the only other roots of (5.8) are

$$(5.10) \quad \text{a) } c^0 = (\zeta_0, 0, \zeta_1), \quad \zeta_0^2 + \zeta_1^2 = 1 \text{ for } v = 0;$$

or

$$(5.10) \quad \text{b) } c^0 = (1, 0, 0) \text{ for } v \neq 0.$$

From Theorem 4.10 together with Lemma 5.5 and Theorem 5.8 we see that there exists a smooth solution branch $(z(s), \beta(s))$ of (5.4) with

$$(5.11) \quad \text{a) } (z(s_0), \beta(s_0)) = ((w(\lambda_0), T_0), \lambda_0)$$

$$\text{b) } (\dot{z}(s_0), \dot{\beta}(s_0)) = \Phi(c^1) = ((a_0, 0), 0) \neq 0.$$

This is the desired periodic solution branch. Since the tangent given in (5.11b) is known, Theorem 4.10 insures that a continuation method using the pseudo-arc length normalization can be used to compute the bifurcating periodic solution branch.

Finally we consider the calculation of paths or families of Hopf bifurcation points. One system that can be used is

$$(5.12) \quad F(z, \beta) \equiv \begin{pmatrix} f(w, \lambda, \alpha) \\ f_w a + w b \\ f_w b - w a \\ l^* a - 1 \\ l^* b \end{pmatrix} = 0 \in \mathbb{R}^{2n+2}$$

Here $z \in (w, a, b, \omega, \lambda) \in \mathbb{R}^{3n+2}$, $\beta \in \alpha$, and $l^* \in \mathbb{R}^n$. For this system we have (see [10])

Theorem 5.13. Let $(w(\lambda_0), \lambda_0, \alpha_0)$ be a Hopf bifurcation point of

(1.1), and let $a_0, b_0, \omega(\lambda_0)$ be as in (5.2). Suppose ℓ^* is chosen so that $\ell^* a_0 = 1$ and $\ell^* b_0 = 0$. Then

$$(5.13) \quad (z_0, \beta_0) = ((\omega(\lambda_0), a_0, b_0, \omega(\lambda_0), \lambda_0), a_0)$$

is a regular point of (5.12).

Acknowledgements. This work was supported by the Department of Energy under Contract No. DE-AS03-76SF 00767 and by the Army Research Office under Contract No. DAAG-29-81-K-0107.

References

- [1] S. Agmon and L. Nirenberg, Properties of solutions of ordinary differential equations in Banach spaces; C.P.A.M. 16 (1963) 121-239.
- [2] W.J. Beyn, Defining equations for singular solutions and numerical applications; these proceedings.
- [3] E.A. Coddington and N. Levinson, Theory of Ordinary Differential Equations; McGraw-Hill, 1955, N.Y.
- [4] D.W. Decker and H.B. Keller, Path following near bifurcation; C.P.A.M. 34 (1981) 149-175.
- [5] D.W. Decker and A.D. Jepson, Convergence cones near bifurcation; in preparation.
- [6] J. Descloux, Two remarks on continuation procedures for solving some nonlinear equations, preprint.
- [7] E.J. Doedel, AUTO: A program for the automatic bifurcation analysis of autonomous systems, Cong. Num. 30 (1981) 265-284 (Proc. 10th Manitoba Conf. Num. Math. and Comp., Winnipeg, Canada).
- [8] J. Fier and H.B. Keller, Follow the folds, in preparation.
- [9] J. Fier, Thesis in Applied Math., Cal Tech, Pasadena, CA 1984.
- [10] A.D. Jepson, Numerical Hopf Bifurcation, Part II, Thesis in Applied Math., Cal Tech, Pasadena, CA 1981.
- [11] A.D. Jepson and A. Spence, Singular points and their computation, these proceedings.
- [12] A.D. Jepson and A. Spence, Folds in solutions of two parameter systems: Part I; Tech. Rept. NA-92-02, Comp. Sci. Dept., Stanford U., Stanford, CA 9182.
- [13] H.B. Keller, Numerical Solution of Two Point Boundary Value Problems, Regional Conf. Ser. in Appl. Math. 24, SIAM, Philadelphia, PA 1976.
- [14] H.B. Keller and W.F. Langford, Iterations, perturbations and multiplicities for nonlinear bifurcation problems, Arch. Rat. Mech. Anal. 48 (1972) 83-108.

- [15] H.B. Keller, Numerical solution of bifurcation and nonlinear eigenvalue problems; in: Applications of Bifurcation Theory (ed. P.H. Rabinowitz) Academic Press, New York, (1977) 359-384.
- [16] J.P. Keener and H.B. Keller, Perturbed bifurcation theory, Arch. Rat. Mech. Anal. 50 (1973) 159-175.
- [17] G. Moore and A. Spence, The calculation of turning points of nonlinear equations, SIAM J. Num. Anal. 17 (1980) 567-576.
- [18] W.C. Rheinboldt, Computation of critical boundaries on equilibrium manifolds, SIAM J. Num. Anal. 19, (1982) 653-669.
- [19] W.C. Rheinboldt and J.V. Burkardt, A locally parametrized continuation process, ACM TOMS 9 (1983) 215-235.
- [20] A. Spence and A.D. Jepson, Numerical computation of cusps, bifurcation points and isola formation points in two parameter problems; these proceedings.
- [21] A. Spence and B. Werner, Non-simple turning points and cusps, IMA J. Num. Anal. 2 (1982) 413-427.
- [22] R.-K. Szeto, The flow between rotating coaxial disks, Thesis in Applied Math., Cal Tech, Pasadena, CA 1978.
- [23] J.J. Todd, The Computation of Fixed Points and Applications, Lect. Notes in Economics and Math. Systems, 124, Springer-Verlag, Berlin, 1976.

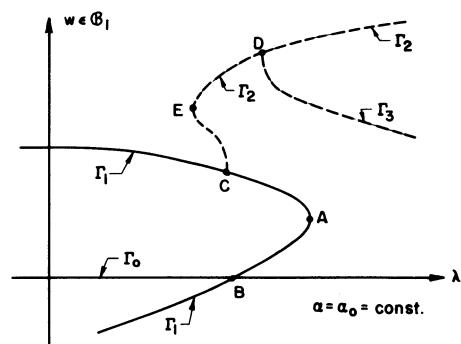


FIGURE 1

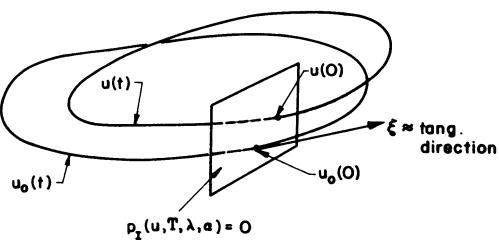


FIGURE 2

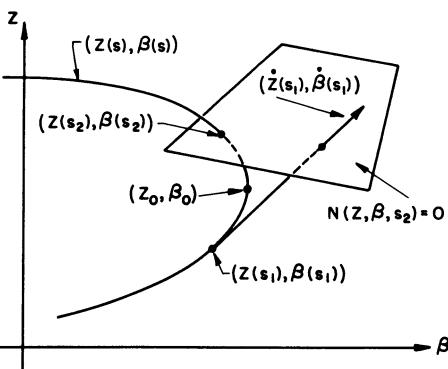


FIGURE 3

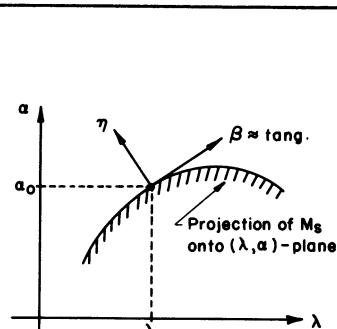
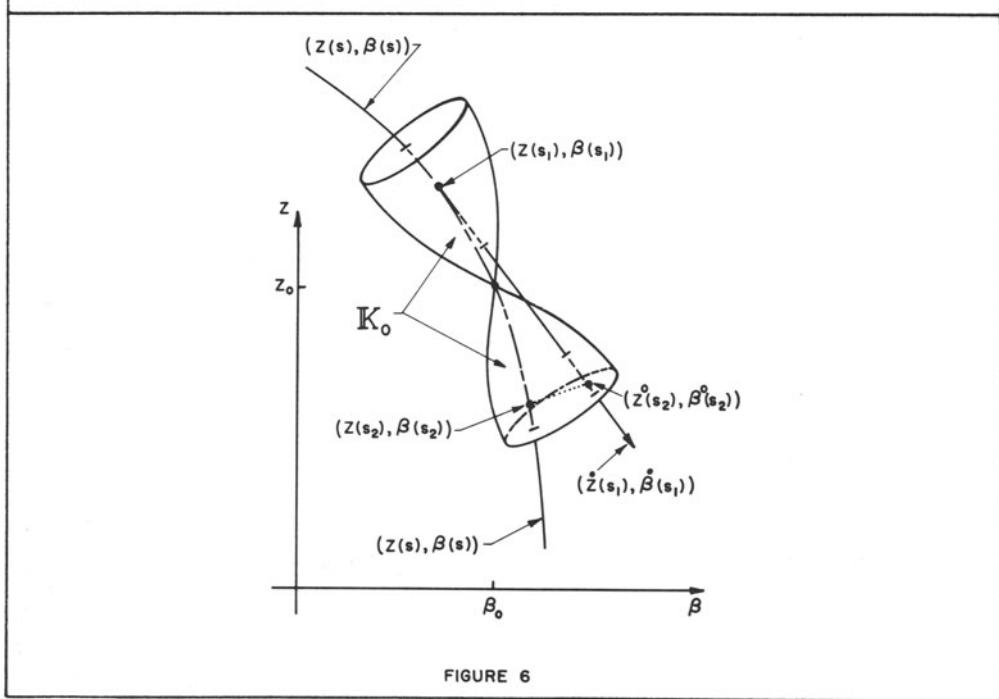
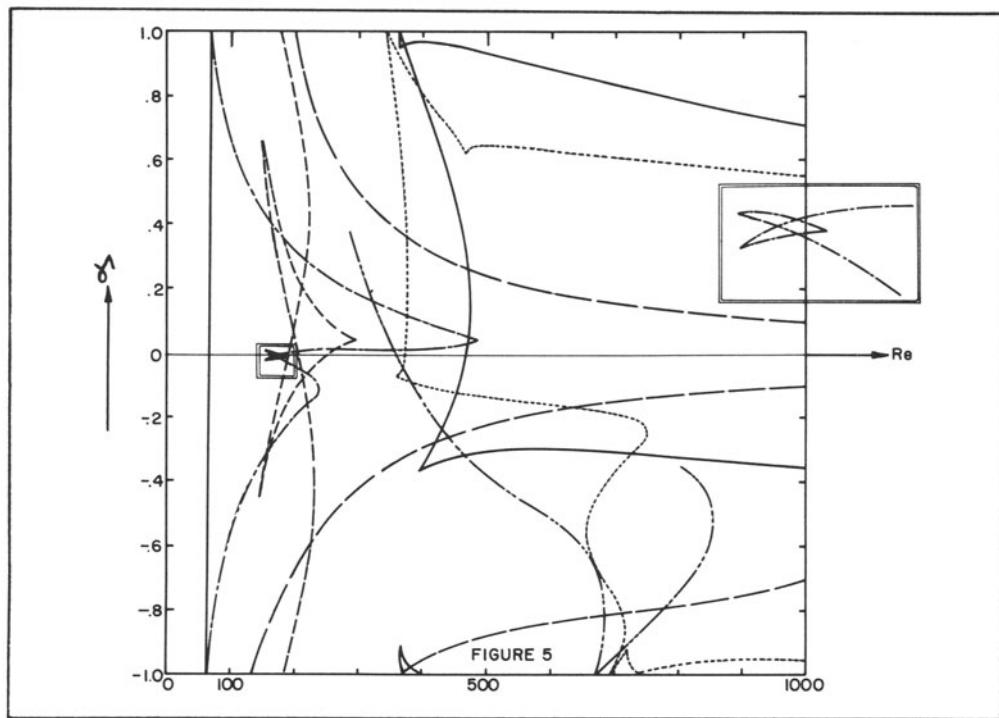


FIGURE 4



NUMERICAL DETERMINATION OF BIFURCATION POINTS IN STEADY STATE AND PERIODIC SOLUTIONS - NUMERICAL ALGORITHMS AND EXAMPLES

M. Kubíček and M. Holodniok

Numerical algorithms for determination of bifurcation points in steady state and periodic solutions are presented. Results of an evaluation of the limit points in a distributed parameter system where shooting method cannot be used are shown in the form of bifurcation diagram. Four direct iteration algorithms for an evaluation of complex (Hopf) bifurcation points in lumped parameter systems (ordinary differential equations) are described and applied to an example taken from the chemical reactor theory. An algorithm for determination of complex bifurcation points in distributed parameter systems (parabolic partial differential equations) is developed and results for a model of tubular reactor with axial dispersion are presented. Two algorithms for evaluation of period doubling bifurcation points in periodic solutions of ordinary differential equations are suggested. Results of the application to a model of two interconnected reaction cells are presented.

I. INTRODUCTION

Behaviour of dynamical systems can usually be described by the differential equation

$$\frac{dX}{dt} = F(X, A), \quad (1)$$

where $X \in B$, B is a suitably defined Banach space, and A is a parameter. We shall deal mainly with autonomous systems, i.e., systems where the time variable t does not appear in the right-hand side of (1) explicitly. It is customary in applied sciences to divide dynamical systems of the type (1) into two groups; the first one, where the space B is of a finite dimension, and the second one, where B is of an infinite dimension. When $B = \mathbb{R}^n$, we speak about lumped-parameter systems (LPS). The system (1) then has the form

$$\frac{dy_i}{dt} = f_i(y_1, y_2, \dots, y_n, \alpha), \quad i = 1, 2, \dots, n. \quad (2)$$

On the other hand, when B is an infinite-dimensional space, e.g. a space of functions $C^2_{[a, b]}$, we speak about the distributed parameter systems (DPS). Let us consider as an example a reaction-diffusion system of the form

$$\frac{\partial Y}{\partial t} = D \frac{\partial^2 Y}{\partial x^2} + f(x, \frac{\partial Y}{\partial x}, Y, \alpha), \quad (3)$$

where $Y = (y_1, \dots, y_n)^T$, $f = (f_1, \dots, f_n)^T$, D is a diagonal matrix with diagonal elements equal to D_1, \dots, D_n and α is a parameter. For the system of partial differential equations of parabolic type (3) the following boundary conditions will be considered:

$$\frac{\partial Y(0, t)}{\partial x} = 0, \quad (4a)$$

$$AY(1, t) + B \frac{\partial Y(1, t)}{\partial x} = 0. \quad (4b)$$

Here A and B are matrices, mostly diagonal.

The number of published papers dealing with the bifurcation theory or the bifurcation analysis of the system (1) is relatively high. Let us mention only several of them [1, 10, 26, 34]. On the other hand, the papers oriented numerically or algorithmically started to appear only ten years ago. A good review of numerical methods used in bifurcation theory was published by Mittelmann and Weber [29]. Numerical algorithms used for the bifurcation analysis of the LPS and DPS are presented in the book [22]. Most papers of the foregoing [28] and this proceedings are also devoted to numerical methods in bifurcation theory. We present here several powerful algorithms for the evaluation of bifurcation points and demonstrate their efficiency on practical examples, especially the algorithms for:
a) evaluation of real bifurcation points (especially limit

- (turning) points) of steady state solutions
- b) evaluation of complex (Hopf) bifurcation points of steady state solutions
 - c) evaluation of bifurcation points of periodic solutions in LPS, especially so called period doubling bifurcation points.

II. EVALUATION OF REAL BIFURCATION POINTS

Steady state solutions of LPS, Eq. (2), fulfil

$$f_i(y_1, y_2, \dots, y_n, \alpha) = 0, \quad i = 1, 2, \dots, n. \quad (5)$$

The necessary condition for a real bifurcation point to occur follows from the implicit function theorem:

$$f_{n+1}(y_1, y_2, \dots, y_n, \alpha) = \det J(y_1, \dots, y_n, \alpha) = 0. \quad (6)$$

Here $J = \{\partial f_i / \partial y_j\}$ is the Jacobi matrix and Eq. (6) also expresses that J has a zero eigenvalue. This eigenvalue moves with changing α along the real axis and crosses the imaginary axis through origin. The bifurcation point is therefore called a real bifurcation point.

Eqs. (5) and (6) form a system of $n+1$ nonlinear equations for $n+1$ unknowns y_1, \dots, y_n, α , coordinates of real bifurcation point in question. The Newton method can be used to solve the system [15]. The convergence is quadratic for the limit point while it is linear (if the Newton method converges at all) for the bifurcation point. To improve the convergence rate for bifurcation points, the equation

$$f_{n+2}(y_1, \dots, y_n, \alpha) = \det \bar{J}(y_1, \dots, y_n, \alpha) = 0, \quad (7)$$

can be added to the system. Here \bar{J} is a modified Jacobi matrix (it is formed from J by substituting $\partial f_i / \partial \alpha$ for, e.g., the last column). The Gauss-Newton method has been suggested [23] for solving this overdetermined system (5), (6), (7). Let us mention that generally $f_{n+2} \neq 0$ at the limit point.

There are other methods for evaluating real bifurcation points which do not require the evaluation of the determinant, see, e.g. [37]. A review of such methods can be found in [29].

Steady state solutions $y(x) = (y_1(x), \dots, y_n(x))$ of DPS, Eq.(3), fulfil (' = d/dx)

$$Dy'' + f(x, y, y', \alpha) = 0, \quad (8)$$

$$y'(0) = 0, \quad (9)$$

$$Ay(1) + By'(1) = 0. \quad (10)$$

We can use the shooting method to solve the nonlinear boundary value problem (8) (9) (10). If we choose initial conditions

$$y_i(0) = \gamma_i, \quad i = 1, 2, \dots, n, \quad (11)$$

and integrate (8), (9), (11) from $x = 0$ to $x = 1$, we obtain n residuals of (10)

$$\varphi(\gamma_1, \dots, \gamma_n, \alpha) = Ay(1) + By'(1) = 0, \quad (12)$$

depending on the choice of γ . For the system (12) we can use the same techniques as described above for LPS [15,23]. This is an advantage of the shooting method, where the DPS is transformed to a finite dimensional problem. However, the integration of relevant initial value problems may be unstable and then the shooting method cannot be used. The multiple shooting method can sometimes avoid this disadvantage. Another possibility is to use a finite-difference approximation for linearized problem, where a necessary condition for the limit point is properly included [19]. Let us demonstrate this technique on the difficult problem of the flow of an incompressible fluid between two rotating disks [25].

$$F'' = \sqrt{Re}HF' + Re(F^2 - G^2 + k), \quad (13a)$$

$$G'' = 2ReFG + \sqrt{Re}G'H, \quad (13b)$$

$$H' = -2\sqrt{Re}F, \quad (13c)$$

$$H(0) = F(0) = H(1) = F(1) = 0, \quad G(0) = 1, \quad G(1) = S. \quad (13d)$$

Here $' = d/dx$. Several authors [27,30,33] discovered multiple solutions for higher values of the Reynolds number Re and different values of the parameter $S \in [-1, 1]$ which characterizes the ratio of angular velocities of both disks. Here the value of k is to be determined. Dependences of the five solutions of the problem on the parameter Re are presented for $S = 0.8$ in [8]. The overall picture of solutions in dependence on the parameter S is for $Re = 625$ presented in [9] together with the resulting profiles. There exist six limit points in the solution diagram (dependence on S) for $Re = 625$. In the following we formulate the necessary conditions for occurrence of limit points, i.e. the points, where the number of solutions of (13) increases or decreases by two when the value of the parameter in question is changed. The approach is similar to that published in [19].

Let us denote the variables:

$$f = \frac{\partial F}{\partial k}, \quad g = \frac{\partial G}{\partial k}, \quad h = \frac{\partial H}{\partial k}, \quad s = \frac{ds}{dk}. \quad (14)$$

After differentiation of Eqs. (13) with respect to k we obtain the following system of differential equations

$$f'' = \sqrt{Re} [hF' + Hf'] + Re(2Ff - 2Gg + 1), \quad (15a)$$

$$g'' = 2Re [Fg + fG] + \sqrt{Re} [g'H + G'h], \quad (15b)$$

$$h' = -2\sqrt{Re}f, \quad (15c)$$

$$h(0) = f(0) = g(0) = 0, \quad f(1) = g(1) = 0. \quad (15d)$$

The condition

$$s = \frac{ds}{dk} = 0 \quad (16)$$

must be fulfilled at a branching (limit) point. The condition

(16) is already included in the boundary conditions (15d).

We have thus obtained a system of six differential equations (13) and (15) of the total order ten together with twelve boundary conditions (13d) and (15d). The problem seems to be overdetermined. However, we have two additional free parameters k and S , values of which have to be evaluated at a branching point. The problem is therefore well determined.

We used a finite difference method to solve the non-linear boundary value problem (13)-(15). Let us denote

$\Delta x = 1/n$, $x_i = i \Delta x$, $i = 0, 1, \dots, n$, $F_i \sim F(x_i)$, $G_i \sim G(x_i)$ etc. We can replace the derivatives F' ; F ; G' ; G ; f' ; f ; g' ; g by central three-point finite difference approximations at the points x_i , $i = 1, 2, \dots, n-1$. Thus, for instance, Eq. (15b) will be approximated by

$$\frac{g_{i-1} - 2g_i + g_{i+1}}{(\Delta x)^2} = 2\operatorname{Re}[F_i g_i + f_i G_i] + \sqrt{\operatorname{Re}} \left[\frac{g_{i+1} - g_{i-1}}{2 \Delta x} H_i + \frac{G_{i+1} - G_{i-1}}{2 \Delta x} h_i \right]. \quad (17a)$$

The derivatives H' and h' will be then replaced by two-point difference formulas centered between two neighbouring grid points, e.g., for Eq. (13c):

$$\frac{H_i - H_{i-1}}{\Delta x} = -\sqrt{\operatorname{Re}}(F_i + F_{i-1}), \quad i = 1, 2, \dots, n. \quad (17b)$$

We arrange the unknowns into a vector X :

$$X = (H_0, h_0, F_0, f_0, G_0, g_0, H_1, h_1, F_1, \dots, H_n, h_n, F_n, f_n, G_n, g_n, k, S)^T. \quad (18)$$

After an appropriate ordering of the finite difference equations (the approximations of boundary conditions are included), we obtain a system of $6(n-1)+2$ equations

$$\varphi(X) = 0 \quad (19)$$

with almost 15 diagonal occurrence matrix. The Newton method has been used to solve Eqs (19); the system of linear algebraic equations arising in each iteration was solved by special software for almost band matrices [16].

The number of iterations in the Newton method (and convergence at all) depends on an initial guess of the limit point. Initial guesses for $Re = 625$ were taken from the dependences of the solution on S [9]. The Newton method then converged in several (5-8) iterations to the chosen limit point. When we did not use this advantageous initial guess, the iteration process mostly diverged.

After evaluating the limit point for a particular value of the parameter Re we can use the results as an initial guess for the iteration for the value $Re + \Delta Re$ or $Re - \Delta Re$. In such a manner we can obtain the dependence of the limit points on the parameter Re . The results of such continuation are presented in the so called bifurcation diagram in the parametric plane " $Re - S$ ", cf. Fig. 1. The grid spacing $n = 100 - 400$ and double precision arithmetics (~ 14 decimal digits of accuracy) has been used to compute the results presented in the figure. Several parts of the curves in the figure were computed in an inverse way: instead of computing the bifurcation value of S for fixed Re the bifurcation value of Re for fixed S was calculated from the equations (19).

The curves in the bifurcation diagram divide the parametric plane " $Re - S$ " into several regions in which the number of solutions of the problem (13) remains constant (several regions are denoted by this number in the Fig. 1). However, Fig. 1 is not a complete bifurcation diagram; there exist additional curves in this figure, cf. paper by Keller and Szeto [14], where the bifurcation diagram has been constructed indirectly by other method.

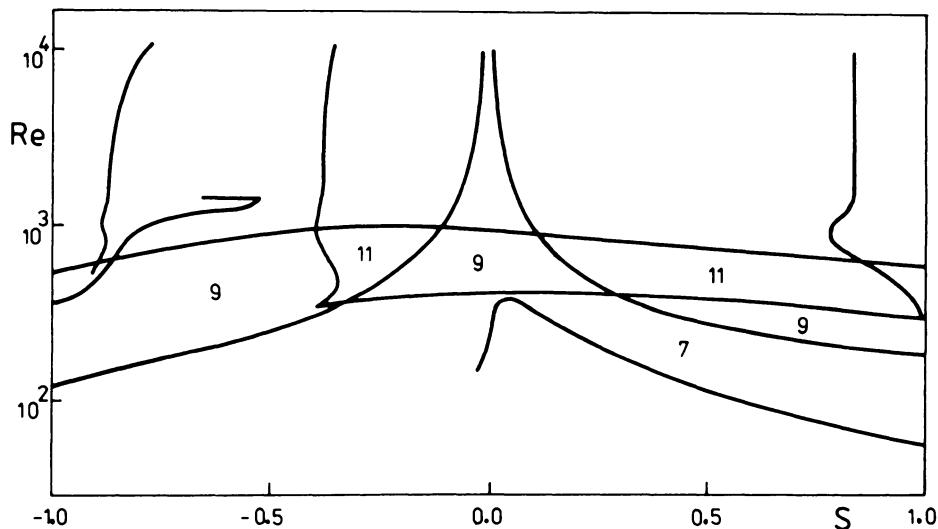


Fig. 1: Bifurcation diagram for the example (13).

III. EVALUATION OF COMPLEX (HOPF) BIFURCATION POINTS

A change of stability of a steady state solution with changing α can occur in two different ways. The first, one, where a real eigenvalue of the linearized problem (Jacobi matrix) crosses imaginary axis, was described above. The second case, so called complex or Hopf bifurcation, occurs when a pair of complex conjugate eigenvalues cross the imaginary axis. A branch of periodic solutions then bifurcates from the branch of steady-state solutions at this bifurcation point, see, e.g., [26] for detailed description. The branches of steady state solutions are connected with branches of periodic solutions at the bifurcation points. In the following we discuss four algorithms for determination of complex bifurcation points in LPS and an algorithm for DPS.

III.1 Algorithms for LPS

The Jacobi matrix J has a pair of complex conjugate pure imaginary eigenvalues at the complex bifurcation point, i.e., it holds

$$\operatorname{Re} \lambda_{1,2} = 0, \quad \operatorname{Im} \lambda_{1,2} = \pm i\sqrt{\omega}, \quad (20)$$

where $\lambda_{1,2}$ are roots of the characteristic polynomial

$$P(\lambda) = (-1)^n \det(J - \lambda I) = \lambda^n + a_1 \lambda^{n-1} + \dots + a_{n-1} \lambda + a_n. \quad (21)$$

The polynomial $P(\lambda)$ can be decomposed into

$$P(\lambda) = (\lambda^2 + \omega)(\lambda^{n-2} + p_1 \lambda^{n-3} + \dots + p_{n-3} \lambda + p_{n-2}) + A\lambda + B. \quad (22)$$

If

$$\begin{aligned} f_{n+1}(y_1, \dots, y_n, \alpha, \omega) &= A(y_1, \dots, y_n, \alpha, \omega) = 0, \\ f_{n+2}(y_1, \dots, y_n, \alpha, \omega) &= B(y_1, \dots, y_n, \alpha, \omega) = 0, \end{aligned} \quad (23)$$

then resulting values of $y_1, \dots, y_n, \alpha, \omega > 0$ determine a complex bifurcation point. The Newton method is used for solving the system of $n+2$ nonlinear equations (5)(23). Detailed description of the algorithm is in [17], a modified algorithm that needs an evaluation of characteristic polynomial of J^2 is presented in [18]. Resulting dimension of the system (again solved by the Newton method) is $n+2$.

Third algorithm [6] does not require evaluation of the characteristic polynomial; similar algorithm has been suggested in [12]. Let us write eigenvalue (20) as, $\lambda = 0 + i\sqrt{\omega} = is$ and let us take a complex eigenvector $u = v + iw$. The equation $Ju = \lambda u$ can be then written as

$$Jv + sw = 0, \quad Jw - sv = 0. \quad (24)$$

Let us consider now Eqs. (24) in the form

$$f_i(y_1, \dots, y_n, \alpha, s, v_1, \dots, v_n, w_1, \dots, w_n) = 0, \quad i = n+1, \dots, 3n. \quad (25)$$

Two components of the vectors v and w can be chosen arbitrarily because every complex multiple of the eigenvector u is

also an eigenvector of J belonging to the eigenvalue $\lambda = i\pi$. The number of unknowns in (25) is therefore equal to $3n$. The system of $3n$ nonlinear equations (5)(25) is then solved by the Newton method. Fourth algorithm lowers dimensionality of the problem to be solved. It follows from (24) that

$$J^2 v + s^2 v = 0. \quad (26)$$

Consider (26) in the form

$$f_i(y_1, \dots, y_n, \alpha, s^2, v_1, \dots, v_n) = 0, \quad i = n+1, \dots, 2n. \quad (27)$$

Two components of the vector v are again chosen arbitrarily so that we have $2n$ equations (5)(27) for $2n$ unknowns. The Newton method is used to solve this system.

Example

Let us consider two interconnected well mixed cells where chemical reaction takes place. As a model chemical reaction the so called Brusselator [31] schema has been chosen. The governing equations have the following form [36] $n=4$, $' = d/dt$:

$$\begin{aligned} y'_1 &= A - (B+1)y_1 + y_1^2 y_2 + \alpha(y_3 - y_1), \\ y'_2 &= By_1 - y_1^2 y_2 + \frac{\alpha}{\rho}(y_4 - y_2), \\ y'_3 &= A - (B+1)y_3 + y_3^2 y_4 + \alpha(y_1 - y_3), \\ y'_4 &= By_3 - y_3^2 y_4 + \frac{\alpha}{\varphi}(y_2 - y_4). \end{aligned} \quad (28)$$

Here A , B , φ and α are parameters of the problem.

Four resulting complex bifurcation points obtained by the algorithms are presented in table 1. Let us note that each solution presented in the table represents in fact two solutions due to the symmetry $y_1 \leftrightarrow y_3$, $y_2 \leftrightarrow y_4$ in the equations (28). Domains of attraction of individual complex bifurcation points are examined in [6] for the model (28) and another model from chemical reactor theory. As follows from the comparison the second algorithm seems to be the worst one. Re-

maining three algorithms are comparable in their performance.

Table 1: Resulting complex bifurcation points for the example
(28), $A = 2$, $B = 6$, $\varphi = 0.1$

i=	1	2	3	4	α	$s = \sqrt{\omega}$
$y_i =$	2.3525	2.6197	1.6474	3.5008		
$v_i =$	9.1665	-60.003	13.862	14.492	0.04349	1.8461
$w_i =$	28.363	-9.3463	-12.431	29.546		
$y_i =$	3.2985	1.9472	0.7015	5.7194		
$v_i =$	38.596	-19.636	-54.584	-4.3262	0.03697	0.3242
$w_i =$	-26.293	-17.891	4.1769	-42.368		
$y_i =$	0.8597	2.59 2	3.1403	2.2395		
$v_i =$	-12.045	37.938	-15.087	51.320	0.9220	2.2101
$w_i =$	4.3113	-34.724	-17.891	-29.136		

III.2 Algorithm for DPS

Two approaches have in principle been used in the literature to determine complex (Hopf) bifurcation point in DPS. The first approach consists in a trial and error technique using test dynamical simulation (numerical) for a sequence of steady state solutions from one branch of solutions. The second approach is based on a "semidiscretization", i.e., obtaining the finite-dimensional approximation of the original dynamic problem (3). The "method of lines" is often used to transform (3) into a system of ordinary differential equations by substituting finite differences for spatial derivatives, e.g. [5,11,32]. In the following a new direct iteration method for computation of the complex bifurcation points is developed [20].

Let us consider the DPS system (3) with boundary conditions (4). Steady state solution fulfills Eqs. (8)(9)(10). Local stability of the steady state solution is determined by the eigenvalues λ of a linearized problem

$$Du'' + \frac{\partial f(x, y', y, \alpha)}{\partial y'} u' + \frac{\partial f(x, y', y, \alpha)}{\partial y} u = \lambda u, \quad (29)$$

$$u(0) = 0, \quad Au(1) + Bu'(1) = 0. \quad (30)$$

At the complex bifurcation point $\lambda =$ is and the eigenfunction $u(x) = p(x) + iq(x)$. Real and imaginary parts of (29) and (30) are then in the form

$$Dp'' + f_{y'}(x, y', y, \alpha)p' + f_y(x, y', y, \alpha)p = -sq, \quad (31a)$$

$$Dq'' + f_{y'}(x, y', y, \alpha)q' + f_y(x, y', y, \alpha)q = sp, \quad (31b)$$

$$p'(0) = 0, \quad q'(0) = 0, \quad (32a)$$

$$Ap(1) + Bp'(1) = 0, \quad Aq(1) + Bq'(1) = 0. \quad (32b)$$

Here the subscripts y and y' denote partial differentiation. The eigenfunction u is determined except for a (non-zero) complex multiplication constant. We can therefore choose two values on the profiles $p(x)$ and $q(x)$ almost arbitrarily (they should be non-zero).

There are in principle two possibilities how to solve resulting system (8)-(10), (31)-(32) for unknowns $(y(x), p(x), q(x), \alpha, s)$. The first one consists in using finite difference methods. The second possibility is to use a shooting method, where we guess the following $3n$ initial conditions

$$y_j(0) = \gamma_j, \quad p_j(0) = \gamma_{n+j}, \quad q_j(0) = \gamma_{2n+j}, \quad j = 1, 2, \dots, n, \quad (33)$$

and the values of α and s . Then we can integrate (8)-(31) as the initial value problem starting from $x = 0$, where the whole set of initial conditions (9)(32a)(33) is known, to $x = 1$. To satisfy remaining boundary conditions, i.e. (10)(32b), the values of γ, α, s should be chosen in such a way that they fulfil a set of $3n$ nonlinear equations

$$\begin{aligned}
 Ay(1) + By'(1) &= 0, \\
 Ap(1) + Bp'(1) &= 0, \\
 Aq(+) + Bq'(1) &= 0.
 \end{aligned} \tag{34}$$

The number of unknowns γ , α , s is $3n+2$ so that the system (34) seems to be underdetermined. However, we can choose two values γ_k , γ_m , $n < k, m \leq 3n$ arbitrarily (see above). Then the system (34) is of a square type of the order $3n$. The Newton method can be used for solution of the system. Jacobian matrix of the system (34) can be computed either by using variational differential equations or by a finite difference approximation. It seems that the latter approach is more suitable because it requires less programming work and was therefore used in the following example.

Example

Axial dispersion of mass and heat in a tubular non-adiabatic reactor where a simple first order chemical reaction takes place can be described by a set of two second order differential equations [4,39]

$$\frac{\partial y_1}{\partial t} = \frac{1}{Pe_M} \frac{\partial^2 y_1}{\partial x^2} + \frac{\partial y_1}{\partial x} + \alpha (1-y_1) \exp(y_2/(1+y_2/\gamma)), \tag{35a}$$

$$\text{Le} \frac{\partial y_2}{\partial t} = \frac{1}{Pe_H} \frac{\partial^2 y_2}{\partial x^2} + \frac{\partial y_2}{\partial x} + \alpha B (1-y_1) \exp(y_2/(1+y_2/\gamma)) - \beta (y_2 - T_c), \tag{35b}$$

with boundary conditions

$$x=0: \quad \frac{\partial y_1}{\partial x} = 0, \quad \frac{\partial y_2}{\partial x} = 0, \tag{36a}$$

$$x=1: \quad \frac{\partial y_1}{\partial x} + Pe_M y_1 = 0, \quad \frac{\partial y_2}{\partial x} + Pe_H y_2 = 0. \tag{36b}$$

Here y_1 and y_2 are conversion and temperature, respectively, Pe_M , Pe_H , Le , β , B , T_c , γ and α are parameters.

Resulting complex bifurcation points are for one set of values of parameters presented in table 2. More detailed results for other parameters and also for another chemical engineering examples are described in [20] together with detailed explanation of the numerical realization.

Table 2: Resulting complex bifurcation points for the example (35)(36). $\gamma = 20$, $P_{eH} = P_{eM} = 3$, $Le = 1.263$, $T_c = 0$, $\beta = 3$, $B = 15$. Two initial conditions, $\gamma_3 = p_1(0) = 1$, $\gamma_4 = p_2(0) = 1$, were chosen fixed

$\gamma_1 = y_1(0)$	$\gamma_2 = y_2(0)$	$\gamma_5 = q_1(0)$	$\gamma_6 = q_2(0)$	α	s
0.9523	2.5124	-0.4453	7.6768	0.1842	2.1921
0.7033	3.3867	1.4356	10.4366	0.1637	1.0375
0.7886	3.5009	0.5429	6.9446	0.1707	1.1742
0.7586	3.5030	0.8646	8.0334	0.1678	1.1714

IV. EVALUATION OF BIFURCATION POINTS OF PERIODIC SOLUTIONS IN LPS

Let us consider an autonomous system (2). A periodic solution with the period T fulfills

$$y_i(t+T) = y_i(t), \quad i = 1, 2, \dots, n. \quad (37)$$

Setting $t = Tz$ we obtain a system of equations

$$\frac{dy_i}{dz} = T f_i(y_1, \dots, y_n, \alpha), \quad i = 1, 2, \dots, n, \quad (38)$$

and mixed boundary conditions (37) are in the form

$$y_i(1) - y_i(0) = 0, \quad i = 1, 2, \dots, n. \quad (39)$$

Considering shooting method we choose initial conditions

$$y_i(0) = x_i, \quad i = 1, 2, \dots, n, \quad (40)$$

and the values of the period T and the parameter α . Then the

system (38) can be integrated starting from $z=0$ to $z=1$. The values of the solution at $z=1$ are obtained from the integration as

$$y_i(1) = \varphi_i(x_1, \dots, x_n, T, \alpha), \quad i = 1, 2, \dots, n. \quad (41)$$

They are dependent on the choice of $x_1, \dots, x_n, T, \alpha$. The relation (39) has to hold for any periodic solution; thus we have to satisfy n equations

$$F_i(x_1, \dots, x_n, T, \alpha) = \varphi_i(x_1, \dots, x_n, T, \alpha) - x_i = 0, \\ i = 1, 2, \dots, n, \quad (42)$$

with $r+1$ unknowns x_1, \dots, x_n, T and one parameter α . To obtain a periodic solution for fixed α we therefore have to fix one variable. It cannot be T because the solution of (42) exists only for discrete (and apriori unknown) values of T . Let us fix x_k for some k . Our choice will be successful if the chosen value actually exists on the trajectory of the k -th component of the wanted periodic solution $y_k(z)$, $z \in [0, 1]$, cf. [7].

The stability of the computed periodic solution can be determined on the basis of characteristic (Floquet) multipliers, e.g. [10], i.e., eigenvalues λ of the monodromy matrix

$$P = \left\{ \frac{\partial \varphi_i}{\partial x_j} \right\}. \quad (43)$$

Elements of the monodromy matrix (and also $\partial F_i / \partial x_j$ for (42)) can be evaluated on the basis of variational differential equations for variational variables

$$p_{ij}(z) = \partial y_i / \partial x_j, \quad i, j = 1, 2, \dots, n. \quad (44)$$

These differential equations are obtained by differentiation of (38) with respect to x_j , i.e.,

$$\frac{dp_{ij}}{dz} = T \sum_{m=1}^n \frac{\partial f_i}{\partial y_m} p_{mj}, \quad p_{ij}(0) = \delta_{ij}, \quad (45)$$

(δ_{ij} is the Kronecker delta). For the elements of the monodromy matrix we then have

$$B = \{ p_{ij}(1) \} . \quad (46)$$

Let us consider a branch of periodic solutions in dependence on the value of parameter α . Such dependence can be obtained by using a continuation technique, e.g., DERPER [7]. The stability of periodic solutions changes at so called bifurcation value of the parameter α . It is the value, where some characteristic multiplier of the corresponding periodic solutions crosses unit circle with changing α . This multiplier can be either +1 or -1 or imaginary. The first case corresponds to the so called limit (turning) points or bifurcation (crossover, symmetry breaking) points on the dependences of periodic solutions on the parameter. Numerical method for determination of such points are subject of, e.g. [38]. The third case (λ imaginary, $|\lambda| = 1$) indicates mostly bifurcation to an invariant torus.

We deal here with the second case, i.e., with the so called double period (period doubling) bifurcation points (sometimes called also Brunovsky bifurcation). The characteristic multiplier goes through -1 at such a point and a new branch of periodic solutions bifurcates with a period which is approximately (asymptotically) doubled in comparison with the period on the original branch. The situation is schematically sketched on Fig. 2, more detailed explanation can be found, e.g. in [10].

The goal of this section is to construct an algorithm for direct (iterative) determination of such period doubling bifurcation points. In fact, we want to have a periodic solution which has -1 as the characteristic multiplier. In the following we describe briefly two iterative algorithms constructed for this purpose.

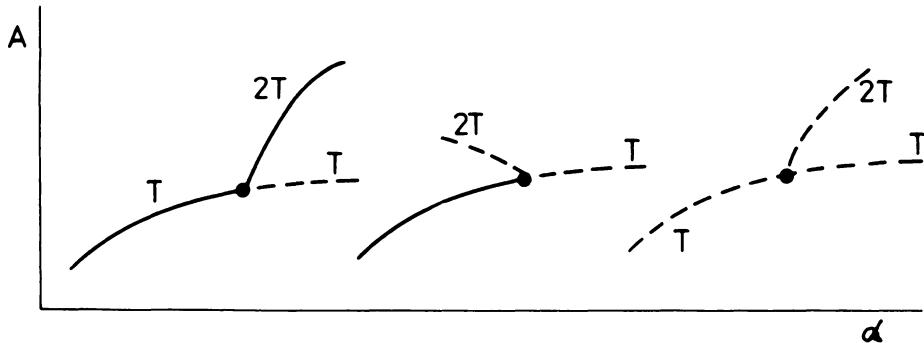


Fig. 2: Schematic representation of period doubling bifurcation point.

— stable, - - - unstable periodic solutions.

A - amplitude or other representative of the periodic solution.

Algorithm 1

Let the characteristic polynomial of the monodromy matrix B be $P(\lambda) = (-1)^n \det(B - \lambda I)$, i.e., the characteristic equation is

$$P(\lambda) = \lambda^n + a_1 \lambda^{n-1} + \dots + a_{n-1} \lambda + a_n = 0. \quad (47)$$

Coefficients a_j can be evaluated by a standard software [40], e.g. by the Krylov method. $\lambda = -1$ is a root of (47) if

$$F_{n+1}(x_1, \dots, x_n, T, \alpha) = 1 + \sum_{i=1}^n (-1)^i a_i = 0. \quad (48)$$

As a result we have $n+1$ nonlinear (algebraic) equations (42) and (48) for $n+1$ unknowns $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n, T, \alpha$. Newton method is used to solve this system. The first n rows of the Jacobi matrix can be evaluated on the basis of the variational variables, i.e.,

$$\frac{\partial F_i}{\partial x_j} = p_{ij}(1) - \delta_{ij}, \quad \frac{\partial F_i}{\partial T} = f_i(y(1), \alpha), \quad \frac{\partial F_i}{\partial \alpha} = q_i(1) \\ i = 1, \dots, n, \quad (49)$$

where $q_i = \partial y_i / \partial \alpha$ fulfil variational equations

$$\frac{dq_i}{dz} = T \sum_{m=1}^n \frac{\partial f_i}{\partial y_m} q_m + T \frac{\partial f_i}{\partial \alpha}, \quad q_i(0) = 0, \\ i = 1, \dots, n. \quad (50)$$

Elements of the last row of the Jacobi matrix, $\partial F_{n+1} / \partial x_j$, $\partial F_{n+1} / \partial T$, $\partial F_{n+1} / \partial \alpha$, can be obtained by using difference formulas. The "analytical" way using variational variables is also possible, however, for higher n it becomes cumbersome.

Algorithm 2

The monodromy matrix B has -1 as an eigenvalue at period doubling bifurcation point, i.e., there exists a non-zero vector $v = (v_1, \dots, v_n)^T$ such that the following system of n equations is fulfilled:

$$(B + I)v = 0. \quad (51)$$

Each non-zero multiple of the vector v is also a solution of (51); we can therefore fix one component of v for whole computation process, e.g.,

$$v_s = 1 \quad (52)$$

for $s \in [1, n]$. Thus we obtain $2n$ equations (42)(51) for $2n$ unknowns $x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n, T, \alpha, v_1, \dots, v_{s-1}, v_{s+1}, \dots, v_n$. The Newton method can be again used to solve this $2n$ by $2n$ system. A part of the Jacobi matrix can be evaluated on the basis of the variational variables, for the remaining part either variational equations for the second order variational variables (e.g. $\partial^2 y_i / \partial x_j \partial x_m$) have to be derived or difference formulas can be used.

Example

Let us consider the model described by Eqs. (28). The course of the Newton iteration process is presented in table 3. Initial guess for the Newton method originated approximately

Table 3: Computation of period doubling bifurcation points for example (28). Course of iterations for the Algorithm 1
 $A=2$, $B=5.9$, $\varphi=0.1$, $k=1$, $x_k=2$, initial guess generated randomly

iteration	x_1	x_2	x_3	x_4	T	α
0	2	3.5229	1.6059	5.5378	11.7588	1.23635
1	2	5.8629	1.6993	5.9794	12.1162	1.24259
2	2	5.9621	1.7065	6.0825	11.9041	1.22265
3	2	5.6563	1.7239	5.7474	11.8191	1.22611
4	2	5.7228	1.7116	5.8312	11.6530	1.22207
7	2	5.7254	1.6980	5.8389	11.5942	1.22382
8	2	5.7254	1.6980	5.8389	11.5942	1.22382

from the results of the continuation of periodic solutions in dependence on the parameter α . Results of one of such continuations (obtained by the DERPER algorithm described in [7]) are presented in Fig. 3. Four period doubling bifurcation points

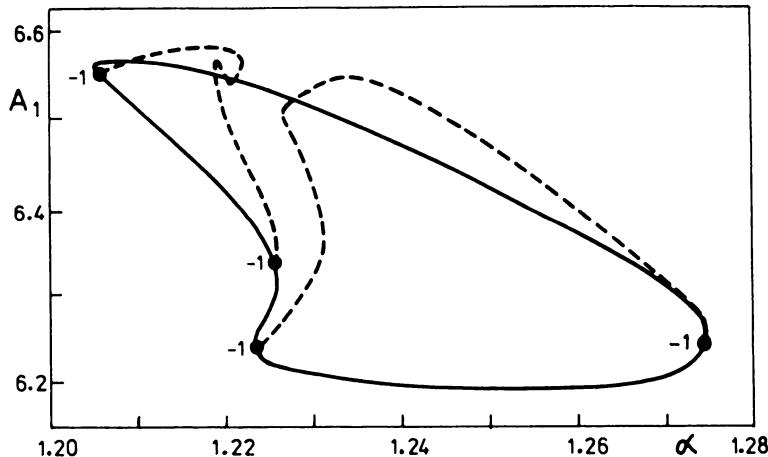


Fig. 3: Solution diagram of periodic solutions of (28) in dependence on the parameter α . $A=2$, $B=5.9$, $\varphi=0.1$.
 A_1 - amplitude of y_1 . Points of period doubling bifurcations are denoted by -1 .

- isolated dependence of periodic solutions with the period $T \sim 11-13$,
- - - branches of periodic solutions with the doubled period $T \sim 22-26$.

exist on the isolated and closed dependence of periodic solutions on α , they are presented in table 4. Every bifurcation

Table 4: Period doubling bifurcation points of the problem (28), $A=2$, $B=5.9$, $\varrho=0.1$, $k=1$, $x_k=2$.

x_2	x_3	x_4	T	α
3.12258	0.86612	3.26192	11.59421	1.22382
3.02534	0.90140	3.14959	12.62766	1.27471
3.25999	0.86603	3.41344	11.47081	1.22556
3.13884	0.85654	3.28176	11.83646	1.20614

point from the table 4 can be obtained four times because the prescribed (choice) $y_k(z) = x_k = 2$, $k=1$, is fulfilled four times in the interval $z \in [0,1]$ for every periodic solution in question. The course of the profile $y_1(z)$ for the periodic solution corresponding to the first bifurcation point in table 4 is presented in figure 4.

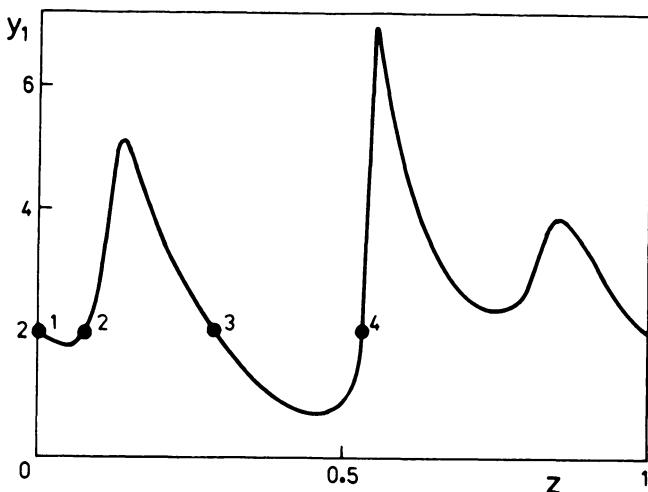


Fig. 4: Course of $y_1(z)$ for first bifurcation point in table 4. Numbering of points $y_1(z) = 2$ is used in table 5.

Four different solutions corresponding to this bifurcation point are presented in table 5. Let us note that values of T and α remain constant for all four solutions.

Table 5: Four different solutions obtained for the first bifurcation point in table 2, $T=11.59421$, $\alpha = 1.22382$

x_2	x_3	x_4	point N° in fig. 5
3.12258	0.86612	3.26192	1
3.68848	0.87736	3.88267	2
3.01182	0.87130	3.13921	3
5.72536	1.69797	5.83894	4

There exists a large number of periodic solutions of the problem (28) and even many period doubling bifurcation points [35]. A domain of attraction of individual bifurcation points is therefore small. The successfullness of the two algorithms presented here and two new algorithms will be tested in a forthcoming paper.

V. DISCUSSION

Evaluation of bifurcation points is an important part of an analysis of a nonlinear dynamical system. Continuation of bifurcation points leads then to a construction of the bifurcation diagram which contains much information about behaviour of the system. Therefore, the algorithms for direct evaluation of bifurcation points play an important role in an ensamble of numerical methods used in the bifurcation theory. Several of recently developed algorithms are subject of this paper. There are two additional techniques connected with bifurcation points. One of them gives directions of branches emanating from the bifurcation point and enables to start the continuation algorithm in its neighbourhood [2,13,21]. The second technique is devoted to an evaluation of a point of an isola formation (birth of isolas) in the solution diagram [3,24].

The algorithms for evaluation of period doubling bifurcation points in periodic solutions can be easily used also for nonautonomous systems of differential equations with perio-

dic right hand sides.

ACKNOWLEDGMENTS

The authors would like to express their thanks to the Deutscher Akademischer Austauschdienst (DAAD) for financial support during the conference. We would like to thank also to prof. Miloš Marek for very useful discussion and help with the preparation of the manuscript.

LITERATURE

- [1] V.I.Arnold: Additional chapters of the theory of ordinary differential equations, Nauka, Moscow, 1978 (in Russian)
- [2] D.W.Decker, H.B.Keller: Solution branching - A constructive technique, in New Approaches to Nonlinear Problems in Dynamics (P.J.Holmes, Ed.). SIAM publ. Philadelphia 1980, p.53
- [3] D.Dellwo, H.B.Keller, B.J.Matkowsky, E.L.Reiss: On the birth of isolas, SIAM, J. Appl. Math., in press
- [4] V.Hlaváček, H.Hofmann, M.Kubíček: Modelling of chemical reactors XXIV., Chem. Engng. Sci. 26 (1971), 1629
- [5] V. Hlaváček, M.Kubíček, M.Marek: Analysis of non stationary heat and mass tranfer in a porous catalyst particle I-II, J. Catal. 15, (1969),17, 31
- [6] M.Holodniok, M.Kubíček: New algorithms for evaluation of complex bifurcation points in ordinary differential equations. A comparative numerical study, Appl. Math. Comput., to be published
- [7] M.Holodniok, M.Kubíček: Continuation of periodic solutions - Algorithm and applications to the Lorenz model, this proceedings
- [8] M.Holodniok, M.Kubíček, V.Hlaváček: Computation of the flow between two rotating coaxial disks, J. Fluid Mech. 81 (1977) 689
- [9] M.Holodniok, M.Kubíček, V.Hlaváček: Computation of the flow between two rotating coaxial disks. Multiplicity of steady state solutions, J. Fluid Mech. 108, (1981), 227
- [10] G.Iooss, D.D.Joseph: Elementary stability and bifurcation theory, Springer Verlag, New York, 1980
- [11] K.F.Jensen, W.H.Ray: The bifurcation behavior of tubular reactors, Chem. Engng, Sci. 37, (1982), 199
- [12] A.D.Jepson : Numerical Hopf bifurcation, PhD Thesis, Dept. of Math. Calif. Inst. of Technology, 1981

- [13] H.B.Keller: Numerical solution of bifurcation and nonlinear eigenvalue problems, in Applications of Bifurcation Theory (P.H.Rabinowitz, Ed.). Academic Press, New York 1977, p.359
- [14] H.B.Keller, R.K.-H.Szeto: Calculations of flows between rotating disks, preprint.
- [15] M.Kubíček: Evaluation of branching points for nonlinear boundary value problems based on the GPM technique, Appl. Math. Comput. 1 (1975), 341
- [16] M.Kubíček: Linear systems with almost band matrix, Sci. papers of the Inst. of Chem. Technol. Prague, K 12 (1977), 5
- [17] M.Kubíček: Algorithm for evaluation of complex bifurcation points in ordinary differential equations. SIAM J.Appl. Math. 38 (1980), 103
- [18] M.Kubíček: Occurrence of oscillatory regimes in lumped parameter systems. Determination of Hopf's bifurcation points. Chem. Engng. Sci. 34 (1979), 1078
- [19] M.Kubíček, V.Hlaváček: Solution of boundary value problems IX. Evaluation of branching points based on the differentiation with respect to boundary condition. Chem. Engng. Sci. 30 (1975), 1439
- [20] M.Kubíček, M.Holodniok: Evaluation of Hopf bifurcation points in parabolic equations describing heat and mass transfer in chemical reactors. Chem. Engng. Sci. in press
- [21] M.Kubíček, A.Klič: Direction of branches bifurcating at a bifurcation point. Determination of starting points for a continuation algorithm, Appl. Math. Comput. 13 (1983), 125
- [22] M.Kubíček, M.Marek: Computational methods in bifurcation theory and dissipative structures. Springer Verlag, New York, 1983
- [23] M.Kubíček, M.Marek: Evaluation of limit and bifurcation points for algebraic and nonlinear boundary value problems. Appl. Math. Comput. 5, (1979), 253
- [24] M.Kubíček, I.Stuchl, M.Marek: "Isolas" in solution diagrams, J.Comput. Phys. 48 (1982) 106
- [25] G.N.Lance, M.H.Rogers: The axially symmetric flow of a viscous fluid between two infinite rotating disks, Proc. Roy, Soc. A 266 (1962), 109
- [26] J.E.Marsden, M.McCracken. The Hopf bifurcation and its applications. Springer Verlag, Berlin 1976
- [27] G.L.Mellor, P.J.Chapple, V.K.Stokes: On the flow between a rotating and a stationary disk, J. Fluid. Mech. 31 (1968), 95
- [28] H.D.Mittelmann, H.Weber, ed.: Bifurcation problems and their numerical solution, Birkhäuser, Basel 1980

- [29] H.D.Mittelmann, H.Weber: Numerical methods for bifurcation problems - A survey and classification, in [28], p. 1
- [30] N.D.Nguyen, J.P.Ribault, F.Florent: Multiple solutions for flow between coaxial disks, *J. Fluid. Mech.* 68 (1975), 369
- [31] G.Nicolis, I.Prigogine: Self-organization in nonequilibrium systems, *J. Wiley*, New York 1977
- [32] J.Peterson, K.A.Overholser, R.F.Heinemann: Hopf bifurcation in a radiating laminar flame, *Chem. Engng. Sci.* 36 (1981), 628
- [33] S.M.Roberts, J.S.Shipman: Computation of the flow between a rotating and a stationary disk, *J. Fluid. Mech.* 73 (1976), 53
- [34] D.H.Sattinger: Topics in stability and bifurcation theory. *Lecture Notes in Math.* 309, Springer Verlag, Berlin 1973
- [35] I.Schreiber, M.Kubíček, M.Holodniok, M.Marek: Periodic phenomena in coupled cells, in preparation
- [36] I.Schreiber, M.Kubíček, M.Marek: On coupled cells, In *New Approaches to Nonlinear Problems in Dynamics*, ed. by P.J. Holmes, SIAM Publ., Philadelphia 1980, p. 496
- [37] R.Seydel: Numerical computation of branch points in nonlinear equations, *Numer. Math.* 33 (1979), 339
- [38] R.Seydel: Numerical computation of branch points in ordinary differential equations, *Numer. Math.* 32 (1979), 51
- [39] A.Varma, N.R.Amundson: Some problems concerning the non-adiabatic tubular reactor: A-priori bounds, qualitative behavior, preliminary uniqueness and stability considerations, *Canad. J. Chem. Engng.* 50 (1972), 470
- [40] J.H.Wilkinson, C.Reinsch: *Handbook for automatic computation II, Linear algebra*, Springer Verlag, Berlin, 1971

Milan Kubíček

Martin Holodniok

Department of Chemical Engineering and Computer Center
 Prague Institute of Chemical Technology
 Suchbátarova 5
 166 28 Praha 6, Czechoslovakia

FEEDBACK STIMULATED BIFURCATION*

Tassilo Küpper and Boguslav Kuszta

1. Introduction

This paper is concerned with an application of bifurcation theory to systems identification. If a technical system is governed by some control parameter it is often easy to determine the bifurcation points experimentally since they are characterized by a significant change in the output. For example think of the change from a constant to a periodic motion which is really evident since it appears as a qualitative rather than a quantitative change. This observation leads to an obvious test for the validity of a mathematical model since the bifurcation points of the system and of the model must coincide.

By its very nature this simple and reliable method seems to be restricted to systems which are in a process of bifurcation. This may be a serious disadvantage as many practical applications deal with tame systems which are in a state far away from bifurcation.

Here we show how this disadvantage can be overcome. Just for the purpose of model identification in tame systems we propose to introduce an artificial bifurcation where we make use of the fact that a system can be forced to change its state by a sufficiently strong feedback procedure. Those artificially created bifurcation points can then be used in the usual way to discriminate among mathematical models. This idea of so called feedback stimulated bifurcation has been used for the first time by Kuszta and Bailey [5] and more recently in several applications [4, 6].

*This research was started while both authors were visiting the California Institute of Technology in Pasadena.

Here we first explain it by a simple heat conduction process in a one-dimensional wire of length 1. The temperature distribution is denoted by $u(x,t)$ with initial distribution $u(x,0) = u_0$. We assume that the temperature at the endpoints is kept fixed at 0. This situation is modelled by the following set of equations:

$$u_t = D u_{xx} + C u + f(u) \quad (1.1)$$

$$u(x,0) = u_0 \quad (1.2)$$

$$u(0,t) = u(1,t) = 0 \quad (1.3)$$

The coefficients D and C and the nonlinearity f are unknown in general and must be determined. In its present form (1.1), (1.2), (1.3) is not a bifurcation problem. Depending on the initial conditions the temperature will develop to a steady state but it will be difficult to use this quantitative information to determine the unknown coefficients. Instead we propose to introduce a feedback procedure: The temperature at the right end is adjusted proportional to the temperature measured at some interior point x_0 . In an experimental device this can be realized for example by a bimetal thermostat. We expect that the feedback will force the system to change its state but of course this must be proved for the corresponding mathematical model:

$$u_t = D u_{xx} + C u + f(u) \quad (1.1)$$

$$u(x,0) = u_0 \quad (1.2)$$

$$u(0,t) = 0, \quad u(1,t) = \gamma u(x_0, t) \quad (1.4)$$

Since γ is a variable parameter it can be considered as a bifurcation problem. We'll see that bifurcation occurs indeed and that it determines in this special example the ratio C/D uniquely.

We'll treat this special example in a more general setting which is formulated in section 2. In section 3 we study the corresponding stationary problem and prove that there is exactly one bifurcation point generated by the feedback procedure. Using the method of lower and upper solutions we further show that there is a global branch of positive solutions. Section 4 is devoted to the time-dependent problem. Using a modified monotone iteration scheme we obtain the existence and uniqueness of solutions as well as stability results. In section 5 we exemplify how to use feedback stimulated bifurcation to determine certain parameters in a mathematical model. We conclude in section 6 with numerical results which illustrate how the feedback mechanism works.

2. The Mathematical Model

Let $\Omega \subseteq \mathbb{R}^n$ denote a bounded domain whose boundary $\partial\Omega$ belongs to the class $C^{2+\alpha}$ for some fixed $\alpha \in (0, 1)$. In $\Omega \times [0, \infty)$ we consider the quasilinear parabolic equation

$$u_t = Lu + F(u) \quad (2.1)$$

together with the linear feedback boundary condition

$$u(x, t) = \gamma \Phi(u)(t) \rho_0(x) \quad (x \in \partial\Omega, t \geq 0) \quad (2.2)$$

and the initial data

$$u(x, 0) = u_0(x) \quad (x \in \Omega). \quad (2.3)$$

We assume that the linear operator L , the nonlinearity F , the feedback-functional Φ and the functions ρ_0 and u_0 satisfy the following hypotheses:

(H1) L is the uniformly elliptic differential operator

$$Lu = - \sum_{i,j=1}^n a_{ij}(x) \frac{\partial^2 u}{\partial x_j \partial x_i} + \sum_{j=1}^n b_j(x) \frac{\partial u}{\partial x_j} + c(x)u$$

with real coefficients $a_{ij} \in C^{2+\alpha}(\bar{\Omega})$, $b_j \in C^{1+\alpha}(\bar{\Omega})$ and $c \in C^\alpha(\bar{\Omega})$. Further, L together with Dirichlet boundary conditions is inverse-positive, i. e. L satisfies

$$\left. \begin{array}{l} Lu(x) \geq o(x \in \Omega) \\ u(x) \geq o(x \in \partial\Omega) \end{array} \right\} \Rightarrow u(x) \geq o(x \in \Omega)$$

(H2) The nonlinearity F is defined by

$$F(u)(x) = f(x, u(x)) u(x)$$

where $f \in C^1(\bar{\Omega} \times [0, \infty))$ is nonnegative, monotonically increasing in its second variable and satisfies $f(x, 0) = 0$ ($x \in \bar{\Omega}$).

(H3) The feedback functional $\Phi : C^0(\bar{\Omega}) \rightarrow \mathbb{R}$ is a positive linear functional; i.e. $\Phi(u) \geq 0$ for nonnegative functions u and $\Phi(u) > 0$ if u is positive in Ω . The function $\rho_0 \in C^{2+\alpha}(\partial\Omega)$ is nonnegative on $\partial\Omega$ and positive in at least one point of $\partial\Omega$.

Remarks. (i) A necessary and sufficient condition for L to be inversepositive is that the solution to $Lz(x) = 1(x \in \Omega)$, $z(x) = 1(x \in \partial\Omega)$ is positive (see Schröder [8]).

(ii) Typical examples for the feedback functional Φ are

$$(a) \quad \Phi(u) = \int_{\Gamma} w(x) u(x, t) dx$$

$$(b) \quad \Phi(u) = \sum_{j=0}^n \gamma_j u(x_j, t)$$

where we assume that $\Gamma \subseteq \Omega$ is a subdomain, $w \in C^0(\Gamma)$ is nonnegative, $\gamma_j \geq 0$ and $x_j \in \Omega$ ($j=0, \dots, n$). The feedback boundary condition could also be replaced by a more general type of the form which for example was used by Triggiani [9]:

$$u(x, t) = \gamma \sum_{j=0}^N \phi_j(u)(t) \rho_j(x)$$

with suitable functionals ϕ_j and functions ρ_j

We also mention the possibility of a time-delay in the feedback which can be of practical meaning.

3. The Stationary Problem: Existence of Solutions and Bifurcation

We first consider the stationary problem associated with (2.1), (2.2) and (2.3) since it plays an important part in stability considerations of the time dependent problem.

$$Lu + F(u) = 0 \quad (3.1)$$

$$u(x) = \gamma \Phi(u) \rho_0(x) \quad (x \in \partial\Omega) \quad (3.2)$$

Let h_0 denote the solution to $Lh_0 = 0$ together with the boundary conditions $h_0(x) = \rho_0(x)$ ($x \in \partial\Omega$). By [1] h_0 is positive in Ω ; hence $\Phi(h_0) > 0$ by (H3).

Since the linear problem

$$Lh = 0$$

$$h(x) = \gamma \Phi(h) \rho_0(x) \quad (x \in \partial\Omega)$$

has the uniquely defined solution $h = \gamma \Phi(h) h_0$ we have $\Phi(h) = \gamma \Phi(h) \Phi(h_0)$ and either $\Phi(h) = 0$ and consequently $h = 0$ due to (H1) and the boundary condition $h|_{\partial\Omega} = 0$ or $1 = \gamma \Phi(h_0)$. Hence $\gamma_0 = 1/\Phi(h_0)$ is the only possible bifurcation point.

Theorem 3.1 There exists a global branch of positive solutions $u_\gamma \in C^{2+\alpha}(\bar{\Omega})$ ($\gamma_0 < \gamma < \gamma_\infty$) to (3.1), (3.2) such that

$$(i) \quad \lim_{\gamma \rightarrow \gamma_0} \|u_\gamma\|_\infty = 0$$

$$(ii) \quad \lim_{\gamma \rightarrow \gamma_\infty} \|u_\gamma\|_\infty = \infty$$

The proof of theorem 3.1 will be based on a series of lemmas. The solution to the linear boundary value problem $Lu = r$, $u(x) = \delta \rho_0(x)$ ($x \in \partial\Omega$) can be represented as $u(x) = \int_{\Omega} G(x,y) r(y) dy + \delta h_0(x)$ where G is the nonnegative Green's function to L with respect to homogeneous Dirichlet boundary conditions. With $r = -F(u)$ and $\delta = \gamma \Phi(u)$ we obtain from (3.1), (3.2):

$$u(x) = - \int_{\Omega} G(x,y) F(u)(y) dy + \gamma \Phi(u) h_{\infty}(x),$$

hence

$$\Phi(u) = - \Phi \left(\int_{\Omega} G(x,y) F(u)(y) dy \right) + \gamma \Phi(u) \Phi(h_{\infty})$$

and

$$\Phi(u) = - \Phi \left(\int_{\Omega} G(x,y) F(u)(y) dy \right) / (1 - \gamma/\gamma_{\infty}).$$

Hence we have shown:

Lemma 1. For a solution to (3.1) the boundary condition (3.2) is equivalent to

$$u(x) = (-\gamma/(1-\gamma/\gamma_{\infty})) \Phi \left(\int_{\Omega} G(x,y) F(u)(y) dy \right) \rho_{\infty}(x) (x \in \partial\Omega)$$

Lemma 2. For each $\delta > 0$ there is a uniquely determined positive solution v_{δ} to the auxiliary problem

$$L v + F(v) = 0$$

$$v(x) = \delta \rho_{\infty}(x) (x \in \partial\Omega).$$

Further, v_{δ} satisfies the estimate

$$\delta h_1(x) \leq v_{\delta}(x) \leq \delta h_{\infty}(x) (x \in \Omega) \quad (3.3)$$

where h_1 is the solution to $(L-\lambda)h_1 = 0$, $h_1(x) = \rho_{\infty}(x)$ ($x \in \partial\Omega$) for a suitable $\lambda < 0$.

Lemma 2 follows from the fact that $\psi = \delta h_{\infty}$ is a supersolution and $\phi = \delta h_1$ is a subsolution if $\lambda < 0$ is chosen appropriately.

Lemma 3. For each $\delta > 0$ the function $u_{\gamma} = v_{\delta}$ is a solution to (3.1), (3.2) for

$$\gamma = \gamma_{\infty} / [1 - \gamma_{\infty} \Phi \left(\int_{\Omega} G(x,y) F(v_{\delta})(y) dy / \delta \right)] \quad (3.4)$$

Proof of Lemma 3. By definition (3.4) γ is a solution to $\delta = -\gamma\Phi(\int_D G(x,y)F(v_\delta)(y)dy)/(1-\gamma/\gamma_0)$ for fixed $\delta > 0$. Hence v_δ satisfies the boundary condition (3.2) by Lemma 1. Since $F(v_\delta)(x) = f(x, v_\delta(x))v_\delta(x)$, $h_1 \leq v_\delta/\delta \leq h_0$ and $f(x,0) = 0$ it follows that $\gamma \geq \gamma_0$ as $\delta \rightarrow 0$. Further $0 \leq u_\gamma(x) = \gamma\Phi(u_\gamma)\rho_0(x)$ implies that $\gamma \geq 0$.

Using $\Phi(\int_\Omega G(x,y)F(v_\delta)(y)dy) \geq 0$ we obtain that $\gamma \geq \gamma_0$. Finally, γ_∞ is defined by

$$\gamma_\infty = \sup_{\delta \geq 0} \gamma_0 / [1 - \gamma_0 (\int_\Omega G(x,y)F(v_\delta)dy/\delta)] .$$

Lemma 4. There is no nonnegative solution $u \neq 0$ to (3.1), (3.2) for $\gamma \leq \gamma_0$.

Proof. Let $u \geq 0$ be a nontrivial solution for $\gamma \leq \gamma_0$. Hence $\Phi(u) \geq 0$ and $\beta = \Phi(\int_\Omega G(x,y)F(u)(y)dy) \geq 0$ and $0 \leq u(x) = -\gamma\beta/(1-\gamma/\gamma_0)\rho_0(x)$. Consequently $\gamma < 0$ in contradiction to $0 \leq u(x) = \gamma\Phi(u)\rho_0(x)$.

Remark. The positive solution u to the stationary problem (3.1), (3.2) is unique if $F(v_\delta)/\delta$ increases monotonically with δ . To see this note that under this hypothesis γ given by (3.4) increases monotonically, too. Hence for each $\delta > 0$ there is a unique $\gamma > \gamma_0$.

4. The Initial Value Problem with Feedback

Solutions to the time-dependent problem can be obtained by a modified monotone iteration scheme. This standard procedure (see for example [7]) requires a few modifications which are due to the feedback boundary conditions. As lower and upper solutions we use as usual the solutions to the stationary problem which also determine domains of stability.

Theorem 4.1. Assume $\gamma_0 < \gamma_1 \leq \gamma \leq \gamma_2 < \gamma_\infty$ and $u_{\gamma_1} \leq u_0 \leq u_{\gamma_2}$.

(i) There exists a solution $u(x,t)$ to (2.1), (2.2), (2.3) such that

$$u_{\gamma_1}(x) \leq u(x,t) \leq u_{\gamma_2}(x) \quad (x \in D, t \geq 0)$$

$$(ii) \quad u_\gamma(x) = \lim_{t \rightarrow \infty} u(x,t)$$

if the stationary problem (3.1), (3.2) has a uniquely determined solution.

Remark. Any solution to (2.1), (2.2), (2.3) with initial data $u_0 > 0$ decays to zero as $t \rightarrow \infty$ if $\gamma \leq \gamma_0$.

Proof of theorem 4.1. The proof follows standard ideas using a monotone iteration scheme. Here we restrict our attention to the modification which are due to the feedback boundary condition. For suitable initial values v_0 and u^0 the Monotone Iteration Scheme is defined by

$$\frac{\partial u^n}{\partial t} = L u^n + F(u^n)$$

$$u^n(x,t) = \gamma \Phi(u^{n-1})(t) \rho_0(x) \quad (x \in \Omega, t \geq 0)$$

$$u^n(x,0) = v_0$$

1) For $u^0 = u_{\gamma_2}$, $v_0 = u_0$ we see that u^0 is a supersolution for $n = 1$, since u^0 satisfies the differential equation and $u_{\gamma_2}^0(x,t) = \gamma_2 \Phi(u_{\gamma_2}^0)(t) \rho_0(x) \geq \gamma \Phi(u_{\gamma_2}^0) \rho_0(x) = \gamma \Phi(u^0) \rho_0(x)$ and $u^0(x,0) \geq u_0(x)$. In the same way we see that u_{γ_1} is a subsolution. By [7, Thm. 2.3.2] there exists a solution $u^1(x,t)$ for the case $n = 1$ satisfying $u^0(x) \geq u^1(x,t)$ ($x \in \Omega, t \geq 0$). Since u^1 satisfies the differential equation, the initial condition and $\gamma \Phi(u^1)(t) \rho_0(x) < \gamma \Phi(u^0) \rho_0(x) = u^1(x,t)$ ($x \in \Omega, t \geq 0$) we obtain that u^1 is a supersolution for the equation in the case $n = 2$; hence there exists a solution u^2 such that $u_{\gamma_1} \leq u^2 \leq u^1$. By the same argument there is a monotone decreasing sequence of solutions u^n

$$u_{\gamma_1}(x) \leq \dots \leq u^n(x,t) \leq u^{n-1}(x,t) \leq \dots \leq u^0 = u_{\gamma_2}$$

converging to a solution u of (2.1), (2.2), (2.3).

2) If we repeat the Monotone Iteration Scheme with the initial data $u^0 = u_{Y_2}$ and $v_0 = u_{Y_2}$ we obtain a sequence u^n converging to a solution \bar{u} satisfying (3.1), (3.3) and $\bar{u}(x,0) = u_{Y_2}^n(x)$ ($x \in \Omega$). In addition we show that $\frac{\partial u^n}{\partial t} \leq 0$ and hence $\frac{\partial \bar{u}}{\partial t} \leq 0$.

For fixed $h > 0$ define $w_h^n(x,t) = [u^n(x,t+h) - u^n(x,t)]/h$. Then w_h^n satisfies the differential equation

$$\frac{\partial w_h^n}{\partial t} = L w_h^n + \xi_h^n(x,t) w_h^n$$

$$\text{where } \xi_h^n(x,t) = \int_0^1 \frac{\partial F}{\partial u} (\tau u(x,t+h) + (1-\tau)u(x,t)) d\tau$$

satisfies $\xi_h^n(x,t) w_h^n(x,t) = F(u^n(x,t+h)) - F(u^n(x,t))$.

Further $w_h^n(x,0) = [u^n(x,h) - u^n(x,0)]/h = [u^n(x,h) - u_{Y_2}(x)]/h \leq 0$ for all n .

Since $u^0(x,t) = u_{Y_2}(x)$ we have $\Phi(u^0)(t+h) \leq \Phi(u^0)(t)$ for all $t \geq 0$, hence

$\Phi(w_h^0) \leq 0$ and $w_h^1(x,t) = \gamma \Phi(w_h^0)(t) \rho_0(x) \leq 0$. The maximum principle applied to w_h^1 gives $w_h^1(x,t) \leq 0$ for all $x \in \Omega$. Consequently we obtain $\frac{\partial w_h^1}{\partial t}(x,t)$

$= \lim_{h \rightarrow 0} w_h^1(x,t) \leq 0$. Repetition of the same argument proves $\frac{\partial w_h^n}{\partial t} \leq 0$ for all n

and $\frac{\partial \bar{u}}{\partial t} \leq 0$. The comparision principle applied to the sequences \bar{u} and u

yields $u \leq \bar{u}$. Since $\bar{u}(x,t)$ is monotonically nondecreasing in t the limit

$\hat{u}(x) = \lim_{t \rightarrow \infty} \bar{u}(x,t)$ exists and is a solution to the stationary problem (for details see for example [7, Thm. 2.6.1].

Starting with $u^0 = u_{Y_1}$, $v_0 = u_{Y_1}$ leads to a solution \underline{u} of (3.1), (3.2) satisfying $\frac{\partial \underline{u}}{\partial t} \geq 0$ and $\underline{u} \leq u$. Again $\tilde{u}(x) = \lim_{t \rightarrow \infty} \underline{u}(x,t)$ is a solution to the stationary problem, hence by uniqueness, $\tilde{u} = \hat{u}$. Because of $\underline{u}(x,t) \leq u(x,t) \leq \tilde{u}(x,t)$

it follows that $\lim_{t \rightarrow \infty} u(x, t) = \hat{u}(x) = u_\gamma(x)$.

5. Parameter Identification via Feedback Stimulated Bifurcation

In section 4 we have seen that the linear part L uniquely determines the bifurcation point γ_0 . In practice, the situation is reverse since we aim to model a given experiment by the equations (2.1), (2.2), (2.3). In the experiment it is the feedback parameter γ which is adjustable, and γ_0 is easily determined since the bifurcation appears as a significant change of the output as γ passes γ_0 . This observation leads to a simple method to identify one parameter in the linear part L of the mathematical model. To illustrate this we consider a simple reaction-diffusion equation.

Example. In equation (2.1) let L be given by $Lu = -\Delta u + cu$ and assume that c is to determine. Let $\gamma_0 > 0$ be the observed bifurcation point and let h^c denote the solution to the linear problem

$$-\Delta h^c + ch^c = 0 \quad (5.1)$$

$$h^c(x) = \rho(x) \quad (x \in \partial\Omega) \quad (5.2)$$

Further let c_0 be the largest number such that the corresponding problem with homogeneous boundary conditions is inversepositive for $c < c_0$; i. e. $-\Delta h + ch \geq 0$, $h(x) \geq 0$ ($x \in \partial\Omega$) implies $h \geq 0$. Then there is a unique $C^* > c_0 < 0$ such that γ_0 is a bifurcation point for problem (3.1), (3.2) with $c = C^*$.

Remark. The parameter C^* is uniquely determined by the corresponding solution h^{C^*} of (5.1), (5.2) which satisfies $\gamma_0^\Phi(h^{C^*}) = 1$. Inversepositivity leads to an error bound: if h^{c_1}, h^{c_2} are 2 solutions of (5.1), (5.2) satisfying $\gamma_0^\Phi(h^{c_1}) < 1 < \gamma_0^\Phi(h^{c_2})$, then $c_1 > C^* > c_2$. The existence of a solution h^{C^*} with $\gamma_0^\Phi(h^{C^*}) = 1$ follows directly from the inversepositivity, the fact,

that h^C depends continuously on C for $C < C_0$ and that for each $x \in \Omega$

$$\lim_{C \rightarrow +\infty} h^C(x) = 0 \quad \lim_{C \rightarrow C_0^-} h^C(x) = \infty$$

For example in case $n = 1$, $\Omega = (0, 1)$ and $\Phi(u) = u(x_0, t)$ the constant C^* is given as the solution of $f^C(1) = f^C(x_0)$ where $f^C(x) := \sinh \sqrt{c} x$ ($c > 0$), $f^0(x) := \sin \sqrt{-c} x$ ($0 > c > -\pi^2$).

6. Numerical Results

The experiment can very well be simulated by numerical calculations. We conclude with some numerical results, since it is interesting to see how the feedback mechanism works.

For illustration we list numbers for the example (1,1) (1,2) (1,4) mentioned in the introduction with $\Omega = (0, 1)$, $D = 1$, $C = +1$, $f = -x^3$, $x_0 = 0,75$.

As discretization we have chosen the explicit leapfrog scheme with stepsize $h_0 = 1/32$ in x -direction and $h_1 = 0,25 \times 10^{-3}$ in time. As bifurcation point we obtain $\gamma_0 = 1, 2 3 4 \dots$.

If the initial value u_0 is taken to be a perturbation of the trivial solution $u \equiv 0$ which deviates only in 1 gridpoint it is first observed that the diffusion process smears this significant perturbation on the whole interval in short time. Only afterwards the feedback becomes dominant. Depending if γ is less or bigger than the bifurcation point the solution approaches 0 or a nontrivial stationary solution. Numerical calculations also show (although not listed here) that the velocity of the feedback stabilization process depends on the functional Φ .

<u>t</u>	<u>x</u>	o.125	o.25	o.375	o.5	o.625	o.75	o.875
<u>$\gamma = 1$</u>								
o.		.o	.o	o.o	1.	.o	.o	.o
o.o2		.o1o	.o29	.o52	.o63	.o52	.o31	.o21
o.5		.4o E-2	.11 E-1	.11 E-1	.14 E-1	.16 E-1	.17 E-1	.18 E-1
1.o		.13 E-2	.26 E-2	.37 E-2	.47 E-2	.54 E-2	.58 E-2	.6o E-2
1.5		.44 E-3	.86 E-3	.12 E-2	.15 E-2	.17 E-2	.19 E-2	.19 E-2
2.o		.14 E-3	.28 E-3	.4o E-3	.58 E-3	.63 E-3	.63 E-3	.65 E-3
<u>$\gamma = 2$</u>								
o.		o.	o.	o.	1.	o.	o.	o.
o.o2		o.o1o	o.o29	o.o52	o.o63	o.o53	o.o34	o.o34
o.5		o.2311	o.4786	o.7597	1.o95	1.511	2.o57	2.834
1.o		o.4296	o.8545	1.277	1.714	2.2o7	2.481	3.811
1.5		o.43o2	o.8555	1.278	1.716	2.2o8	2.842	3.812
2.o		o.43o2	o.8555	1.278	1.716	2.2o8	2.842	3.812

References:

- [1] H. Amann: On the existence of positive solutions of nonlinear elliptic boundary value problems.
Ind. Univ. Math. J. 21, 125 - 146 (1971)
- [2] K. Glashoff, J. Sprekels: An application of Glicksberg's theorem to set-valued integral equations arising in the theory of thermostats.
SIAM J. Math. Anal. 12, 477-486 (1981)
- [3] K. Glashoff, J. Sprekels: The regulation of temperature by thermostats and set-valued integral equations.
J. Integral Eq. 4, 95-112 (1982)
- [4] T. Küpper, B. Kuszta: Verzweigung bei Rückkopplungsproblemen
to appear in ZAMM
- [5] B. Kuszta, J. E. Bailey: Nonlinear Model Identification by analysis of feedback-stimulated bifurcation.
IEEE Transactions on Automatic Control. AC -27, 227-228
(1982)
- [6] G. Lyberatos, B. Kuszta, J. Bailey: Discrimination and identification of dynamic catalytic reaction models via introduction of feedback.
Preprint
- [7] D. Sattinger: Topics in stability and bifurcation theory.
Springer Lecture Notes 309, 1973
- [8] J. Schröder: Operator Inequalities.
Academic Press New York 1980

[9] R. Triggiani: Well-posedness and regularity of boundary feedback parabolic systems.

J. Diff. Eq. 36, 347 - 362 (1980)

Tassilo Küpper
Abt. Mathematik
Universität Dortmund
Postfach 50 05 00
D-4600 Dortmund 50

Boguslav Kuszta
Institute of Control
and Electronics
Technical University
Warsaw

NUMERICAL STUDIES OF TORUS BIFURCATIONS

W.F. Langford

The normal form equations for the interactions of a Hopf bifurcation and a hysteresis bifurcation of stationary states can give rise to an axisymmetric attracting invariant torus. Nonaxisymmetric perturbations are found to produce phase locking, period doubling, bistability, and a family of strange attractors.

1. Introduction

The major unresolved problem of bifurcation theory today may well be that of understanding the transition to turbulence. This problem has been the subject of several recent books and conferences, see [1-4]. The so-called "main sequence" [5,6] of bifurcations leading to turbulence begins with one or more bifurcations of stationary states, followed by a Hopf bifurcation to a periodic orbit, then a Naimark-Sacker torus bifurcation (and possibly a period doubling cascade), and finally the appearance of a strange attractor representing turbulent flow. Such sequences of transitions have been observed in experiments having simple geometries, for example Rayleigh-Bénard convection [7], Taylor vortices [8], and oscillating chemical reactions [9]. Mathematically, the first two or three bifurcations in this sequence are fairly well understood, and even explain some of the experimental data, but from the torus bifurcation onward very little is understood.

Recent studies of certain vector fields close to a singularity with eigenvalues $\{0, +iw, -iw\}$ have revealed the existence in this context of the following bifurcation sequence: stationary bifurcation, Hopf bifurcation, bifurcation to an invariant torus; see [10-20]. It is convenient to think of these singularities as resulting from the coalescence of two "primary" bifurcations: a Hopf bifurcation (corresponding to the eigenvalues $+iw, -iw$) and a bifurcation of stationary states (corresponding to the eigenvalue 0). Then it is not surprising to find stationary and Hopf bifurcations on unfolding the singularity, however the appearance of an

invariant torus was unexpected. Its existence is guaranteed in a small open region of the unfolding-parameter space when certain inequalities involving the low order nonlinear terms are satisfied. The primary bifurcation of stationary states entering into this coalescence may be any from a growing list of cases: saddlenode [12], transcritical [10], pitchfork [11,13], hysteresis or cusp [17], and isola. Which case is to be expected in a given problem depends on the number of parameters in the problem and on the presence or absence of symmetries.

This paper is a continuation of the work in the previous paragraph, in that it presents studies of bifurcations that occur after the first appearance of an invariant torus. The approach is numerical, in contrast to the analytical results referenced above. The bifurcations in question involve global dynamics, so complex that analytical techniques are inadequate to give a fully detailed description. One quite recent analytical result, established for the transcritical-Hopf case, is that in a neighborhood of the singular point the relative measure of parameter values corresponding to quasiperiodic flow on the torus approaches unity as the neighborhood shrinks to zero [18]. Another is the observation that, if as the torus grows it approaches a heteroclinic saddle connection, then a theorem of Silnikov may imply the presence of Smale's horseshoes and hence chaotic dynamics [12]. Between these two extremes, very little is known analytically. The present numerical studies help fill this gap and may point the way for future theoretical work. Numerically we find strange attractors with large basins of attraction, in contrast to the "leaky" chaos of the Silnikov mechanism.

2. The Model Equations

The numerical results presented here are for the case of interactions of hysteresis and Hopf bifurcations, a case which is very rich in structure and is still under investigation, but which promises to have important applications. We will assume that the spectrum contains the simple eigenvalues $\{0, +i\omega, -i\omega\}$ with the remainder in the negative half-plane, and that a two-step preliminary transformation of the system has been performed, consisting of first a reduction to the 3-dimensional center

manifold, and then a transformation of this system to its Poincaré-Birkhoff normal form. The resulting equations can be written:

$$(1) \quad \begin{aligned} x' &= (z - \beta)x - \omega y + \text{h.o.t.} \\ y' &= \omega x + (z - \beta)y + \text{h.o.t.} \\ z' &= \lambda + \alpha z + az^3 + b(x^2 + y^2) + \text{h.o.t.} \end{aligned}$$

Here λ is the bifurcation parameter, and α, β are unfolding parameters, all near 0. The frequency ω comes from the original pure imaginary eigenvalue $i\omega$. The a and b terms are "resonant", and "h.o.t." stands for higher order terms, which do not affect the classification of stationary and periodic solutions, but do influence the full dynamics and especially the flow on and near the torus. In principle, the higher order terms can be transformed to be axisymmetric about the Z-axis up to arbitrarily high order, but the procedure does not converge in general; the neighborhood of validity shrinks to zero as the order is increased. Furthermore, axisymmetry is a highly nongeneric condition. Therefore it seems essential that nonaxisymmetry be retained if the conclusions are to have any practical applications.

The model equations used for numerical investigation were chosen for computational economy while retaining the essential features of (1). The variables and parameters were assumed rescaled to make the leading terms in (1) of order one, then a and b were assigned values $-(1/3)$ and -1 respectively corresponding to one of the most interesting cases in the general classification. The remaining resonant cubic term in the z' equation was retained along with a simple nonaxisymmetric fourth degree monomial to incorporate important nonaxisymmetric effects. Higher order terms in the x' and y' equations have been dropped for the results presented here, their presence does not seem to give bifurcation phenomena qualitatively different from what has been found for (2).

$$(2) \quad \begin{aligned} x' &= (z - \beta)x - \omega y \\ y' &= \omega x + (z - \beta)y \\ z' &= \lambda + \alpha z - z^3/3 - (x^2 + y^2)(1 + \rho z) + \epsilon zx^3 \end{aligned}$$

Here ρ and ϵ are "small", ρ determines the location of the Naimark-Sacker torus bifurcation, and ϵ controls the nonaxisymmetry. Analogous model equations have been derived from normal forms for the transcritical-Hopf and pitchfork-Hopf cases, and are described in [19,20] and [11,17] respectively.

3. Numerical Results

Since it is the qualitative behaviour of solutions that is of interest, the results are presented here in graphical form. The computations were performed on an IBM personal computer, working in compiled BASIC in double precision with an 8087 numerical coprocessor, and the graphics were produced on a Hewlett-Packard 7470A Plotter. The accuracy of the computations is considered to be sufficient to show the location and qualitative features of attracting sets (ω -limit sets), however the flow inside a strange attractor is characterized by divergence of trajectories and sensitive dependence on initial conditions, so that the numerical solutions do not accurately predict the location of a point on a trajectory inside a strange attractor over long time intervals. Similarly the initial-value methods used here (Runge-Kutta and predictor-corrector) are inadequate for computing unstable orbits of saddle type.

Figures 1 to 13 show a series of computed solutions of system (2) with parameter values $\alpha = 1$, $\beta = 0.7$, $\lambda = 0.6$, $\omega = 3.5$, and $\rho = 0.25$ all held fixed, while the axisymmetry-breaking parameter ϵ was slowly increased. Each figure shows a single trajectory, plotted as a solid line for $X > 0$ and a broken line for $X < 0$, after initial transients have died away. For each figure there is a large basin of attraction within which different choices of initial point all lead to the same attractor. Previous studies of (1) and analogous systems have traced the succession of bifurcations leading up to an invariant torus [17,19], so we begin here with the axisymmetric invariant torus in Figure 1. In Figure 2 with $\epsilon = 0.025$, one sees that the trajectory has become more concentrated in one band around the torus and is less concentrated in an adjacent band. This may be interpreted as a weak resonant interaction between the two oscillations on the torus. As ϵ increases, there appears to be a saddlenode bifurcation of periodic orbits

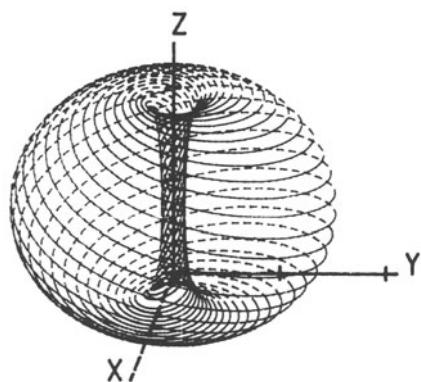


Fig. 1. Torus. Eps=0.0

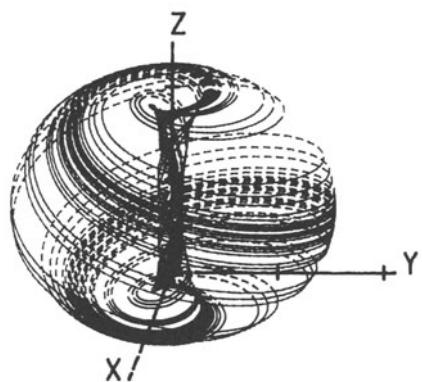


Fig. 2. Torus. Eps=0.025

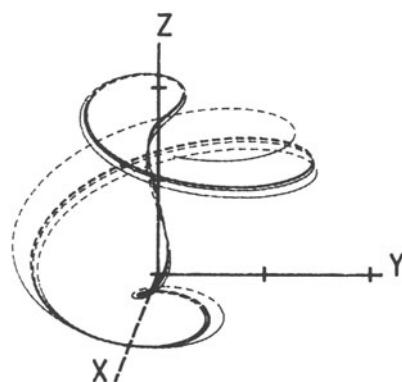


Fig. 3. Period 4. Eps=0.04

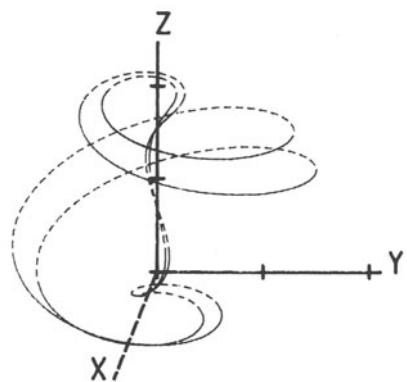


Fig. 4. Period 8. Eps=0.06

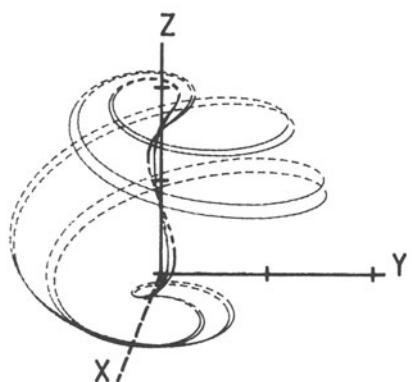


Fig. 5. Period 16. Eps=0.0675

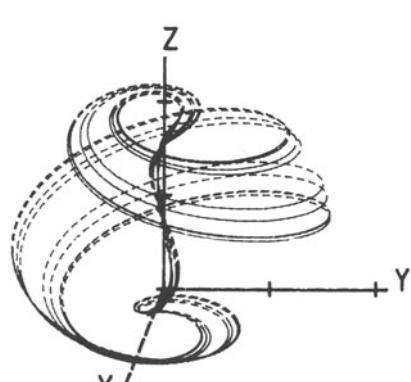


Fig. 6. Period 32. Eps=0.07

within the torus, giving rise to a stable limit cycle and an (unobserved) unstable cycle. Figure 3 shows the period 4 limit cycle observed for $\epsilon = 0.04$, together with an initial transient portion of a typical solution trajectory. We say that the system is now phase-locked, because the period 4 cycle is preserved if we vary the "forcing" frequency ω over a small interval; from the point of view of dynamical systems theory, the system in Figure 3 is structurally stable. Increasing ϵ to 0.06 gives the new attractor in Figure 4, a cycle of period 8. The period 4 cycle evidently still exists (between the two loops of the 8-cycle) but is now unstable, and a period-doubling bifurcation has occurred between $\epsilon = 0.04$ and $\epsilon = 0.06$. In the process, the smooth toroidal manifold of Figures 1 and 2 has been destroyed, because now a small section containing the period 4 and 8 cycles is folded on itself by 180° every four revolutions, rather like a Möbius band. Further increases of ϵ produce a period doubling cascade, with bifurcation values of ϵ spaced more and more closely, see Figure 5 for period 16 with $\epsilon = 0.0675$, and Figure 6 for period 32 with $\epsilon = 0.07$. We have not tried to compute the Feigenbaum constant for this cascade, however see [21].

Beyond this period doubling cascade, the solutions become even more complicated. For $\epsilon = 0.09$ there appears to be a "chaotic band" of period 4, see Figure 7. Solutions from initial points in a large basin approach this band quickly, but within the band the motion is aperiodic and effectively unpredictable over long time intervals. Figure 8 shows the same chaotic band in cross-section, cut by the $X=0$ plane. Figure 8 extends over a much longer time interval than Figure 7; the trajectory (after transients) has intersected the $X=0$ plane 1000 times (500 Poincaré maps), always falling in one of eight "islands" (four for the Poincaré map) which resemble segments of a curve. The trajectory moves among the islands in the sequence indicated by the numbers 1 to 4 in Figure 8. Note that in this sequence the island undergoes an S-folding, then the "S" is flattened vertically onto itself and stretched horizontally, and finally it is mapped onto the original island 1. In this process the central portion of the island reverses orientation, i.e. the outside becomes the inside, and the two ends are folded inward. If the attractor is the limit of this sequence as $t \rightarrow \infty$ then it can not be simply the short curve segment which it appears to be at

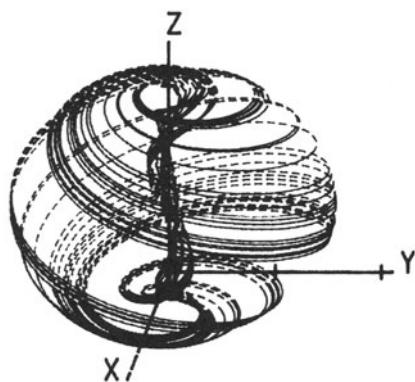


Fig. 7. Chaotic Band. 0.09

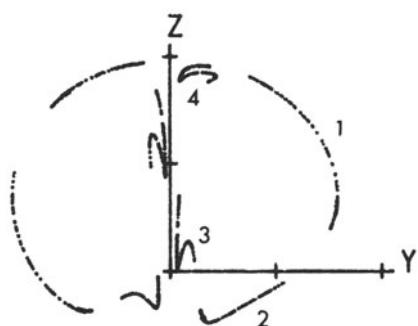


Fig. 8. 500 Poincaré Maps.

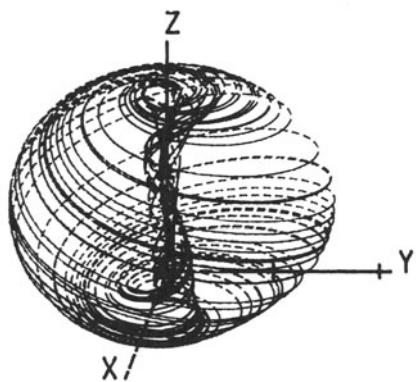


Fig. 9. Folded Torus. 0.1

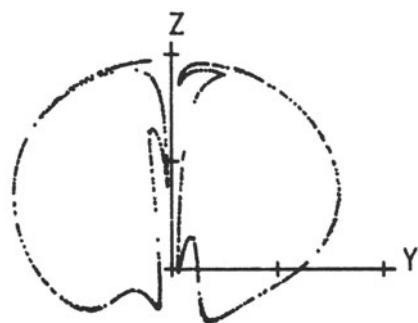


Fig. 10. 500 Poincaré Maps.

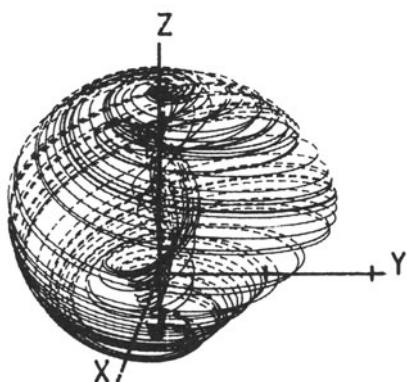


Fig. 11. Turbulence. Eps=0.25

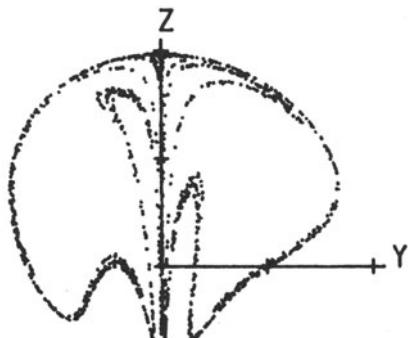


Fig. 12. 700 Poincaré Maps.

first glance, but rather an infinitely folded curve of infinite length, in fact a fractal object [22]. A blown up portion of an island in Figure 8 looks like a product of an interval with a Cantor set, see [21] for an analogous case.

Further increasing ϵ causes the chaotic bands to widen and appear to merge, eventually recreating the torus as in Figures 9 and 10, but now we have not a smooth manifold as in Figures 1 and 2, but a fractal object: a thick torus or bagel [23]. As ϵ increases further, the torus visibly thickens, and the flow becomes more and more "turbulent" or chaotic, see Figures 11 and 12. Still larger ϵ causes the attractor to contact its basin boundary, but it is not clear whether Newhouse sinks [24] are created in the process. The attractor bursts through its basin boundary, and due to the nonaxisymmetry this occurs at first in a very small region which a trajectory may fail to find for a long time. The result is "transient chaos", see Figure 13, with $\epsilon = 0.28$. A typical trajectory now resembles the chaotic trajectory of Figure 11 for a long but finite time, but finally escapes from the chaotic region and is drawn to another attractor, in this case a stable node on the negative Z-axis. This stable node coexists with the torus and chaotic attractors above for smaller values of ϵ , and the boundary separating their basins of attraction can be extremely complicated, probably another fractal set. Another example of coexistence of attractors with complicated (fractal?) basin boundaries is shown in Figure 14, where $\epsilon = 0.07$ and $w = 5$. Two limit cycles of periods 5 and 6, represented by the symbols X and O respectively, coexist within the "ghost" of the former invariant torus. Very small changes in initial points affect which cycle a trajectory eventually approaches, and in fact preliminary work indicates that the basin boundaries may be as complicated as the Julia sets of iterated mappings of the complex plane.

4. Conclusions

Numerical computations have shown evidence of phase locking, period doubling, coexistence of attractors, strange attractors varying from a chaotic band to a thick torus, and transient chaos, all resulting from nonaxisymmetric perturbations of an invariant torus. This provides new

information on the unfoldings of the hysteresis-Hopf singularity. In addition these model equations may help in understanding the general problem of bifurcations from invariant tori in 3D flows. Considerable recent effort has gone into studies of bifurcations of maps of an interval [25], and of a plane [26]; much of the motivation for that work comes from Poincaré maps of flows, but flows themselves have been considered too expensive for direct study. The simple model equation (2) and those in [17,19] remove that obstacle, opening the way to detailed computer-assisted direct studies of the bifurcations from tori to strange attractors.

Figures 7 to 12 show only a few samples of the variety of phase portraits of chaotic or strange attractors for this system. It seems likely that these attractors are topologically inequivalent and thus not structurally stable in the classical sense. Yet in a practical sense they form a continuum, a small perturbation of one of these strange attractors yields a new strange attractor with similar qualitative behavior. This suggests that it may be necessary to devise a new more global definition of structural stability to deal with such strange attractors. The strange attractors studied here withstood perturbations far greater than those that destroyed the invariant tori which gave rise to them. Thus the local existence of quasiperiodic tori proven in [18] may be very local indeed, while these strange attractors, whose existence has not yet been proven rigorously, may in fact persist more globally.

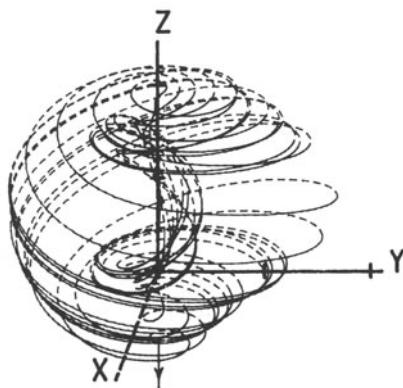


Fig. 13. Transient Chaos.

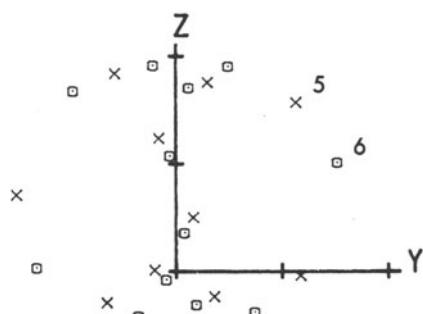


Fig. 14. Coexistence.

REFERENCES

- [1] H.L. Swinney and J.P. Gollub. Hydrodynamic Instabilities and the Transition to Turbulence. Springer-Verlag, New York (1981).
- [2] G. Iooss and D.D. Joseph. Nonlinear Dynamics and Turbulence, Pitman Press. To appear.
- [3] R.E. Meyer. Transition and Turbulence. Academic Press, New York (1981).
- [4] Proceedings of the International Conference on Order in Chaos, Los Alamos, NM. Physica D, V. 7D (1983) Nos. 1-3.
- [5] R. Abraham and J.E. Marsden. Foundations of Mechanics, 2nd Ed. Benjamin/Cummings Reading MA (1978).
- [6] G. Iooss and W.F. Langford. Conjectures on the Routes to Turbulence via Bifurcations. Ann. N.Y. Acad. Sci., V. 357 (1980) pp. 489-505.
- [7] J.P. Gollub and S.V. Benson. Many routes to turbulent convection. J. Fluid Mech., V. 100 (1980) pp. 449-470.
- [8] P.R. Fenstermacher, H.L. Swinney and J.P. Gollub. Dynamical instabilities and the transition to chaotic Taylor vortex flow. J. Fluid Mech., V. 94 (1979) 103-128.
- [9] C. Vidal, J.-C. Roux, S. Bachelart and A. Rossi. Experimental study of the transition to turbulence in the Belousov-Zhabotinsky reaction. Ann. N.Y. Acad. Sci., V. 357 (1980) pp. 377-396.
- [10] W.F. Langford. Periodic and steady-state mode interactions lead to tori. SIAM J. Appl. Math., V. 37 (1979) pp. 22-48.
- [11] W.F. Langford and G. Iooss. Interactions of Hopf and pitchfork bifurcations. Bifurcation Problems and their Numerical Solution, H.D. Mittelmann and H. Weber (Eds). ISNM 54, Birkhauser Verlag, Basel (1980) pp. 103-134.
- [12] J. Guckenheimer. On a codimension two bifurcation. Dynamical Systems and Turbulence, Warwick 1980, D.A. Rand and L.S. Young (Eds). Lecture Notes in Mathematics No. 898, Springer-Verlag, New York (1981) pp. 99-142.
- [13] P.J. Holmes. Unfolding a degenerate nonlinear oscillator: a codimension two bifurcation. Ann. N.Y. Acad. Sci., V. 357 (1980) pp. 473-488.
- [14] S.-N. Chow and J.K. Hale. Methods of Bifurcation Theory. Springer-Verlag, New York (1982).

- [15] J. Guckenheimer and P. Holmes. *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*. Springer-Verlag, New York (1983).
- [16] F. Spirlig. Sequence of bifurcations in a three-dimensional system near a critical point. *J. Appl. Math. Mech. (ZAMP)* V. 34 (1983) pp. 259-276.
- [17] W.F. Langford. A review of interactions of Hopf and steady-state bifurcations. To appear in [2].
- [18] J. Scheurle and J. Marsden. Bifurcation to quasiperiodic tori in the interaction of steady-state and Hopf bifurcations. Preprint, Berkeley, Calif. (1982).
- [19] W.F. Langford. Unfoldings of degenerate bifurcations. *Dynamical Systems, Fractals and Chaos*, P. Fischer and W.R. Smith (Eds), Marcel Dekker. To appear.
- [20] W.F. Langford. Chaotic dynamics in the unfoldings of degenerate bifurcations. *Proceedings of the International Symposium on Applied Mathematics and Information Science*, Kyoto University, Japan (1982).
- [21] J. Perreault. M.Sc. Thesis, Dept. of Mathematics, McGill University, Montreal (1983).
- [22] B.B. Mandelbrot. *The Fractal Geometry of Nature*. W.H. Freeman, San Francisco (1983).
- [23] R.H. Abraham and C.D. Shaw. *Dynamics - The Geometry of Behaviour*. Part 2: *Chaotic Behaviour*. Aerial Press, Santa Cruz (1983).
- [24] S.E. Newhouse. The abundance of wild hyperbolic sets and nonsmooth stable sets for diffeomorphisms. *Publ. Math. IHES*, V. 50 (1979) pp. 101-151.
- [25] P. Collet and J.-P. Eckmann. Iterated Maps on the Interval as Dynamical Systems. *Prog. Phys.* V. 1, Birkhauser Boston (1980).
- [26] D.G. Aronson, M.A. Chory, G.R. Hall, and R.P. McGehee. Bifurcations from an invariant circle for two-parameter families of maps of the plane: a computer assisted study. *Commun. Math. Phys.*, V. 83 (1982) pp. 303-354.

W.F. Langford
 Department of Mathematics and Statistics
 University of Guelph
 Guelph, Ontario
 Canada N1G 2W1

Numerical Treatment of Bifurcation Branches by Adaptive Condensation

Helmut Jarausch and Wolfgang Mackens

1. Introduction

We report on the numerical computation of solution branches $(u, \lambda) \in \mathbb{R}^n \times \mathbb{R}$ of large finite dimensional nonlinear systems

$$(1) \quad Au = F(u, \lambda) \quad ,$$

$F \in C^2(\mathbb{R}^n \times \mathbb{R}, \mathbb{R}^n)$
 A, F_u symmetric, A positive definite

by use of a certain condensation technique, the Condensed Newton with Supported Picard iteration ([4]). Making efficient use of a given (fast) solver for linear systems involving the fixed positive definite $n \times n$ -matrix A the CNSP-approach adaptively reduces problem (1) to a decoupled pair of equations. One of these is high dimensional but controllable by a modified Picard iteration while the other one is low dimensional and may thus be treated by expensive methods such as Newton's iteration.

The small system carries the major information of (1) and thus procedures to compute initial solutions on branches, to follow solution branches and to detect singular points on these branches need only be implemented for this small system. In this way algorithms become applicable that have often been judged as being too expensive for large systems.

The second section contains a description of the CNSP-strategy and its main features for the λ -independent case. (The convergence analysis of CNSP will be published elsewhere. For the moment being cf. [4].)

Section 3 deals with the λ -dependent case. The implementation of branch tracing by arclength continuation is discussed.

Section 4 shows that CNSP can produce information on singular points during branch tracing in a natural unexpensive way.

Each section contains a numerical example demonstrating the performance of the algorithm.

2. The CNSP-strategy

First we want to solve the nonlinear system

$$(2) \quad Au = F(u) \quad , \quad A, F \text{ as above,}$$

by use of a given solver " A^{-1} " of the linear system

$$Au = f \quad , \quad f \in \mathbb{R}^n .$$

Remark: This situation is not too seldom within applications: Imagine A^{-1} to be a fast solver like MG, PCCG, FFT or a fast direct solver incorporating a lot of sophisticated acceleration techniques. Alternatively suppose A^{-1} is a solver for a complicated linear finite element system needing a lot of programming effort. Then, certainly, one is willing to use it.

The first attempt, Picard's iteration,

$$(3) \quad u_{n+1} = A^{-1} F(u_n) \quad ,$$

will diverge in general for the problems of interest here, say the calculation of unstable bifurcation branches, because the spectral radius of $A^{-1} F_u$ will exceed 1. However, very often only a few eigenvalues of $A^{-1} F_u$ are greater than 1 in modulus and are thus responsible for the failure of (3). The idea of CNSP is to extract their influence from (3) to treat it within a separate low dimensional system.

The first step to achieve this resembles the "Alternative Problem"-approach of Cesari [1], Hale [2] and others (cf. the references in [1]). We endow \mathbb{R}^n with a Hilbert space structure, $X := (\mathbb{R}^n, \langle \cdot, \cdot \rangle, \| \cdot \|)$, where the inner product is defined by $\langle u, v \rangle := u^T A v$ and $\| u \| := \langle u, u \rangle^{1/2}$

By means of any linear orthogonal projector

$$P : X \rightarrow PX \quad , \quad u \rightarrow p = Pu$$

and its complementary projector

$$Q := I - P : X \rightarrow QX = P X^\perp \quad , \quad u \rightarrow q = Qu$$

one can split the original equation

$$(4) \quad u - A^{-1} F(u) = 0$$

into the two coupled equations

$$(5p) \quad p - PA^{-1} F(p+q) = 0 ,$$

$$(5q) \quad q = QA^{-1} F(p+q)$$

The "Alternative-Method"-approach chooses P (once and for ever) such that (5q) becomes solvable for the implicit function $q = q(p)$ in a region of interest, and arrives at the equivalent "Alternative Problem" to solve the bifurcation equation

$$(6) \quad p - PA^{-1} F(p+q(p)) = 0$$

in the low dimensional variable p .

Contrary to exploiting the intrinsic coupling of (5p) and (5q) by solving (5q) the present approach tries to locally decouple (5p) and (5q) by a special adaptive choice of P :

Let u be a (starting) approximation for the solution of (4). Then we choose a set of "supports", i. e. a set of orthonormal vectors

$$Z := (z_1, \dots, z_m) \in X^m$$

which approximate the eigenspace corresponding to the m dominating eigenvalues of $A^{-1} F_u(u)$:

$$(7) \quad |\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_m| \geq \gamma_{\max} \geq \lambda_{m+1} \geq \dots$$

The constant $\gamma_{\max} < 1$, which has to be chosen in advance, is the overall linear convergence rate of the final CNSP iteration procedure. The number m is determined by γ_{\max} via (7). (The convergence rate γ_{\max} is limited from below only by the number m of supports one is willing to store.)

Now with

$$P := ZZ^T A : X \rightarrow PX = \text{span } Z$$

$$Q := I - P : X \rightarrow QX , \quad X = PX \oplus QX$$

and $p = Pu =: Zc$, $c \in \mathbb{R}^m$, system (5) reads

$$(8p) \quad \phi(c) := c - Z^T F(Zc + q) = 0$$

$$(8q) \quad q = QA^{-1} F(Zc + q) .$$

For fixed q equation (8p) is m dimensional (in general $m \ll n$), condensing u to its essentially nonlinear components $c := Z^T Au$. Notice that (8p) does not involve A^{-1} any longer. Since (8p) is low dimensional we can try to reduce its error by any expensive iteration, say Newton's iteration

CN	Condensed Newton Step	$c \leftarrow c - (I - Z^T F_u (Zc + q) Z)^{-1} \phi(c)$
----	--------------------------	--

Remark: The executability of this step near a solution will become clear in Section 3.

On the other hand, for fixed c , equation (8q) is a contractive fixed point equation (contraction rate $\sim \gamma_{\max}$), which can be treated by Picard's iteration supported by Q :

SP	Supported Picard Step	$q \leftarrow QA^{-1} F(zc + q)$
----	--------------------------	----------------------------------

Notice that the application of $Q := I - ZZ^T A$ essentially requires only the additional work of evaluating m inner products in n -space.

To see that a combination of these steps leads to a convergent process we introduce the residual vector $R(u) := u - A^{-1} F(u)$, the total residuum $r(u) := \|R(u)\|^2$, the Q -residuum $r_Q(u) := \|QR(u)\|^2$ and the P -residuum $r_P(u) := \|PR(u)\|^2$ ($= \|\phi(c)\|^2$). Observe that

$$(9) \quad r = r_P + r_Q .$$

It is easily seen that (8p), (8q) and the original problem (5) may equivalently be expressed as $r_P = 0$, $r_Q = 0$ and $r = 0$, respectively. A CN-step will reduce r_P , an SP-step will reduce r_Q . The key observation that allows to integrate both steps by means of Seidel-steering (reduction of the dominating residual part) into a complete CNSP-scheme is the following

Decoupling Lemma

With orthonormal supports $Z := (z_1, \dots, z_m) \in X^m$ let $P := ZZ^T A$, $Q := I - P$. Let $u(t) := u + tg$ with $u \in X$, $g \in QX$, $\|g\| = 1$, $0 \leq t \leq s$. Furthermore let L be the Lipschitz constant of $A^{-1} F_u$ on $\{u(t) \mid t \in [0, s]\}$. Then for $r_p(t) := \|PR(u(t))\|^2$ we have

$$(10) \quad |r_p(t) - r_p(0)| \leq (2 \vee r_p(0) + M(t)) M(t).$$

Here $M(t) := t(\frac{t}{2} L + E)$, with

$$(11) \quad 0 \leq E \leq \|A^{-1} F_u(u)Z - ZH\|,$$

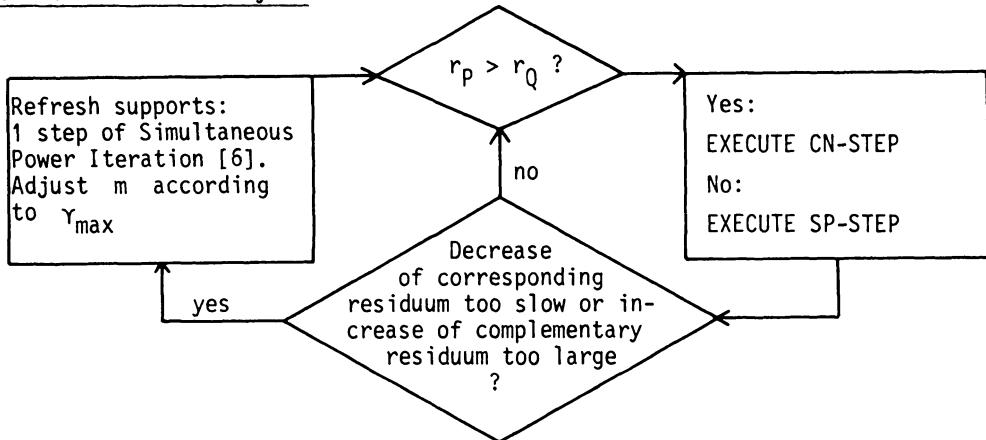
where $H := Z^T F_u(u)Z$ is the Rayleigh quotient matrix of $A^{-1} F_u(u)$ with respect to the Ritz system Z in X .

The same is true for P and Q interchanged.

Thus a small change of $q = Qu$ ($c = Z^T Au$) will not lead to a significant increase of r_p (r_Q) and will thus not spoil the gain of each specific step by increasing the complementary residuum, if E is small. However, the right hand side of (11) is the optimal error bound for the Z -Ritz approximation of eigenvalues of $A^{-1} F_u(u)$, which is zero iff Z spans an eigenspace.

The control of the eigenvector qualities of Z can be done by testing the right hand side of (11), of course. A far more practical procedure is to read the Decoupling Lemma backwards: Z is to be judged as being poor if either a step does not lead to a sufficient reduction of the corresponding residuum or to an intolerable increase of the complementary residuum. In that case Z has to be "refreshed" by one step of simultaneous vector iteration [6] leading to a more adequate $PX \oplus QX$ -decomposition. Notice that we are free to adapt Z at any time without changing the total residuum (9).

The basic CNSP-cycle thus reads as follows.

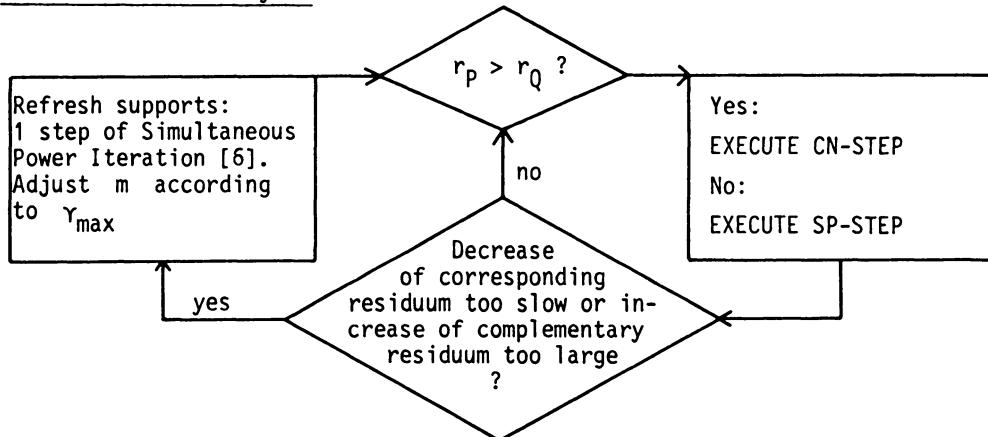
Sketch of the CNSP-Cycle

For a more elaborated version of CNSP - using Powell's [7] method for the (8p)-system and damped Picard iteration for the (8q)-system - global convergence (either to a solution or to a stationary point) can be proved [4] under mild assumptions on the nonlinearity.

To demonstrate the performance of that algorithm we include a convergence history (Fig. 1) for the calculation of the upper (unstable) solution of a discretized version of $-\Delta u = e^u$ on $\Omega = [-1,1]^2$, $u|_{\partial\Omega} = 0$. Usual 5-point differencing on a regular 16×16 grid was used. CNSP was started at a solution of $-\Delta u = \text{constant}$, $u|_{\partial\Omega} = 0$ with a total residuum $r = 7 \cdot 10^{-8}$.

For every step r_P (square) and r_Q (triangle) are given in logarithmic scale. The nature of each specific step is described in the bottom line. A digit denotes refreshment of the corresponding number of supports by one step of the simultaneous power iteration (increasing or decreasing numbers coming from the automatic adjustment of m to $\gamma_{\max} = 0.8$); P denotes a Powell-step, Q a Q-supported Picard-step; X and Y mean rejection of a P- or Q-step, respectively.

The diagram ends at a total residuum of $3 \cdot 10^{-12}$. The first time Picard's iteration is invoked by the algorithm (Step 10) the total residuum has already dropped to $4 \cdot 10^{-3}$. After 27 steps the error has been reduced to the discretization error. We include the rest of the history to show the regular behaviour of the iteration near the solution. Notice also that finally only one support vector is needed.

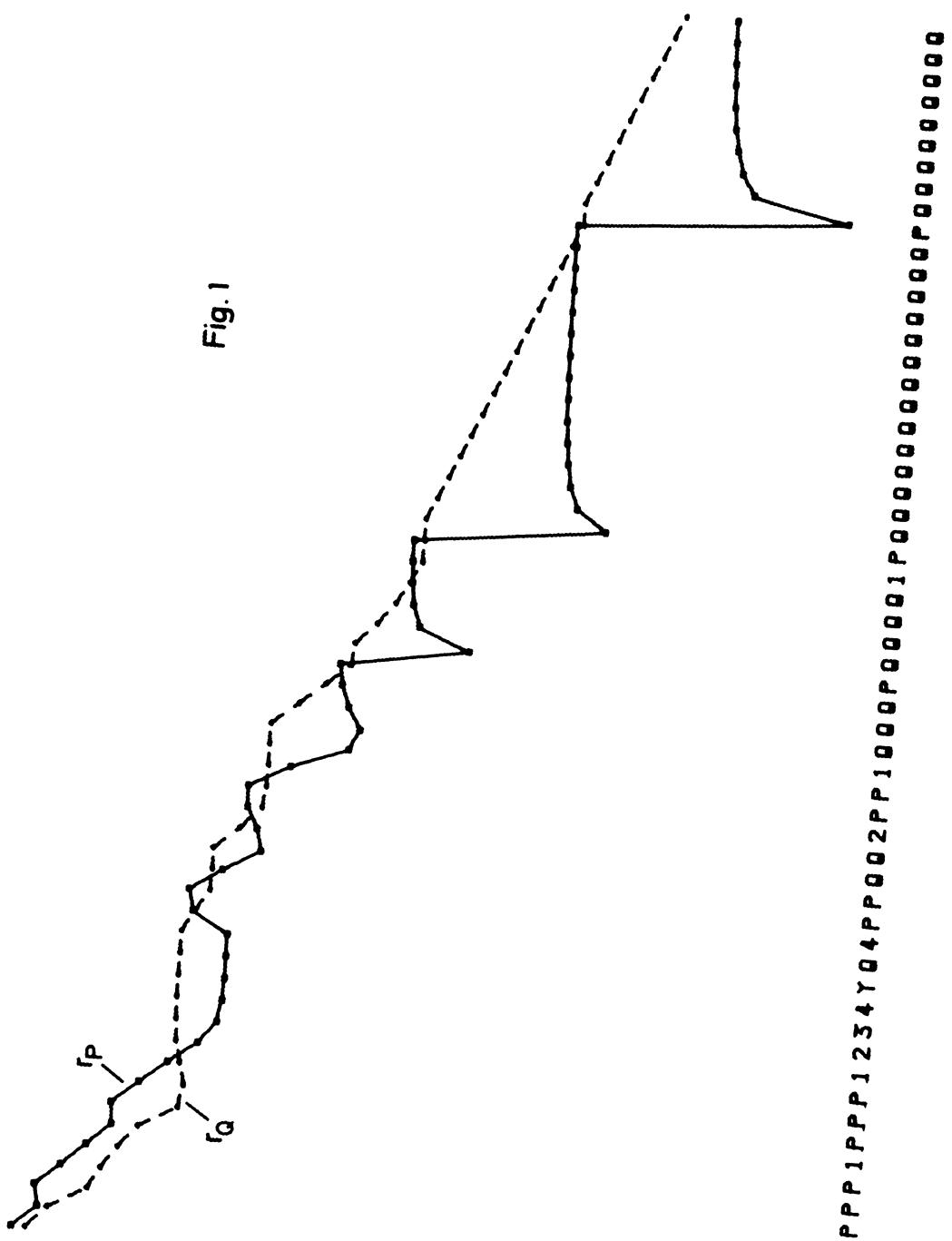
Sketch of the CNSP-Cycle

For a more elaborated version of CNSP - using Powell's [7] method for the (8p)-system and damped Picard iteration for the (8q)-system - global convergence (either to a solution or to a stationary point) can be proved [4] under mild assumptions on the nonlinearity.

To demonstrate the performance of that algorithm we include a convergence history (Fig. 1) for the calculation of the upper (unstable) solution of a discretized version of $-\Delta u = e^u$ on $\Omega = [-1,1]^2$, $u|_{\partial\Omega} = 0$. Usual 5-point differencing on a regular 16×16 grid was used. CNSP was started at a solution of $-\Delta u = \text{constant}$, $u|_{\partial\Omega} = 0$ with a total residuum $r = 7 \cdot 10^{-8}$.

For every step r_p (square) and r_Q (triangle) are given in logarithmic scale. The nature of each specific step is described in the bottom line. A digit denotes refreshment of the corresponding number of supports by one step of the simultaneous power iteration (increasing or decreasing numbers coming from the automatic adjustment of m to $\gamma_{\max} = 0.8$); P denotes a Powell-step, Q a Q-supported Picard-step; X and Y mean rejection of a P- or Q-step, respectively.

The diagram ends at a total residuum of $3 \cdot 10^{-12}$. The first time Picard's iteration is invoked by the algorithm (Step 10) the total residuum has already dropped to $4 \cdot 10^{-3}$. After 27 steps the error has been reduced to the discretization error. We include the rest of the history to show the regular behaviour of the iteration near the solution. Notice also that finally only one support vector is needed.



3. CNSP for the λ -dependent equation

Now we turn to the problem to trace regular solution branches of

$$(12) \quad R(u, \lambda) := u - A^{-1}F(u, \lambda) = 0 \quad (\Leftrightarrow r(u, \lambda) := \|R(u, \lambda)\|^2 = 0).$$

For branches that can be parametrized by λ , within a λ -continuation process, one could certainly use the above CNSP-iteration to get back onto solution branches in u -direction. However, we prefer arclength continuation, since this is even simpler with CNSP than without.

Normally any continuation step to follow a solution branch $\varphi^{-1}(0)$ of an equation $\varphi(x) = 0$ ($\varphi \in C^2(\mathbb{R}^{n+1}, \mathbb{R})$, rank $\varphi' = n$) consists of two parts.

A. Predictor step: At the initial point $x_0 \in \varphi^{-1}(0)$ calculate the tangential vector $t \in \mathbb{R}^{n+1}$ of $\varphi^{-1}(0)$ from

$$(13a) \quad \varphi'(x_0)t = 0, \quad \|t\| \neq 0.$$

Determine by some controlled step width h_t the point

$$(13b) \quad x_p := x_0 + h_t t$$

as a prediction of the new point $x_1 \in \varphi^{-1}(0)$.

B. Corrector step: Go back to the branch by means of an iterative solver of $\varphi(x) = 0$, restricting x to some hypersurface intersecting $\varphi^{-1}(0)$ near x_p (see the relevant papers of this volume).

From a geometric standpoint the best corrector one probably can use is a Pseudo-Inverse-Newton step

$$(14) \quad x_p^0 = x_p, \quad x_p^{k+1} = x_p^k - (\varphi'(x_p^k))^+ \varphi(x_p^k)$$

$(\varphi')^+$ denotes the Moore-Penrose inverse of φ' , since its continuous version [8] finds its way onto $\varphi^{-1}(0)$ orthogonally.

For large systems the Moore-Penrose step (14) is very time consuming and thus usually replaced by some other device. With CNSP there is no objection to the direct use of (14).

We split the problem (12) as before

$$(15p) \quad \phi(c, \lambda) := c - Z^T F(Zc + q, \lambda) = 0 \iff r_p(u, \lambda) := \|PR(u, \lambda)\|^2 = 0$$

$$(15q) \quad q = QA^{-1}F(Zc + q, \lambda) \iff r_Q(u, \lambda) := \|QR(u, \lambda)\|^2 = 0$$

and claim that we have to implement (13) and (14) only for the small system (15p) with $x = (c, \lambda)$. This would be clear, if instead of ϕ we used the "Alternative Method function" (cf. (6))

$$\tilde{\phi}(c, \lambda) := c - Z^T F(Zc + q(c, \lambda), \lambda) ,$$

taking into account the implicit function $q(c, \lambda)$ which solves (15q).

It turns out that - without having aimed at this initially - ϕ actually is a good approximation to $\tilde{\phi}$ if only $q = q(c, \lambda)$ at the evaluation points x_0, x_p, x_p^k of (13) and (14).

Approximation Lemma

If $q = q(c, \lambda)$, i. e. q solves (15q) for fixed c, λ , and $Z \in X^m$ is a system of eigenvectors of $A^{-1}F_u(Zc + q, \lambda)$ then

$$\frac{d^k}{dc^k} \phi(c, \lambda) = \frac{d^k}{dc^k} \tilde{\phi}(c, \lambda) \quad , \quad k = 0, 1, 2 ;$$

$$\frac{d^k}{d\lambda^k} \phi(c, \lambda) = \frac{d^k}{d\lambda^k} \tilde{\phi}(c, \lambda) \quad , \quad k = 0, 1 .$$

Thus $\phi(c, \lambda)$ gives us all the information we need for a continuation of $\tilde{\phi}^{-1}(0)$ in (c, λ) -space provided only we guarantee $q \sim q(c, \lambda)$ at points where ϕ and ϕ' have to be evaluated and that Z is a good eigensystem. The first of these qualities, " $r_Q(u, \lambda)$ small" may be controlled by SP-iteration. The control of Z by "reading the Decoupling Lemma backwards" does no longer work, since the Lemma does not apply to r_Q if λ is changed. However, we have the following

FACT

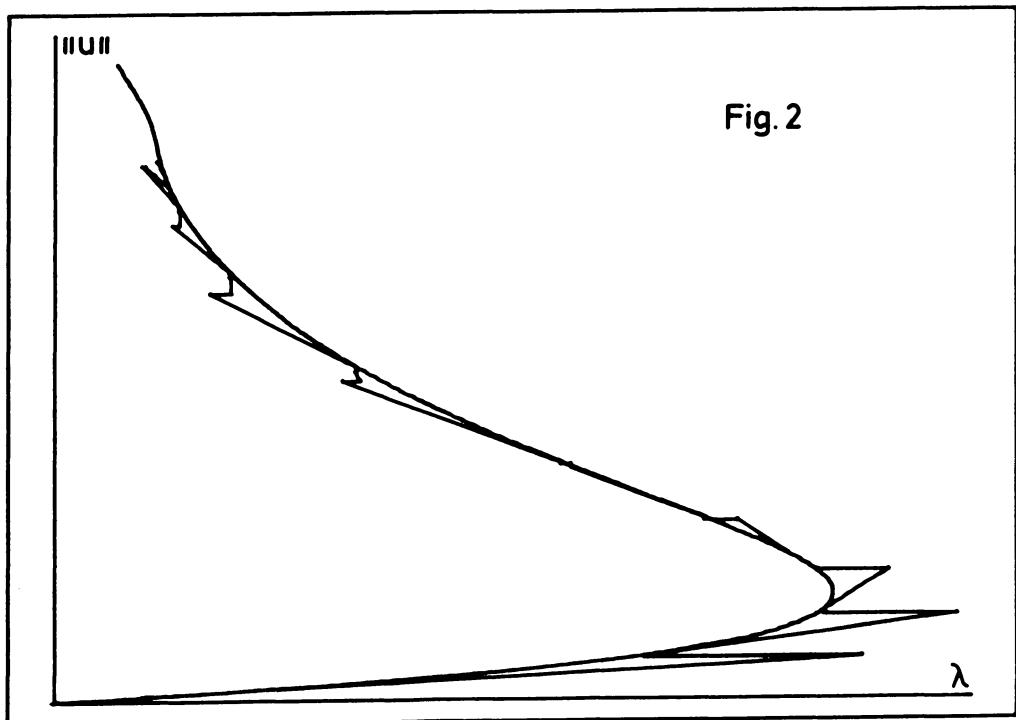
The conclusion of the Decoupling Lemma pertains for r_Q if every change $\Delta\lambda$ of λ is accompanied by the first order q -adaption

$$(16) \quad q \leftarrow q + \Delta\lambda \cdot [I - QA^{-1}F_u]^{-1} QA^{-1}F_\lambda .$$

Numerical evidence shows that a sufficiently good approximation of (16), saving the iterative approximation of $[I - QA^{-1}F_u]^{-1}$, is given by

$$(17) \quad q \leftarrow q + \Delta\lambda \cdot QA^{-1}F_\lambda .$$

Fig. 2 shows two traces of the solution branch of $-\Delta u = \lambda e^u$ on $\Omega = [-1,1]^2$, $u|_{\partial\Omega} = 0$ for 5-point FD-discretisation on a 16×16 grid. (13) and (14) are used for $\phi(c,\lambda)$ applying (17) for every change of λ . The steps of the corrector (14) are interwoven with SP-steps to control $r_Q(u,\lambda)$ and $r_P(u,\lambda)$ to zero by Seidel-steering. (Formulas may be read up in [5].) Control of the step-size h_t in (13) is implemented by requiring r_P to have a prescribed starting value TOL_{start} at prediction points. Iteration (14) is stopped whenever $r(u,\lambda)$ drops below TOL_{stop} . The smooth line consists of 105 trace points



derived with $TOL_{start} = 5 \cdot 10^{-3}$, $TOL_{stop} = 5 \cdot 10^{-7}$. The coarser trace was produced with $TOL_{start} = 2 \cdot 10^1$, $TOL_{stop} = 2 \cdot 10^{-4}$.

Notice that the evaluation of ϕ and

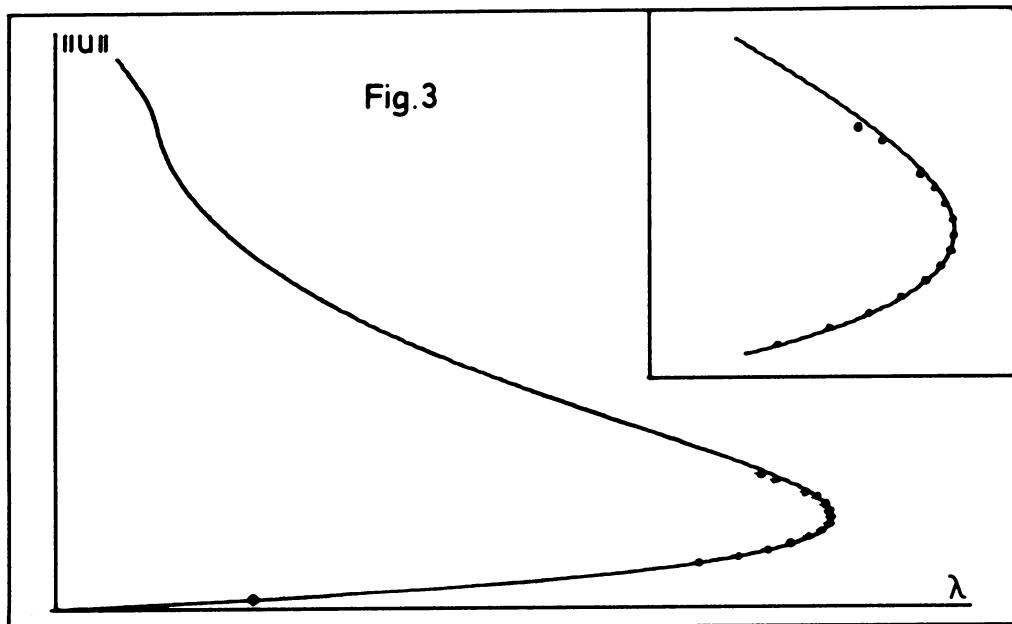
$$(18) \quad \phi' = (I - Z^T F_u(Zc+q, \lambda)Z, -Z^T F_\lambda(Zc+q, \lambda))$$

only needs the evaluation of F and F' . Moreover, if ϕ' has to be approximated by finite differencing, all the additional work is the evaluation of $m+1 \ll n$ additional F -values. For the above example the maximal m for $\gamma_{max} = 0.8$ was found to be $m = 4$.

How well the approximation

$$\tilde{\phi} \sim \phi \quad , \quad (16) \sim (17)$$

works is shown by the last example of this section. The same problem as above is considered. For $\lambda = 0.5$ the lower solution is calculated together with one support vector (exactly). Branch following without SP-iteration produces the result of Fig. 3. The only adaption of q is done by (17). The couple $(Z, Q\Lambda^{-1}E)$ may be rendered as an adaptive reduced basis [3].

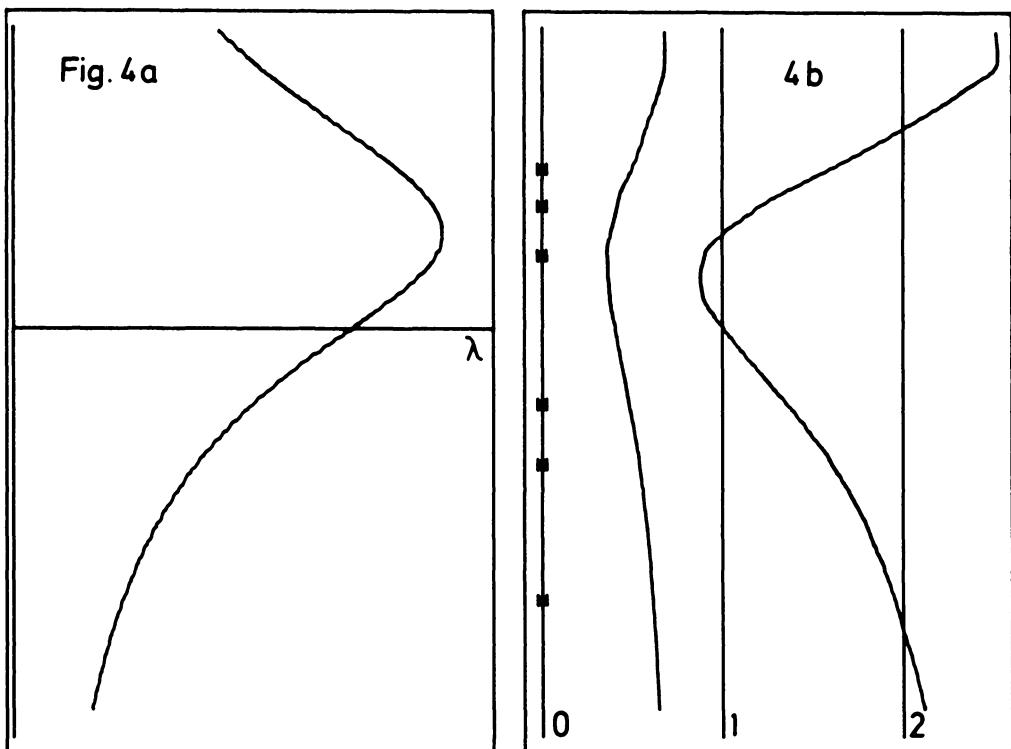


4. Detection of singular points

Notice that the Rayleigh quotient matrix $Z^T F_u Z$ of $A^{-1} F_u$ with respect to the Ritz system Z is permanently available, cf. (18). The control mechanism of CNSP guarantees that those eigenvectors with corresponding eigenvalues close to 1 are always included in the supports Z . Thus singular points, where 1 is an eigenvalue of $A^{-1} F_u$, may be detected by computing the eigenvalues of the tiny $m \times m$ matrix $Z^T F_u Z$.

To give an example we treat the 5-point FD-discretisation on a 16×16 grid of $-\Delta u = \lambda u ((1 - \sin(u)) + u^2)$ on $\Omega = [-1, 1]^2$, $u|_{\partial\Omega} = 0$.

Fig. 4a shows the first nontrivial solution branch bifurcating from the trivial solution. For this trace only two supports have been used. Fig. 4b gives a plot of the eigenvalues of the 2×2 -matrix $Z^T F_u Z$. The turning point and the bifurcation point of the branch in Fig. 4a show up precisely when an eigenvalue crosses the value 1. The x-marks in Fig. 4b indicate where refreshment of Z has been done.



References

- [1] Cesari, L.: Functional Analysis, Nonlinear Differential Equations, and the Alternative Method. pp 1 - 197 in "Nonlinear Functional Analysis and Differential Equations", L. Cesari, R. Kannan and J. D. Schuur (eds.), Marcel Dekker Inc., New York 1976.
- [2] Chow, S.-N. and J. K. Hale: Methods of Bifurcation Theory. Springer-Verlag, New York - Heidelberg - Berlin 1982.
- [3] Fink, J. P. and W. C. Rheinboldt: On the Error Behavior of the Reduced Basis Technique for Nonlinear Finite Element Approximations, ZAMM 63 (1983) 21 - 23.
- [4] Jarausch, H. and W. Mackens: CNSP - A fast globally convergent scheme to compute stationary points of elliptic variational problems. Bericht Nr. 15 des Instituts für Geometrie und Praktische Mathematik der RWTH Aachen, 1982.
- [5] Jarausch, H. and W. Mackens: Computing solution branches by use of a Condensed Newton - Supported Picard iteration scheme. To appear in Tagungsband ZAMM 1984.
- [6] Parlett, B. N.: The Symmetric Eigenvalue Problem. Prentice Hall, Englewood Cliffs, 1980.
- [7] Powell, M. J. D.: A Hybrid Method for Nonlinear Equations. pp 87 - 114 in "Numerical Methods for Nonlinear Algebraic Equations", Ph. Rabinowitz (ed.), Gordon and Breach, London - New York - Paris 1970.
- [8] Tanabe, K.: Continuous Newton-Raphson method for solving an underdetermined system of nonlinear equations. Nonlinear Analysis, Theory, Methods and Applications 3 (1979) 495 - 503.

H. Jarausch, W. Mackens, Institut für Geometrie und Praktische Mathematik,
Techn. Hochschule Aachen, Templergraben 55, 5100 Aachen

NUMERICAL DETERMINATION OF MULTIPLE BIFURCATION POINTS

Reinhard Menzel

1. Introduction

Consider the finite-dimensional nonlinear system

$$G(x, t) = 0 \quad (1)$$

of n equations in n variables $x = (x_1, \dots, x_n)^T$ depending on an additional real parameter t and let $G: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ be a sufficiently smooth mapping. In some neighbourhood $D \subset \mathbb{R}^n \times \mathbb{R}$ of (x^*, t^*) with $G(x^*, t^*) = 0$ we assume the system to have a one-dimensional smooth manifold

$$\mathcal{C} := \{(x, t) \in D : G(x, t) = 0\}$$

of solutions.

It is well known that if the Jacobian $G_x(x^*, t^*)$ is nonsingular the implicit function theorem ensures the existence of a unique branch $(x(t), t)$ with $G(x(t), t) = 0$ defined for all t , $|t - t^*|$ sufficiently small. On the other hand in case of a singular Jacobian $G_x(x^*, t^*)$ in a neighbourhood of (x^*, t^*) we have a variety of possible sets of solutions of (1). The point (x^*, t^*) could be a so-called turning point, i.e. a point at which the t -component of the branch becomes extremal. More exactly, in this situation the following conditions hold:

$$\text{rank}(G_x(x^*, t^*)) = n-1 \quad (2)$$

$$G_t(x^*, t^*) \notin \mathbb{R}(G_x(x^*, t^*)) \quad (3)$$

$$G_{xx}(x^*, t^*)v^*v^* \notin \mathbb{R}(G_x(x^*, t^*)) \quad (4)$$

where $\mathbb{R}(.)$ denotes the range and v^* the (except for the sign) unique solution of

$$G_x(x^*, t^*)v = 0, \|v\| = 1. \quad (5)$$

The point (x^*, t^*) also could be a bifurcation point where two or more different branches of solutions intersect. So-called simple bifurcation points can be characterized by (2) and the conditions

$$G_t(x^*, t^*) \in \mathbb{R}(G_x(x^*, t^*)) \quad (6)$$

$$H''(u^*)(\begin{smallmatrix} v^* \\ 0 \end{smallmatrix})w^* \notin \mathbb{R}(G_x(x^*, t^*)) \quad (7)$$

where for simplification as in the following the mapping $H: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^n$ is defined by

$$H(u) := G(x, t), u := (x, t)^T;$$

the vectors $(\begin{smallmatrix} v^* \\ 0 \end{smallmatrix})$ and w^* span the null space $\mathcal{E}(H'(u^*))$ with $u^* := (x^*, t^*)^T$.

For several applications the numerical determination of turning and bifurcation points is of special interest. In the last few years effective methods for computing turning and simple bifurcation points have been proposed, see, e.g., [6, 12, 13] and [8, 11, 17].

The present communication is concerned with the numerical determination of multiple bifurcation points where $(x^*, t^*) \in \mathfrak{C}$ here will be called in such a manner if the condition

$$\text{rank}(G_x(x^*, t^*)) < n-1 \quad (8)$$

holds. As defining equations of multiple bifurcation points overdetermined nonlinear systems

$$F(z) = 0 \quad (9)$$

are presented. Sufficient conditions can be given such that the GAUSS-NEWTON-Iteration

$$z^{k+1} := z^k - [F'(z^k)^T F'(z^k)]^{-1} F'(z^k)^T F(z^k) \quad (10)$$

is at least Q-linearly convergent to the solution z^* of (9). In case of small rank deficiency of $F'(z^*)$ one is even able

to ensure Q-superlinear convergence by solving an auxiliary problem of higher dimension. These results on solving over-determined consistent systems can be considered as generalizations of those obtained in [15] and [18].

2. Defining equations of multiple bifurcation points

In the following for the point $(x^*, t^*) \in \mathfrak{C}$ we suppose that the condition

$$\text{rank}(G_x(x^*, t^*)) = n-r \quad (11)$$

holds. The null spaces of $G_x(x^*, t^*)$ and $G_x(x^*, t^*)^T$ are assumed to be spanned by the vectors $\{v^{1,*}, \dots, v^{r,*}\}$ and $\{\psi^{1,*}, \dots, \psi^{r,*}\}$, respectively, where $(v^{i,*})^T v^{j,*} = (\psi^{i,*})^T \psi^{j,*} = \delta_{ij}$, $i, j = 1, \dots, r$ and δ_{ij} denotes the Kronecker-symbol. Therefore we have

$$\begin{aligned} \mathcal{E}(G_x(x^*, t^*)) &= \text{span}\{v^{1,*}, \dots, v^{r,*}\} \\ \mathcal{E}(G_x(x^*, t^*)^T) &= \text{span}\{\psi^{1,*}, \dots, \psi^{r,*}\} . \end{aligned}$$

Moreover, we suppose that $H'(u)$ is of full rank for all u in a neighbourhood of u^* with $u \neq u^*$.

Now consider the overdetermined nonlinear systems of equations

$$\begin{aligned} F_1(\bar{z}) := \begin{bmatrix} G(x, t) \\ G_x(x, t)v^i \\ \vdots \\ (v^i)^T v^j - \delta_{ij} \end{bmatrix} &= 0, \quad i, j = 1, \dots, r \\ \bar{z} := (x, t, v^1, \dots, v^r)^T \end{aligned} \quad (12)$$

and

which in the present case is a rank deficient one. This type of problems has been analyzed, e.g., in [1,2,3] and only under stringent assumptions the convergence of the BEN-ISRAEL-Iteration

$$z^{k+1} := z^k - F'(z^k)^+ F(z^k) \quad (17)$$

could be shown; $F'(z^k)^+$ denotes the MOORE-PENROSE-Inverse of $F'(z^k)$. For the special case of a consistent least-squares problem, however, which is of interest here, under more natural assumptions one is able to get at least Q-linear convergence.

3.1. Theorem: Let $F: R^S \rightarrow R^m$, $m \geq s$ be a C^3 -mapping and suppose the existence of a $z^* \in R^S$ with $F(z^*) = 0$. Assume

$$\text{rank}(F'(z^*)) < s$$

and let $R^S = N \oplus Y$ where N denotes the null space of $F'(z^*)$. Define like in [15] for positive constants ρ, θ, Φ the quantities

$$B_\rho(z^*) := \{z \in R^S : \|z - z^*\| \leq \rho\}$$

$$C_\theta(z^*) := \{z \in R^S : \|P_Y(z - z^*)\| \leq \theta \|P_N(z - z^*)\|\}$$

$$T_\Phi(z^*) := \{z \in R^S : \|(I - P_L)P_N(z - z^*)\| \leq \Phi \|P_N(z - z^*)\|\}$$

$$W_{\rho\theta\Phi}(z^*) := B_\rho(z^*) \cap C_\theta(z^*) \cap T_\Phi(z^*)$$

where L is a one-dimensional subspace of N and P_N , $P_Y = I - P_N$, P_L denote ortho-projectors on N , Y , L , respectively.

For L let exist positive constants $c_1, c_2, c_3, \bar{\rho}, \bar{\theta}, \bar{\Phi}$ such that

(i) $\|F''(z^*)py\| \geq c_1 \|p\| \|y\|$ for all $p \in L$ and $y \in R^S$

(ii) $\|F'(z)v + F'(z)w\| \geq c_2 \|F'(z)v\| + c_3 \|F'(z)w\|$ for all $v \in N$, $w \in Y$, $z \in W_{\rho\theta\Phi}(z^*)$

are fulfilled.

Then there are positive constants $\rho^*, \theta^*, \Phi^*, \lambda^*$ such that $F'(z)$ for $z \in W_{\rho^*\theta^*\Phi^*}(z^*)$, $z \neq z^*$ is of full rank, the iterates (10) remain in $W_{\rho^*\theta^*\Phi^*}(z^*)$ for all k and converge to z^* provided that $z^0 \in W_{\rho^*\theta^*\Phi^*}(z^*) \cap T_{\lambda^*}(z^*)$ holds. The convergence is at least Q-linear with asymptotical rate 1/2.

For the proof of the theorem see [10].

In the special case

$$\text{rank}(F'(z^*)) = s-1 \quad (18)$$

by means of an extension of WEBER/WERNER's method (see [18]) for overdetermined consistent nonlinear systems of equations one is even able to get Q-superlinear convergence.

3.2. Theorem: Let $F: R^s \rightarrow R^m$, $m \geq s$ be a C^2 -mapping and suppose the existence of a $z^* \in R^s$ with $F(z^*) = 0$. Assume (18) and

$F''(z^*)v^*v^* \notin R(F'(z^*))$
where $v^* \in R^s$ spans the null space of $F'(z^*)$.

Then the auxiliary problem

$$Q(y) = 0 \quad (19)$$

with

$$Q(y) := \begin{bmatrix} F(z) \\ F'(z)v \\ v^Tv - 1 \end{bmatrix}, \quad y := \begin{bmatrix} z \\ v \end{bmatrix} \quad (20)$$

$Q: R^{2s} \rightarrow R^{2m+1}$ has the solution $y^* := (z^*, v^*)^T$ and the Jacobian $Q'(y^*)$ is of full rank.

This assertion is a special case of a more general result obtained in [10].

In consequence of the preceding theorem the auxiliary problem (19) is a well posed one and therefore the Q-superlinear convergence of a sequence $\{y^k\}$ generated by a GAUSS-NEWTON-like method to y^* under suitable conditions

$$F_2(\tilde{z}) := \begin{pmatrix} G(x, t) \\ G_x(x, t)^T \psi^i \\ \vdots \\ (\psi^i)^T \psi^{j-\delta_{ij}} \\ \vdots \end{pmatrix} = 0, \quad i, j = 1, \dots, r \quad (13)$$

$\tilde{z} := (x, t, \psi^1, \dots, \psi^r)^T$

defined by the mappings $F_i: \mathbb{R}^{n_1} \rightarrow \mathbb{R}^{n_2}$, $i=1,2$, $n_1:=n(r+1)+1$, $n_2:=(r+1)(n+r/2)$. Obviously, the vectors $\bar{z}^*:=(x^*, t^*, v^1, \dots, v^r)^T$ and $\tilde{z}^*:=(x^*, t^*, \psi^1, \dots, \psi^r)^T$ are isolated solutions of (12) or (13), respectively, in the sense that there are neighbourhoods of \bar{z}^* and \tilde{z}^* in which these solutions are unique. Note that by solving the systems (12), (13) beside the bifurcation point (x^*, t^*) simultaneously we get the vectors spanning the null spaces of $G_x(x^*, t^*)$ and $G_x(x^*, t^*)^T$. These vectors especially are of interest for solving the algebraic bifurcation equations to approximate the bifurcating from (x^*, t^*) solution branches of (1).

For the numerical determination of \bar{z}^* and \tilde{z}^* the rank deficiency of the Jacobians $F'_1(\bar{z}^*)$ and $F'_2(\tilde{z}^*)$, respectively, is of importance. In the following we will get general assertions on this quantity. First of all we consider the situation in which the conditions (11) and (6) hold. Then the null space of $H'(u^*)$ has the form

$$N_1 := \mathcal{E}(H'(u^*))_{1,1} = \text{span}\{(v_0^1, \dots, v_0^r, v_1^{r+1})\} \quad (14)$$

with some $v^{r+1} \in \mathbb{R}^n$ and $v^i, i=1, \dots, r$ defined like above. Throughout this paper by (V) we will denote the following assumption.

Assumption (V): Relative to a given linear manifold $M \subset \mathbb{R}^{n+1}$ for all $p \in M$ with $p \neq 0$ and at least for one k , $1 \leq k \leq r$ the condition

$$H''(u^*)(v_0^k, p) \notin \mathbb{R}(G_x(x^*, t^*)) \quad (15)$$

is fulfilled.

About rank deficiency of the Jacobians $F'_1(\bar{z}^*)$ and $F'_2(\tilde{z}^*)$ we have the following theorem.

2.1. Theorem: For $r > 1$ suppose that the conditions (11), (6) and the assumption (V) with $M = N_1$ are satisfied.

Then $F'_1(\bar{z}^*)$ and $F'_2(\tilde{z}^*)$ have rank deficiency $r(r-1)/2$.

The proof of the theorem has been given in [9].

Now we consider the situation characterized by the conditions (11) and (3). Here the null space of $H'(u^*)$ is of the form

$$N_2 := \text{span}\{v_0^{1,*}, \dots, v_0^{r,*}\}. \quad (16)$$

Analogously to Theorem 2.1 in this case we obtain the following assertion.

2.2. Theorem: Suppose that the conditions (11),(3) and the assumption (V) with $M = N_2$ are satisfied.

Then $F'_1(\bar{z}^*)$ and $F'_2(\tilde{z}^*)$ have rank deficiency $r(r-1)/2$.

For the proof again compare [9]. Note that the defining equations (12),(13) can be considered as generalizations of those proposed in [11,16,17] for turning and simple bifurcation points, respectively.

3. Numerical solution of the defining equations

The defining equations (12),(13) are overdetermined nonlinear systems of the form (9) having a solution z^* at which in consequence of the preceding theorems the Jacobian $F'(z^*)$ is not of full rank.

In general, an approximation to z^* is computed by solving the corresponding nonlinear least-squares problem

can be ensured.

References

- [1] BEN-ISRAEL, A.: A Newton-Raphson method for the solution of systems of equations.
J.Math.Anal.Appl. 15, 243-252(1966).
- [2] BOGGS, P.T.: The convergence of the Ben-Israel-Iteration for nonlinear least-squares problems.
Math.Comp. 30, 512-522(1976).
- [3] BOGGS, P.T., DENNIS, J.E.: A stability analysis for perturbed nonlinear iterative methods.
Math.Comp. 30, 199-215(1976).
- [4] DECKER, D.W., KELLEY, C.T.: Newton's method at singular points.
SIAM J.Numer.Anal. 17, 66-70(1980).
- [5] GRIEWANK, A., OSBORNE, M.R.: Newton's method for singular problems when the dimension of the null space is >1 .
SIAM J.Numer.Anal. 18, 145-149(1981).
- [6] MELHEM, R.G., RHEINBOLDT, W.C.: A comparision of methods for determining turning points of nonlinear equations.
Computing 29, 201-226(1982).
- [7] MENZEL, R., SCHWETLICK, H.: Zur Lösung parameterabhängiger nichtlinearer Gleichungen mit singulären Jacobi-Matrizen.
Numer.Math. 30, 65-79(1978).
- [8] MENZEL, R., PÖNISCH, G.: A quadratically convergent method for computing simple singular roots and its application to determining simple bifurcation points.
To appear in Computing.
- [9] MENZEL, R.: Ob odnom podchode k čislennomu opredeleniju mnogokratnyh toček wewlenija w konečnomernom slučae.
In preparation.
- [10] MENZEL, R.: On solving overdetermined nonlinear systems of equations in case of rankdeficient Jacobians.
In preparation.
- [11] MOORE, G.: The numerical treatment of nontrivial bifurcation points.
Numer.Funct.Anal.Optimiz. 2, 441-472(1980).
- [12] PÖNISCH, G., SCHWETLICK, H.: Computing turning points of curves implicitly defined by nonlinear equations depending on a parameter.
Computing 26, 107-121(1981).

- [13] PÖNISCH, G., SCHWETLICK, H.: Ein lokal überlinear konvergentes Verfahren zur Bestimmung von Rückkehrpunkten implizit definierter Raumkurven.
Numer.Math. 38, 455-466(1982).
- [14] REDDIEN, G.W.: On Newton's method for singular problems.
SIAM J.Numer.Anal. 15, 993-996(1978).
- [15] REDDIEN, G.W.: Newton's method and high order singularities.
Comp.Math.Appl. 5, 79-86(1979).
- [16] SEYDEL, R.: Numerical computation of branch points in nonlinear equations.
Numer.Math. 33, 339-352(1979).
- [17] WEBER, H.: On the numerical approximation of secondary bifurcation problems.
In: Allgower,E.L., Glashoff,K., Peitgen,H.O. (eds.):
Numerical solution of nonlinear equations. Lect. Notes in Math. 878, Springer-Verlag, 1981, pp. 407-425.
- [18] WEBER, H., WERNER, W.: On the accurate determination of nonisolated solutions of nonlinear equations.
Computing 26, 315-326(1981).

Reinhard Menzel
Technische Universität Dresden
Sektion Mathematik
DDR-8027 Dresden, Mommsenstraße 13

CONTINUATION NEAR SYMMETRY-BREAKING BIFURCATION POINTS

H. D. Mittelmann

Abstract One class of examples of nonlinear boundary value problems whose discretizations in general inherit bifurcation properties from the continuous case are problems with symmetries. We analyze the behaviour of a generalized inverse iteration method for the numerical solution of these problems near a symmetry-breaking pitchfork bifurcation point. While the computation of symmetric solutions does not represent any difficulties it is shown that a suitable form of the algorithm may be used to determine the bifurcating branch of non-symmetric solutions in a very stable and efficient way. The theoretical results are confirmed by numerical results for a classical example.

1. Introduction

We study a constructive method for determining paths of solutions of a class of nonlinear functional equations, say

$$g(u, \lambda) = 0, \quad (1.1)$$

where $g : X \times \mathbb{R} \rightarrow X$, X a Banach space.

It is often not advantageous, not uniquely possible or not possible at all to parametrize the solutions u of (1.1) by λ . Hence it was suggested (see, for example, [6]) to introduce an additional parameter $s \in \mathbb{R}$ and add a suitable normalization condition

$$\begin{aligned} g(u(s), \lambda(s)) &= 0, \\ N(u(s), \lambda(s), s) &= 0. \end{aligned} \quad (1.2)$$

An algorithm of inverse iteration type was introduced in a more special setting in [6] to solve the following form of (1.2).

$$g(u(\rho), \lambda(\rho)) = \lambda L u - f(u) = 0, \quad (1.3a)$$

$$N(u(\rho), \rho) = \|u(\rho)\|^2 - \rho^2 = 0, \quad (1.3b)$$

where $L : X \rightarrow X$ is a linear selfadjoint positive definite operator, $f : X \rightarrow X$ a nonlinear operator and $\|\cdot\|$ the norm introduced by L . A finite-dimensional version of this algorithm was successfully applied in [7,8] to compute solution branches for discretizations of boundary value problems for ordinary and partial differential equations. The algorithm compared favourably with

several standard methods to solve (1.1) in the case that simple or nonsimple turning points are present, near primary bifurcation points and for following spurious solution curves. Multigrid versions of the algorithm were developed and tested in [9,10]. A continuation strategy for those parts of the solution paths that may be parametrized by ρ was given in [10].

The normalization (1.3b) is, of course, well-known. The success of the generalized inverse iteration is not so much caused by using that but seems to be due to the fact that an inverse iteration is used to solve (1.3) instead of Newton's method. We restrict ourselves now to the case $L = I$, I the identity operator on X . Let $u^* \in X'$ be such that $u^*u = \|u\|$. For a given $\rho > 0$ and $u^{(0)} \in X$, $\|u^{(0)}\| = \rho$, the generalized inverse iteration generates the iterates $(u^{(k)}, \lambda^{(k)})$ according to

$$u^{(k)} = \rho \frac{\tilde{u}^{(k-1)}}{\|\tilde{u}^{(k-1)}\|}, \quad \lambda^{(k)} = u^{(k)*} f(u^{(k)}) / \rho^2$$

$$\tilde{u}^{(k-1)} = u^{(k-1)} + \delta u^{(k-1)}, \quad k = 1, 2, \dots, \text{ where} \quad (1.4)$$

$$\begin{bmatrix} g_u^{(k-1)} & g_\lambda^{(k-1)} \\ u^{(k-1)*} & 0 \end{bmatrix} \cdot \begin{bmatrix} \delta u^{(k-1)} \\ * \end{bmatrix} = - \begin{bmatrix} g^{(k-1)} \\ 0 \end{bmatrix}$$

Here $g^{(k-1)} = g(u^{(k-1)}, \lambda^{(k-1)})$ and $g_u^{(k-1)}, g_\lambda^{(k-1)}$ are the Fréchet derivatives evaluated at the same arguments.

It may be shown (cf. [7]) that (1.4) is quadratically convergent to regular or simple turning points $u(\rho)$. In addition to this local behaviour the algorithm is also robust in the sense that larger stepsizes for the continuation in ρ may be taken without causing lack of convergence, divergence or convergence to a solution on a different branch. In the continuation from ρ to $\rho + \delta\rho$ the 'predictor' step used is simply renormalization of $u(\rho)$. The matrix in (1.4) is regular in the above mentioned points but in general singular in bifurcation points. Numerical experience shows that there as expected the performance of (1.4) deteriorates. We shall analyze this and propose a modification of (1.3) in the case that the singular point is a symmetry-breaking pitchfork bifurcation point. We show theoretically that the behaviour has improved considerably and present finally some numerical

results.

The contents of the following sections are

2. Simple bifurcation points
3. Symmetry-breaking bifurcation
4. The augmented system
5. The inverse iteration
6. An example
7. Numerical implementation
8. Results

2. Simple bifurcation points

In a simple bifurcation point $(u_0, \lambda_0) = (u(\rho_0), \lambda(\rho_0))$ of (1.3a) we have

$$\begin{aligned} N(g_u^0) &= \text{span}\{\phi_0\}, \quad \|\phi_0\| = 1, \\ R(g_u^0) &\text{ is closed and codim } R(g_u^0) = 1. \end{aligned} \tag{2.1}$$

$g_u^0 = g_u(u_0, \lambda_0)$ is thus a Fredholm operator of index zero and its adjoint operator $g_u^{0*} : X' \rightarrow X'$ satisfies

$$N(g_u^{0*}) = \text{span}\{\phi_0^*\}, \quad R(g_u^{0*}) = \{x \mid \phi_0^* x = 0\}. \tag{2.2}$$

In addition we assume the zero eigenvalue of g_u^0 to have algebraic multiplicity one and thus we may take

$$\phi_0^* \phi_0 = 1. \tag{2.3}$$

We call (u_0, λ_0) a pitchfork bifurcation point if

$$\phi_0^* g_\lambda^0 = 0, \quad \phi_0^* g_{uu} \phi_0 \phi_0 = 0, \tag{2.4}$$

$$\phi_0^* (g_{u\lambda}^0 \phi_0 + g_{uu}^0 v_0 \phi_0) \neq 0 \tag{2.5}$$

where v_0 is the unique solution of

$$g_u^0 v_0 + g_\lambda^0 = 0, \quad \phi_0^* v_0 = 0. \tag{2.6}$$

In order to analyze the behaviour of problems (1.3a) respectively (1.3a), (1.3b) in the neighbourhood of pitchfork bifurcation points we proceed as in [3] to which we shall repeatedly refer. In particular we study the growth of the Frechét derivatives as (u_0, λ_0) is approached along one of the branches intersecting there.

Lemma 2.7 Suppose $g(u, \lambda)$ is twice continuously differentiable with respect to both u and λ and $g_u^0 = g_u(u_0, \lambda_0)$ is a Fredholm operator of index zero with a simple zero eigenvalue and

$$g_u^0 \phi_0 = 0, \quad g_u^{0*} \phi_0^* = 0. \quad (2.8)$$

Then for $|\rho - \rho_0| < \delta$, some $\delta > 0$, there exists a twice continuously differentiable pair $(\beta(\rho), \phi(\rho))$ such that

$$g_u(u(\rho), \lambda(\rho))\phi(\rho) = \beta(\rho)\phi(\rho), \quad (2.9a)$$

$$\beta(\rho_0) = 0, \quad \phi(\rho_0) = \phi_0. \quad (2.9b)$$

Proof We consider the system of equations

$$H(\beta, \phi, \rho) \equiv \begin{bmatrix} g_u(u(\rho), \lambda(\rho))\phi - \beta\phi \\ \phi_0^*\phi - 1 \end{bmatrix} = 0. \quad (2.10)$$

This system has a solution $(0, \phi_0, \rho_0)$ and the Frechét derivative there with respect to (ϕ, β) is

$$H'(0, \phi_0, \rho_0) = \begin{bmatrix} g_u^0 & \phi_0 \\ \phi_0^* & 0 \end{bmatrix}. \quad (2.11)$$

This operator is nonsingular by Lemma I in [3] and thus the result follows from the implicit function theorem.

We can state and prove now for the sake of completeness a result on the growth of g_u^{-1} .

Theorem 2.12 Let $g(u, \lambda)$ be twice continuously differentiable with respect to u and λ . Let $u(\rho)$ be a twice continuously differentiable solution arc through the pitchfork bifurcation point (u_0, λ_0) . Then, for some $\delta > 0$ and $K > 0$ and each ρ in $0 < |\rho - \rho_0| < \delta$

$$\|g_u^{-1}(u(\rho), \lambda(\rho))\| \leq \frac{K}{|\rho - \rho_0|^p}, \quad (2.13)$$

where $p = 1$ if $\dot{\lambda}_0 = \frac{d\lambda}{d\rho}(\rho_0) \neq 0$ and $p \geq 2$ if $\dot{\lambda}_0 = 0$.

Proof Differentiating (1.3a) with respect to ρ and evaluating at ρ_0 yields

$$g_u^0 \dot{u}_0 + g_\lambda^0 \dot{\lambda}_0 = 0. \quad (2.14)$$

If $\dot{\lambda}_0 \neq 0$ we have $\dot{u}_0 = \dot{\lambda}_0 v_0$ from (2.6). Differentiating (2.9a) with respect to ρ , evaluating at ρ_0 and multiplying by ϕ_0^* we obtain

$$\begin{aligned} & \phi_0^* g_u^0 \dot{\phi}_0(\rho_0) + \phi_0^*(g_{uu}^0 \dot{u}_0 + g_u^0 \lambda \dot{\lambda}_0) \phi_0 \\ &= \dot{\beta}(\rho_0) \phi_0^* \phi_0 + \beta(\rho_0) \phi_0^* \dot{\phi}_0(\rho_0). \end{aligned} \quad (2.15)$$

Recalling (2.1) - (2.3) and (2.9b) we have

$$\dot{\beta}(\rho_0) = \dot{\lambda}_0 \phi_0^*(g_{uu}^0 v_0 + g_u^0 \lambda) \phi_0 \neq 0. \quad (2.16)$$

But then the smallest eigenvalue of $g_u(u(\rho), \lambda(\rho))$ decreases for $\rho \rightarrow \rho_0$ as $|\rho - \rho_0|$ and (2.13) follows. If $\dot{\lambda}_0 = 0$ then $\dot{u}_0 = c\phi_0$ from (2.14), $c \in \mathbb{R}$. Then $\dot{\beta}(\rho_0) = 0$ follows in (2.15) using in addition (2.4).

3. Symmetry-breaking bifurcation

We assume that problem (1.1) satisfies the following symmetry condition:

$$\begin{aligned} & \text{There exists } S \in L(X) \text{ with } S \neq I, S^2 = I \\ & \text{and } g(Su, \lambda) = Sg(u, \lambda) \text{ for } u \in X, \lambda \in \mathbb{R}. \end{aligned} \quad (3.1)$$

This is a very common condition. This and other symmetries have been considered in [12] and we refer in particular to [5] where it is shown that the presence of symmetries reduces the codimension of singularities. In [13] it was shown that symmetry-breaking pitchfork bifurcation points may be computed as isolated solution of an inflated system. We intend to compute solutions near such a point without necessarily knowing it. We note as in [13] that (3.1) implies the decompositions

$$X = X_s \oplus X_a, \quad X' = (X')_s \oplus (X')_a, \quad (3.2)$$

where

$$X_s = \{x \in X \mid Sx = x\}, \quad X_a = \{x \in X, Sx = -x\}, \quad (3.3a)$$

$$(X')_s = \{\psi \in X' \mid \psi = \psi\delta\}, \quad (X')_a = \{\psi \in X', \psi = -\psi\delta\}. \quad (3.3b)$$

We call the pitchfork bifurcation point (u_0, λ_0) of (1.1) symmetry-breaking, if

$$u_0 \in X_S, \phi_0 \in X_A. \quad (3.4)$$

The solution set of (1.1) near (u_0, λ_0) consists of two smooth, transversally intersecting branches of the form

$$\begin{aligned} u_1(\xi) &= u_0 + w_1(\xi), \quad \lambda_1(\xi) = \lambda_0 + \xi, \quad \xi \in [-\xi_0, \xi_0], \\ u_2(\xi) &= u_0 + \xi \phi_0 + w_2(\xi), \quad \lambda_2(\xi) = \lambda_0 + O(\xi^2), \\ \|w_i(\xi)\| &= O(\xi^i), \quad i = 1, 2, \quad w_1(\xi) \in X_S. \end{aligned} \quad (3.5)$$

In addition to (3.4) we assume

$$\phi_0^* \in (X')_A \quad (3.6)$$

which, however, is always satisfied for finite-dimensional X (cf. [13]). The branch u_1 consists of symmetric solutions, while the symmetry is lost along u_2 .

4. The augmented system

In the following we shall see that the system (1.3) has no better behaviour than stated in Theorem 2.12 for (1.3a). Nevertheless the inverse iteration (1.4) for the solution of (1.3) performs reasonably well in the neighborhood of a bifurcation point (cf. 8).

We rewrite (1.3) as

$$F(x, \rho) = 0, \quad x \in X \times R \quad (4.1)$$

and denote its Fréchet derivative with respect to x in $x_0 = (u_0, \lambda_0)$ by

$$F_x^0 = \begin{bmatrix} g_u^0 & g_\lambda^0 \\ u_0^* & 0 \end{bmatrix}. \quad (4.2)$$

Theorem 4.3 Let $F(x, \rho)$ in (4.1) be continuously differentiable in the symmetry-breaking pitchfork bifurcation point x_0 and let

$$u_0^* v_0 \neq 0, \quad v_0^* u_0 \neq 0, \quad v_0^* \text{ as in (4.8)} \quad (4.4)$$

then F_x^0 is a Fredholm operator of index zero with

$$N(F_x^0) = \text{span}\{\phi_0\}, \quad \phi_0 = \begin{bmatrix} \phi_0 \\ 0 \end{bmatrix} \quad (4.5)$$

$$R(F_x^0) = \{x \in X_R \mid \phi_0^* x = 0\}, \quad \phi_0^* = (\phi_0^*, 0). \quad (4.6)$$

The zero eigenvalue of F_x^0 is of algebraic multiplicity one.

Proof To solve $F_x^0 \phi_0 = 0$, $\phi_0 = \begin{bmatrix} v \\ \alpha \end{bmatrix}$, we must have

$$g_u^0 v + \alpha g_\lambda^0 = 0, \quad u_0^* v = 0. \quad (4.7)$$

The first equation has the solution $v = \alpha v_0 + \beta \phi_0$, $\beta \in R$. We have $u_0^* \in (X')_S^*$ and hence $u_0^* w = u_0^* S^2 w = (u_0^* S)(Sw) = -u_0^* w$ for $w \in X_a$. The second equation then yields $0 = u_0^* v = \alpha u_0^* v_0$ and thus $\alpha = 0$.

There cannot be a vector $\phi_1 = \begin{bmatrix} v \\ \alpha \end{bmatrix}$ with $F_x^0 \phi_1 = \phi_0$ because then

$$g_u^0 v + \alpha g_\lambda^0 = \phi_0$$

but $\phi_0 \in X_a$ and the left-hand side is in X_S . To solve

$$\phi_0^* F_x^0 = 0, \quad \phi_0^* = (v^*, \alpha)$$

we get

$$g_u^{0*} v^* + \alpha u_0^* = 0, \quad v^* g_\lambda^0 = 0.$$

The first equation has the solution $v = \alpha v_0^* + \phi_0^*$, where v_0^* is the unique solution of

$$g_u^{0*} v_0^* + u_0^* = 0, \quad v_0^* \phi_0 = 0. \quad (4.8)$$

The second equation then yields $\alpha v_0^* u_0 = 0$ and hence $\alpha = 0$.

The analogue of Theorem 2.12 holds. For the sake of brevity we only note that $\|F_x^{-1}(x(\rho), \rho)\|$ grows at least as fast as $\|g_u^{-1}(u(\rho), \lambda(\rho))\|$ in (2.13) and that this result also follows essentially from (2.4) and (2.5).

5. The inverse iteration

In computing the branches intersecting in a symmetry-breaking pitchfork bifurcation point the problem is not to determine the symmetric (primary) branch. This branch may be obtained by solving

$$g(u, \lambda) = 0, \quad g: X_S \times R \rightarrow X_S. \quad (5.1)$$

This problem does not have the non-symmetric branch anymore and the question arises if analogously a problem may be found that does not have the symmetric branch.

We consider the augmented problem

$$g(u(\rho), \lambda(\rho)) = \lambda u - f(u) = 0, \quad (5.2a)$$

$$N(u(\rho), \rho) = \|u_a(\rho)\| - \rho = 0, \quad (5.2b)$$

where $u = u_s + u_a$, $u_s \in X_s$, $u_a \in X_a$.

Obviously for $\rho > 0$ in a neighbourhood of (u_0, λ_0) (5.2) has only the non-symmetric (secondary) branch.

In generalization of (1.4) we propose the following algorithm for the numerical solution of (5.2).

For a given $\rho > 0$, $u^{(0)} \in X$, $\|u_a^{(0)}\| = \rho$, compute $(u^{(k)}, \lambda^{(k)})$ according to

$$u^{(k)} = \tilde{u}_s^{(k-1)} + \rho \frac{\tilde{u}_a^{(k-1)}}{\|\tilde{u}_a^{(k-1)}\|}, \quad \lambda^{(k)} = u^{(k)*} f(u^{(k)}) / \|u^{(k)}\|$$

$$\tilde{u}^{(k-1)} = u^{(k-1)} + \delta u^{(k-1)}, \quad k = 1, 2, \dots, \text{ where} \quad (5.3)$$

$$\begin{bmatrix} g_u^{(k-1)} & g_\lambda^{(k-1)} \\ u_a^{(k-1)*} & 0 \end{bmatrix} \begin{bmatrix} \delta u^{(k-1)} \\ * \end{bmatrix} = - \begin{bmatrix} g^{(k-1)} \\ 0 \end{bmatrix}.$$

Here $u_a^{(k-1)*} \in (X')_a$ is such that $u_a^{(k-1)*} u_a^{(k-1)} = \|u_a^{(k-1)}\|$. From (3.5) we have $u_a^{(k-1)}/\rho \rightarrow \phi_0$ for $\rho \rightarrow 0$ along the non-symmetric branch, hence $u_a^{(k-1)*} \rightarrow \phi_0^*$

If we rewrite (5.2) as

$$F(x(\rho), \rho) = 0, \quad x \in X \times R, \quad (5.4)$$

then the limit along the non-symmetric branch of the Fréchet derivative with respect to x in x_0 is now in contrast to (4.2)

$$F_x^0 = \begin{bmatrix} g_u^0 & g_\lambda^0 \\ * & 0 \end{bmatrix}. \quad (5.5)$$

As in Theorem 4.3 it may be shown that F_x^0 is Fredholm of index zero but that the zero eigenvalue is of algebraic multiplicity two. Using Theorem 1 of [4] the following result may be shown. We note that it represents an improvement over the corresponding behaviour of F_x^{-1} mentioned at the end of 4. along the branch of non-symmetric solutions.

Theorem 5.6 Let X be a Hilbert space and let $F(x, \rho)$ in (5.4) be twice continuously differentiable with respect to x and ρ . Let $x(\rho)$ be a twice continuously differentiable arc of solutions through the symmetry-breaking bifurcation point x_0 . Then for some $\delta > 0$, $K > 0$ and each ρ in $0 < \rho < \delta$

$$\|F_x^{-1}(x(\rho), \rho)\| \leq \frac{K}{\rho}. \quad (5.7)$$

We finally note that in the case of a Hilbert space X quadratic convergence of the algorithm (5.3) to any solution u_0 for which the matrix in (5.5) is regular may be proven as in [6] by writing (5.3) in the form

$$u^{(k)} = \Phi(u^{(k-1)})$$

and showing that the Jacobian satisfies

$$\Phi'(u_0)P_{u_0} = 0,$$

where P_{u_0} is the orthogonal projector on $(\text{span } \{u_0\})^\perp$.

6. An example

The following example of a symmetry-breaking bifurcation point was already given in [11].

$$\lambda u(s) - \int_0^\pi k(s,t)h(u(t))dt = 0, \quad 0 \leq s \leq \pi, \quad (6.1)$$

where $k(s,t) = 2(3\sin s \sin t + 2 \sin 2s \sin 2t)/\pi$ and $h(u) = u + u^3$.

We rewrite (6.1) as

$$g(u, \lambda) = \lambda u - Kh(u) = 0. \quad (6.2)$$

where $g: \mathbb{R} \times X \rightarrow X$ with $X = C[0, \pi]$,

endowed with the sup-norm. Defining $S \in L(X)$ by

$$Su(s) = u(\pi-s)$$

we see that (3.1) is satisfied for g in (6.2).

The above problem has the trivial solution $u_0 = 0$ and the branches of symmetric ($Su = u$) and anti-symmetric ($Su = -u$) solutions

$$u_i(\lambda; s) = \pm \frac{2}{\sqrt{3}} \sqrt{\frac{\lambda}{4-i} - 1} \sin is, \quad i = 1, 2,$$

bifurcating from the trivial solution at $\lambda = 3$ and $\lambda = 2$, respectively. At $\lambda = 6$ the branches

$$\begin{aligned} u_3^+(\lambda; s) &= \frac{2}{3} \sqrt{\frac{2}{3}\lambda - 1} \sin s \pm \frac{2}{3} \sqrt{\frac{\lambda}{6} - 1} \sin 2s \\ u_3^-(\lambda; s) &= -\frac{2}{3} \sqrt{\frac{2}{3}\lambda - 1} \sin s \pm \frac{2}{3} \sqrt{\frac{\lambda}{6} - 1} \sin 2s \end{aligned}$$

of non-symmetric solutions bifurcate from $u_1(6; s) = \frac{2}{\sqrt{3}} \sin s$.

We consider a standard discretization of (6.1). The integral is replaced by the composite trapezoidal rule and collocation is used in the nodes of an equidistant grid to obtain

$$\underline{g}(\lambda, \underline{u}) = \lambda \underline{u} - \underline{M} \underline{h}(\underline{u}) \quad (6.3)$$

where $\underline{g}: \mathbb{R} \times \mathbb{R}^N \rightarrow \mathbb{R}^N$, $\underline{M}_{ij} = w_j k(t_i, t_j)$, $t_i = (i-1)h$, $h = \pi/(N-1)$, $w_1 = w_N = h/2$, $w_i = h$, $i = 2, \dots, N-1$. The linear operator in the symmetry-relation is chosen as

$$S = \begin{bmatrix} 0 & & & 1 \\ & \ddots & & \\ 1 & & 0 \end{bmatrix}.$$

7. Numerical Implementation

The generalized inverse iteration (5.3) will now be applied to the discretization (6.3) of problem (6.1). This serves, of course, as a simple example for the application to finite-dimensional problems with symmetry-breaking bifurcation points in general.

We decompose $X = \mathbb{R}^N$ as

$$\mathbb{R}^N = \mathbb{R}_s^N \times \mathbb{R}_a^N, \quad \dim \mathbb{R}_s^N = N_s, \quad \dim \mathbb{R}_a^N = N_a. \quad (7.1)$$

If bases of \mathbb{R}_s^N and \mathbb{R}_a^N are known these spaces may be identified with \mathbb{R}^{N_s} and \mathbb{R}^{N_a} . We assume that N in (6.3) is odd and thus $N_s = (N+1)/2$, $N_a = N - N_s$. We have

$$\begin{aligned} X_s &= \text{span } \{e_i + e_{N+1-i}, i = 1, \dots, N_s\} , \\ X_a &= \text{span } \{e_i - e_{N+1-i}, i = 1, \dots, N_a\} . \end{aligned}$$

Any $\underline{u} \in X$ may be decomposed into its symmetric and anti-symmetric parts
 $\underline{u} = (\underline{u}_s, \underline{u}_a)^T$ where $\underline{u}_s = I_s \underline{u}$, $\underline{u}_a = I_a \underline{u}$ and $I_s \in \text{Isom } (R_s^N, R_s^N)$,
 $I_a \in \text{Isom } (R_a^N, R_a^N)$.

We may take

$$I_s = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & & & & 0 & & & \\ & \bullet & & & & & & \\ & & \bullet & & & & & \\ & & & 1 & & & & \\ 0 & & & & \sqrt{2} & & & \\ & & & & & 1 & & \\ & & & & & & 0 & \\ & & & & & & & 0 \end{bmatrix} , \quad I_a = \frac{\sqrt{2}}{2} \begin{bmatrix} 1 & & & & 0 & & & \\ & \bullet & & & & & & \\ & & \bullet & & & & & \\ & & & \bullet & & & & \\ 0 & & & & 1 & & & \\ & & & & & 0 & & \\ & & & & & & -1 & \\ & & & & & & & 0 \end{bmatrix} .$$

Algorithm (5.3) may now be implemented as

$$\begin{aligned} \underline{u}^{(k)} &= \begin{bmatrix} \underline{u}_s^{(k-1)} + \frac{\delta \underline{u}_s^{(k-1)}}{\rho} \\ \rho \underline{u}_a^{(k-1)} + \frac{\delta \underline{u}_a^{(k-1)}}{\|\underline{u}_a^{(k-1)} + \delta \underline{u}_a^{(k-1)}\|} \end{bmatrix} , \quad \lambda^{(k)} = \frac{\underline{u}^{(k)T} M_h(\underline{u}^{(k)})}{\underline{u}^{(k)T} \underline{u}^{(k)}} , \\ \begin{bmatrix} I_{sa} g_u^{(k-1)T} I_{sa} & I_{sa} g_\lambda^{(k-1)} \\ \frac{1}{\rho} (0, \underline{u}^{(k-1)T} I_a^T) & 0 \end{bmatrix} \begin{bmatrix} \delta \underline{u}_s^{(k-1)} \\ \delta \underline{u}_a^{(k-1)} \\ * \end{bmatrix} &= \begin{bmatrix} -I_{sa} g^{(k-1)} \\ 0 \end{bmatrix} , \end{aligned} \tag{7.2}$$

where $I_{sa} = (I_s^T, I_a^T)^T$.

Algorithm (7.2) will be used in the following section to compute the solutions u_3 of problem (6.1). It is obvious that by working in R_s^N respectively R_a^N the solutions u_1 and u_2 may be obtained.

8. Results

In 5. we have seen that the Jacobian in the symmetry-breaking bifurcation point for continuation along the branch of non-symmetric solutions is the same for algorithm (5.3) as for pseudo-arc length continua-

tion. We apply this result now to algorithm (7.2) respectively the pseudo-arclength method for (6.3). Another method which has been proposed for the solution of problems of the type (6.3) is continuation along certain components of the solution.

In the following we compare the generalized inverse iteration with the above mentioned algorithms by applying them to (6.3) with $N = 5$. No predictor-step is used for continuation along a component while the anti-symmetric part of the solution is just renormalized for the methods (1.4) and (7.2). The trapezoidal rule is of arbitrary order for the periodic functions to be integrated, so the obtained value will be nearly exact independent of N . The behaviour of the algorithms did also not strongly depend on N .

The following table contains the number of steps needed for algorithms (1.4), (7.2) and continuation along the components \underline{u}_3 and \underline{u}_4 . The starting solution $\underline{u}^{(0)}$ was that on the secondary branch \underline{u}_3^+ with $\rho = \|\underline{u}_3^{(0)}\| = 1$. Then this branch was followed towards the bifurcation point by continuing to $\rho = 1/2, 1/4, \dots$

ρ	λ	\underline{u}_3 -cont.	\underline{u}_4 -cont.	(1.4)	(7.2)
1/2	7.5	6	4	4	3
1/4	6.375	5	4	5	3
1/8	6.09375	4	4	5	2
1/16	6.02344	4	4	4	2
1/32	6.00587	3	4	4	1
1/64	6.00147	3	4	3	1
1/128	6.00037	3	3	2	1
1/256	6.00009	3	3	1	0
1/512	6.00002	(3)	3	0	0

Table 8.1 Continuation with different algorithms, starting at $\rho = 1$, $\lambda = 12$.

A number in parentheses indicates convergence in the given number of steps but to a different branch, while zero steps were necessary if the renormalization of the previous solution already satisfied the stopping criterion which was taken as $\|g(\lambda^{(k)}, \underline{u}^{(k)})\| < 10^{-6}$ for all methods.

We see that away from the bifurcation point the algorithm (1.4) behaves very similar as continuation along one of the components. In the neighbour-

hood of the bifurcation point, however, generalized inverse iteration has advantages and in the special form (7.2) it is superior also away from the singularity. We point out again that for a given value of the continuation parameter only this latter method does not have the solutions on the primary branch. This suggests that it should be possible as for (1.4) applied to problems without secondary bifurcation (cf. [8]) to use considerably larger continuation steps.

The next table shows what happens for continuation from the solutions corresponding to $\rho = \rho_s$ to ρ_t .

ρ_s	ρ_t	u_3 -cont.	u_4 -cont.	(1.4)	(7.2)
1	1/4	(8)	5	6	4
1/4	1/16	6	5	6	2
1/16	1/64	6	5	5	1
1/64	1/256	(3)	4	2	1
1	1/16	(11)	7	8	3
1/16	1/256	(3)	6	5	1
1	1/256	(16)	10	11	3

Table 8.2 Continuation with larger stepsizes.

We see that u_4 - continuation is still comparable to (1.4) but that (7.2) is so robust that it is tempting to use extremely large steps.

ρ_s	ρ_t	u_4 -cont.	(7.2)
2	1/256	8	4
1/156	2	(11)	5
4	1/256	10	5
1/256	4	(15)	(3)

Table 8.3 Continuation with very large steps.

For $\rho = 2$ (4) λ has the value 30 (102). While the number of iterations needed for the continuation steps in Tables 8.1 and 8.2 in the opposite direction, i.e. away from the bifurcation point instead of towards it, is essentially the same, we see in Table 8.3 that algorithm (7.2) finally breaks

down if an extremely large step from the bifurcation point is taken.

We present now a few typical results obtained with the pseudo-arclength method [6] with predictor step and weight-factor θ , $0 \leq \theta \leq 2$. Steps of different size δs are taken from three different points on the non-symmetric branch.

θ	λ_0	δs	λ	pseudo-arclength	(7.2)
1 1.5	12	4.55	7.5121	6	3
		1.4	6.9338	6	3
1.5 1.5	6.375	.125	6.0937	4	2
		.165	6.0239	6	2
1.3 1.5 1.7	6.02344	.073	6.000166	4	1
		.08	6.0003	5	1
		.09	6.000589	6	1

Table 8.4 Continuation with the pseudo-arclength method.

A greater θ allowed larger steps but only away from the bifurcation point. Numerical instabilities could be observed in its neighbourhood.

We conclude with the remarks that generalized inverse iteration in the form of (5.3) allows to compute solutions on the branch bifurcating at a symmetry-breaking bifurcation point in a very robust and efficient way. It is not necessary to know the bifurcation point. Since the points on the symmetric branch are not solutions of (5.2) any starting guess composed of a symmetric part which is not too different from the symmetric solution near the bifurcation point and a suitable nonzero anti-symmetric part will lead to convergence. If however, the bifurcation point is known then appropriate modifications of the basic algorithm (1.4) should have much better properties than that also in the case that the bifurcation is not symmetry-breaking.

Although the asymptotic behaviour of the systems used in (5.3) and in the pseudo-arclength method is identical the algorithms perform quite differently and a deeper analysis also including the predictor step is needed to explain this. In particular we note that for the predicted u the λ given by the Rayleigh-quotient minimizes the residual $\|g(u, \lambda)\|$ w.r.t. λ . In the general case, however, (8.1d) below should be replaced by

$$R(u, \lambda) = g_\lambda^T(u, \lambda)g(u, \lambda).$$

We finally compare the continuation method used in [14] with the generalized inverse iteration (1.4). In [14] the equation (1.1) is augmented by

$$N(u, \lambda) = \theta \dot{\rho}_0 (\rho - \rho_0) + (2-\theta) \dot{\lambda}_0 (\lambda - \lambda_0) - \delta s = 0 , \quad (8.1a)$$

where $\rho = \text{null}$ and $\theta \in (0,2)$. The predictor step is

$$u(\alpha) = u_0 + \alpha \dot{u}_0, \quad \lambda(\beta) = \lambda_0 + \beta \dot{\lambda}_0 , \quad (8.1b)$$

where α, β satisfy

$$N(u(\alpha), \lambda(\beta)) = 0, \quad R(u(\alpha)), \lambda(\beta)) = 0 \quad (8.1c)$$

and here

$$R(u, \lambda) = u^T g(u, \lambda) \quad (8.1d)$$

denotes the generalized Rayleigh-quotient. The augmented system is then solved by Newton's method. This continuation strategy is in the predictor as well as in the corrector step closely related to (1.4).

θ	λ_0	δs	λ	(8.1)	(1.4)
1.5	12	2.972	6.3787	5	6
		3.163	6.0239	7	8
		3.176	6.00005	9	10
1.5	6.375	.33	6.09134	4	6
		.45	6.00099	9	7
1.5 2	6.0234 8	7.9 E-3 8 E-3	6.00007 6.00002	8 5	5 5

Table 8.5 Comparison of (8.1) and (1.4)

We see from the results that the method is in fact comparable to (1.4) and hence allows larger steps than the pseudo-arc length method. The number of iterations did not strongly depend on θ . Near the bifurcation point, where $\dot{\lambda}_0 = 0$ along the branch of nonsymmetric solutions, it was necessary to choose $\theta = 2$ to obtain the same number of iterations as for (1.4). Since there also $\dot{\rho}_0 = 0$ holds it is not possible to continue with this method from the bifurcation point.

References

- [1] BREZZI, F., RAPPAZ, J. and RAVIART, P.A., Finite dimensional approximations of nonlinear problems. Part III. Simple bifurcation points, *Numer. Math.* 38, 1-30 (1981).
- [2] DECKER, D.W. and KELLER, H.B., Multiple limit point bifurcation, *J. Math. Anal. Appl.* 75, 417-430 (1980).
- [3] DECKER, D.W. and KELLER, H.B., Path following near bifurcation, *Comm. Pure Appl. Math.* 34, 149-175 (1981).
- [4] DESCLOUX, J., Two remarks on continuation procedures for solving some non-linear problems, manuscript (1983).
- [5] GOLUBITSKY, M. and SCHAEFFER, D., Imperfect bifurcation in the presence of symmetry, *Comm. Math. Phys.* 67, 205-232 (1979).
- [6] KELLER, H.B., Numerical solution of bifurcation and nonlinear eigenvalue problems, in "Applications of bifurcation theory", ed. P.H. Rabinowitz, Academic Press, New York, 1977.
- [7] MITTELMANN, H.D., An efficient algorithm for bifurcation problems of variational inequalities, *Math. Comp.* (to appear).
- [8] MITTELMANN, H.D., A fast solver for nonlinear eigenvalue problems, in "Iterative solution of nonlinear systems", R. Ansorge, T. Meis and W. Törnig (eds.), Springer Lecture Notes in Mathematics, vol. 953, 1982.
- [9] MITTELMANN, H.D., Multi-grid methods for simple bifurcation problems, in "Multi-grid methods", W. Hackbusch and U. Trottenberg (eds.), Springer Lecture Notes in Mathematics, vol. 960, 1982.
- [10] MITTELMANN, H.D. and WEBER, H., Multi-grid solution of bifurcation problems, SIAM J. Sci. Stat. Comp. (to appear).
- [11] PIMBLEY, G.H., Eigenfunction branches of nonlinear operators and their bifurcation, Springer Lecture Notes in Mathematics, vol. 104, 1969.
- [12] SATTINGER, P.H., Group theoretic methods in bifurcation theory, Springer Lecture Notes in Mathematics, vol. 762, 1979.
- [13] WERNER, B. and SPENCE, A., The computation of symmetry-breaking bifurcation points, preprint 83/5, Universität Hamburg (1983).
- [14] BANK, R.E. and CHAN, T.F., PLTMGC: A multi-grid continuation program for solving parametrized nonlinear elliptic systems. Report #261, Yale Computer Science Department (1983).

H.D. Mittelmann

Department of Mathematics
 Arizona State University
 Tempe, AZ 85287, USA

and

Abt. Mathematik
 Universität Dortmund
 4600 Dortmund, F.R.G.

The Numerical Buckling of a Visco-elastic Rod

Gerald Moore

Introduction

We consider the deformation of a thin, inextensible rod under axial thrust P . The rod has pinned ends and is assumed to deform in a plane with one end free to move horizontally. It is of unit length and if $x \in [0,1]$ denotes the material points in the initial undeformed configuration then $u(x)$ denotes the angle between the tangent to the rod at x and the horizontal.

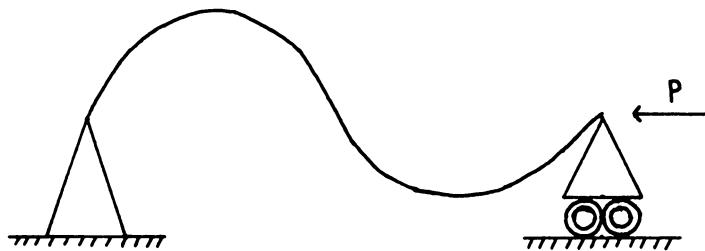


FIG. 1

We work within the quasi-static theory so that, if $m(x)$ denotes the bending moment, the balance of moments is given by

$$(1) \quad m'(x) + P \sin u(x) = 0$$

and the pinned end conditions imply

$$(2) \quad m'(0) = m'(1) = 0.$$

If the rod is assumed to be linearly elastic, the bending moment and curvature are related by

$$(3) \quad m(x) = \beta u'(x), \quad \beta > 0$$

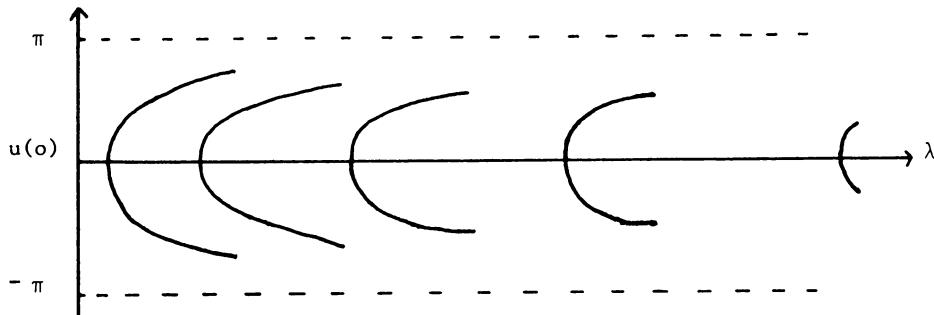
and combining (1), (2) and (3) gives

$$(4) \quad \begin{aligned} u''(x) + \lambda \sin u(x) &= 0 \\ u'(0) = u'(1) &= 0 \end{aligned}$$

where λ is proportional to the thrust. This equation was solved in terms of elliptic functions by Euler [6] and is one of the most commonly used examples of bifurcation from the trivial solution $u(x) \equiv 0$ [3,9,10]. It is clear that $u(x) \equiv 0$ is a solution of (4) for all λ and the linearisation about this solution,

$$(5) \quad \begin{aligned} v''(x) + \lambda v(x) &= 0 \\ v'(0) = v'(1) &= 0, \end{aligned}$$

is singular for $\lambda_n = (n\pi)^2$, $n = 0, 1, \dots$. Standard results in bifurcation theory [1] show that $(\lambda_n, 0)$, $n \geq 1$, is a bifurcation point of (4) and the global solution graph is displayed in Fig. 2.



The two most notable features of the above solutions are that:-

- (i) there is no secondary bifurcation:
- (ii) the branch from $(\lambda_n, 0)$ maintains the same nodal structure and symmetry as the eigenfunction corresponding to the zero eigenvalue of (5) at λ_n , $\cos n\pi x$; thus $u(x)$ has n simple zeros in $(0, 1)$.

If we consider the shape of the rod on each branch, the horizontal and vertical co-ordinates being given by

$$(6) \quad X(x) = \int_0^x \cos u(x) dx$$

$$Y(x) = \int_0^x \sin u(x) dx$$

respectively, then symmetry and nodal-structure is inherited from u , i.e. on the n th. primary branch $Y(x)$ has $n-1$ simple zeros. In Fig. 3 below we show how the rod changes with increasing λ for $n = 1$ and $2[6,9]$.

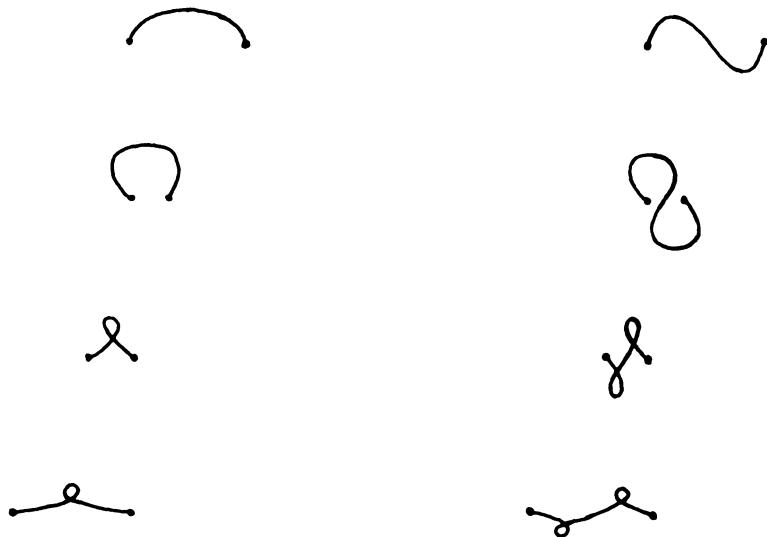


FIG. 3

In this paper, however, we wish to consider the deformation of a linearly visco-elastic rod, a problem on which much less work has been done. In this case the constitutive relation replacing (3) is

$$(7) \quad m(x,t) = g(0)u'(x,t) + \int_0^t g(t-s)u'(x,s)ds.$$

Now the bending moment and curvature are explicit functions of time and the former depends on the history of the latter. The "relaxation function" $g(\cdot)$ is the stress required to maintain constant strain and physically it is required that

$$(8) \quad g(\cdot) \geq 0, \quad \dot{g}(\cdot) \leq 0, \quad \ddot{g}(\cdot) \geq 0.$$

Thus $g(\cdot)$ would typically behave like a linear combination of negative exponentials with positive coefficients.

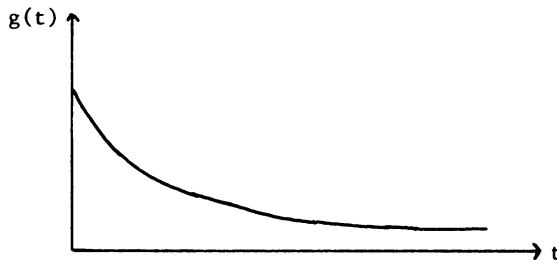


FIG. 4

It should be noted that we are dealing with a visco-elastic solid which is "partially relaxing", $\lim_{t \rightarrow \infty} g(t) > 0$, in contrast to a "completely relaxing" visco-elastic liquid, $\lim_{t \rightarrow \infty} g(t) = 0$.

Combining (1) and (2) with (7), and dividing through by $g(0)$, gives the fundamental equation for the deformation of a visco-elastic rod,

$$(9) \quad u''(x,t) + \int_0^t K(t-s)u''(x,s)ds + \lambda(t)\sin u(x,t) = 0$$

$$u'(0,t) = u'(1,t) = 0 \quad \blacktriangledown t.$$

Thus we have a second kind nonlinear Volterra partial integro-differential equation with convolution kernel. This leads to somewhat unusual bifurcation problems in which time is the independent parameter and $\lambda(\cdot)$ a given function of t , proportional to the axial load. For any $\lambda(\cdot)$, $u(\cdot) \equiv 0$ is a possible solution and the uniqueness theorem for Volterra equations shows that it is only possible for a non-trivial solution to branch away at values of t for which $\lambda(t) = (n\pi)^2$ $n = 1, 2, \dots$. This is considered in the next section.

In [4] a number of open problems are posed about the behaviour of the visco-elastic rod. Some of these will be discussed in [8] but in this paper we shall merely present some numerical results about the existence

and form of secondary bifurcation for various choices of $g(\cdot)$ and $p(\cdot)$.

2. Bifurcation from the Trivial Solution

In this section we briefly consider the condition under which bifurcation from the trivial solution may take place and the resulting form of the non-zero branch. A more detailed analysis will be given in [8] and we note also the independent work on buckling at $\lambda(t) = \pi^2$ contained in [7].

We consider equation (9) in its weak form and so for each t $u(\cdot, t) \in H^1(0, 1)$ and u describes a continuous mapping $R \rightarrow H^1(0, 1)$. Since the trivial solution does not contribute to the integral term, t may be renormalised so that $\lambda(0) = (n\pi)^2$. Letting Q be the orthogonal L^2 projection onto $\{\cos n\pi x\}$ we may split $u(x, t)$, in standard Liapunov-Schmidt fashion, into

$$(10) \quad u(x, t) = \varepsilon(t) \cos n\pi x + w(x, t) \\ \int_0^1 w(x, t) \cos n\pi x \, dx = 0 \quad \forall t$$

and decompose (9) into

$$(11) \quad \begin{aligned} a) \quad & w''(x, t) + \int_0^t K(t-s) w''(x, s) ds + \lambda(t) Q \sin u(x, t) = 0 \\ & w'(0, t) = w'(1, t) = 0 \quad \forall t, \quad w(x, 0) = 0 \\ b) \quad & -(n\pi)^2 (1 + \int_0^t K(t-s) \varepsilon(s) ds) + \lambda(t) \int_0^1 \sin u(x, t) \cos n\pi x \, dx = 0 \\ & \varepsilon(0) = 0 \end{aligned}$$

Now (11a) is a regular equation for $w(x, t)$, with $\varepsilon(t)$ as a parameter, whose solution satisfies

$$(12) \quad \|w(\cdot, t)\|_{H^1} \leq K_t \|\varepsilon\|^3 \quad \|\varepsilon\|_t = \sup_{s \in [0, t]} |\varepsilon(s)|.$$

Inserting this result into (11b), and assuming $\lambda(t)$ and $K(t)$ are sufficiently smooth, gives

$$(13) \quad \frac{-(n\pi)^2 K(0)}{2} \int_0^t \varepsilon(s) ds + \frac{\dot{\lambda}(0)t\varepsilon(t)}{2} - \frac{(n\pi)^2 \varepsilon^3(t)}{16} + B(t, \varepsilon) = 0$$

where

$$(14) \quad |B(t, \varepsilon)| = O(t^2 \|\varepsilon\|_t + t \|\varepsilon\|^3 + \|\varepsilon\|^5).$$

This is a somewhat usual bifurcation equation to deal with, but using the Newton polygon method [11] it can be shown that, if

$$(15) \quad \dot{\lambda}(0) > -\frac{2}{3} K(0)(n\pi)^2 ,$$

a solution of (13) may be developed in terms of $t^{\frac{1}{2}}$ and

$$(16) \quad \varepsilon(t) = At^{\frac{1}{2}} + O(t)$$

with $A = \frac{4}{n\pi} \frac{(3\lambda'(0) - 2(n\pi)^2 K(0))^{\frac{1}{2}}}{6}$. Inserting this solution back into (12)

and (10) finally gives

$$(17) \quad u(x,t) = At^{\frac{1}{2}} \cos n\pi x + O(t)$$

as a bifurcating solution of (9) from $\lambda(0) = (n\pi)^2$, these solutions correspond to those for the elastic problem (4) where

$$(18) \quad u(x,\lambda) = \frac{2(2(\lambda - (n\pi)^2))^{\frac{1}{2}}}{(n\pi)^2} \cos(n\pi x) + O(\lambda - (n\pi)^2).$$

It is noteworthy that, in the visco-elastic case, $\lambda(t)$, which is proportional to the axial thrust, may decrease (but not too much) as it leaves $(n\pi)^2$.

3. Numerical Formulation

In this section we describe the numerical methods applied to obtain the results given in 4. A discrete Galerkin method is used to reduce (9) to a system of nonlinear equations and thus $u(x,t)$ is approximated spatially by

$$(19) \quad u_h(x,t) \equiv \sum_j \alpha_j(t) \phi_j(x),$$

where the ϕ_j are suitable piecewise polynomials in $H^1(0,1)$. The continuous Galerkin approximation of (9) is therefore

$$(20) \quad A\dot{q}(t) + \int_0^t K(t-s)Aq(s)ds - \lambda(t)f(q(t)) = 0,$$

where

$$(21) \quad A_{ij} = \int_0^1 \phi_i(x) \phi_j'(x) dx \quad \text{and}$$

$$f_i(q(t)) = \int_0^1 \sin(u_h(x,t)) \phi_i(x).$$

This system of coupled Volterra equations can be discretized by standard integration methods. For example, if the trapezoidal rule is used, (20) becomes

$$(22) \quad \begin{aligned} A\tilde{\alpha}_{n+1} - \lambda(t_{n+1})f(\tilde{\alpha}_{n+1}) + \frac{\delta t_{n+1}}{2} \{ K(0)A\tilde{\alpha}_{n+1} + K(\delta t_{n+1})A\tilde{\alpha}_n \} \\ + \sum_{p=1}^n \delta t_p \{ K(t_{n+1-s_p})A\tilde{\alpha}_p + K(t_{n+1-s_{p-1}})A\tilde{\alpha}_{p+1} \} = 0, \\ t_{n+1} = t_n + \delta t_{n+1}. \end{aligned}$$

Thus we arrive at a system of nonlinear equations

$$(23) \quad F(\delta t_{n+1}, \tilde{\alpha}_{n+1}) = 0,$$

whose solution branches may be followed by known methods [5], to provide approximations to the bifurcating solutions of (9). Since $u(x,t) = 0(t^{\frac{1}{2}})$, it is also necessary, in order to maintain accuracy, to take account of this singularity.

If a more accurate quadrature rule is required we prefer to use block methods rather than the complexities of Runge-Kutta methods or the step-changing difficulties of linear-multistep methods [2].

In applying these quadrature rules it is possible, if $K(\cdot)$ is a linear combination of exponentials, to maintain running totals of the integrals (or rather their approximations.) Thus if

$$(24) \quad K(t) = \sum_{j=1}^m c_j e^{-b_j t},$$

we have

$$(25) \quad \begin{aligned} \int_0^{t_{n+1}} K(t_{n+1}-s) g(s) ds &= \sum_{j=1}^m c_j e^{-b_j (t_{n+1}-t_n)} \int_0^{t_n} e^{-b_j (t_n-s)} \tilde{g}(s) ds \\ &\quad + \sum_{j=1}^m c_j \int_{t_n}^{t_{n+1}} e^{-b_j (t_{n+1}-s)} \tilde{g}(s) ds, \end{aligned}$$

and only approximations to

$$(26) \quad \int_0^{t_n} e^{-b_j (t_n-s)} g(s) ds \quad j = 1, \dots, m$$

need be stored.

4. Numerical Results

In this section we present a few numerical results illustrating the occurrence of secondary bifurcation for the visco-elastic problem (in contrast to the elastica). More detailed results will be given in [8].

λ_n	1	2	3	4	5
$\lambda_n - t$	-	4.0	1.6	1.1	0.8
λ_n	-	3.5	1.6	1.1	0.8
$\lambda_n + t$	7.0	3.2	1.6	1.1	0.8
$\lambda_n + t^2$	3.5	2.9	1.6	1.0	0.8

The above table shows, for various axial thrusts $P(t)$, the time, after primary buckling, at which secondary bifurcation occurred. In each case the relaxation function is $g(t) = 1 + e^{-t}$.

The form of the solution is similar in each case. The secondary buckling is of pitch-fork type in which the symmetry of the primary branch is lost. In fact an $(n-1)$ -node solution is superimposed on the primary n -node branch.

REFERENCES

1. Crandall M.G. and Rabinowitz P.H.: Bifurcation from Simple Eigenvalues. J.Func. Anal. 8(1971), 321-340.
2. Delves L.M. and Walsh J.: Numerical Solution of Integral Equations. Oxford University Press, 1974.
3. Dickey R.W.: Bifurcation Problems in Nonlinear Elasticity. Pitman. London 1976.
4. Gurtin M.E.: Some Questions and Open Problems in Continuum Mechanics and Population Dynamics. J.Diff.Equ. (to appear).
5. Keller H.B.: Numerical Solution of Bifurcation and Non-linear Eigenvalue Problems. In Applications of Bifurcation Theory, Rabinowitz P.H. (ed.), Academic Press, New York, 1977.
6. Love A.E.H.: A Treatise on the Mathematical Theory of Elasticity. Dover. New York. 1944.
7. Mignot F. and Puel J.P.: Buckling of a Viscoelastic Rod. Arch. Rat. Mech. Anal. (to appear).
8. Moore G. and Reynolds D.W.: The Buckling of a Viscoelastic Rod. (In preparation).
9. Reiss E.L.: Column Buckling - An Elementary Example of Bifurcation. In Bifurcation Theory and Nonlinear Eigenvalue Problems, Keller J.B. and Antman S. (ed), Benjamin, New York, 1969.
10. Stakgold I.: Branching of Solutions of Nonlinear Equations. SIAM Rev. 13(1971), 289-332.
11. Vainberg M.M. and Trenogin V.A.: Theory of Branching of Solutions of Nonlinear Equations. Noordhoff. Leiden. 1974.
Gerald Moore
School of Mathematical Sciences,
National Institute for Higher Education, Dublin 9, Eire.

ASYMPTOTIC ERROR EXPANSION FOR FINITE DIFFERENCE
SCHEMES FOR ELLIPTIC SYSTEMS NEAR TURNING POINTS

Harry Munz

In this note we consider the convergence near turning points of a class of finite difference schemes for a one-parameter family of semilinear elliptic systems

$$\begin{aligned} -\Delta u_j(\lambda, x) &= f_j(\lambda, x, u_1(\lambda, x), \dots, u_m(\lambda, x)) \\ 1 \leq j \leq m, \quad x \in \Omega \\ u_j(\lambda, x) &= 0 \quad 1 \leq j \leq m, \quad x \in \Gamma \end{aligned} \tag{1}$$

where $\lambda \in \mathbb{R}$, Ω is a bounded region in \mathbb{R}^n with boundary $\Gamma \in C^{2,\alpha}$, $0 < \alpha < 1$, and $f \in C^{1,0}(\mathbb{R} \times \bar{\Omega} \times \mathbb{R}^m, \mathbb{R}^m)$.

It is shown that a curve $(\lambda(s), u(s))$, $s \in [s_1, s_2]$, of solutions of (1) that only contains regular and turning points is uniformly approximated by a solution curve of the discrete problem (Theorem 1). Furthermore it is proved, that simple turning points are inherited by the discrete problem (Theorem 2). In both cases the existence of asymptotic error expansions with respect to the L_∞ -norm is asserted.

The Difference Schemes

Let \mathbb{R}_h^n denote a uniform mesh of meshsize $h > 0$ on \mathbb{R}^n .
 $\Omega_h := \Omega \cap \mathbb{R}_h^n$ and $\Gamma_h := \{x \in \Gamma : \exists y \in \Omega_h, \exists \zeta \in \mathbb{R}, \exists i \in \{1, \dots, n\} \text{ s.t. } x = y + \zeta e_i\}$, where e_i is the i -th coordinate vector.

For the FD-approximation of (1) on Ω_h we replace the Laplacian by the standard $(2n+1)$ -point-difference-star.

Let $\mathring{\Omega}_h$ be the set of regular meshpoints, i. e. of those points $x \in \Omega_h$, which have all the neighbours $x \pm he_i$, $1 \leq i \leq n$, in Ω and $\Gamma_h^x := \Gamma_h \setminus \mathring{\Omega}_h$. For $x \in \Gamma_h^x$ one needs auxiliary functional

values in those grid points, which occur in the FD-star, but which lie outside $\bar{\Omega}_h$. These are obtained by one-dimensional polynomial extrapolation of degree k of the functional values in the $k+1$ nearest points in $\bar{\Omega}_h := \Omega_h \cup \Gamma_h$, which are on the corresponding grid line section. The extrapolation requires, that there are sufficiently many points from $\bar{\Omega}_h$ on this section. Generally, for Ω given this is a condition on the choice of h and of the 'origin' of \mathbb{R}_h^n .

Difference schemes using this type of boundary approximation are considered in [1], [4], [5], [6], and [7]. For a detailed description of the method we refer to [4], [5].

In the sequel we assume that k is chosen independently of x and h , that each grid line section contains sufficiently many points from $\bar{\Omega}_h$ for the extrapolation process and that $\bar{\Omega}_h$ is grid connected.

Solution arcs through turning points

In [5], the described FD-approximation of equation (1) has been considered for fixed λ . It has been proved that for $1 \leq k \leq 4$ isolated solutions of (1) are uniformly approximated by solutions of the discrete problem and that the global error admits an asymptotic expansion in h .

Definition: A solution u of (1) is called isolated iff for $g \equiv 0$ the linearized problem

has only the trivial solution $w = 0$.

As the error bounds in [5] depend on the smoothness of u and on the norm of the solution operator of (2), solution curves containing non-isolated solutions of (1) cannot be treated by the methods of [5] directly. It is possible, however, to adapt the method of proof in [5] to non-isolated solutions of turning point type.

Definition (cf. [3]): A solution (λ^*, u^*) of (1) is called a turning point iff

- (i) equation (2) with $g=0$, as well as its adjoint, possess a 1-dimensional solution space spanned by ϕ^* resp. ψ^* .
- (ii) equation (2) is not solvable for $g = \frac{\partial}{\partial \lambda} f(\lambda^*, \cdot, u^*(\cdot))$.

It is well known that solution arcs through turning points should not be parametrized by λ but by some auxiliary parameter, say s . This may be done (cf. e.g. [3]) by introducing a continuously differentiable functional ℓ on $\mathbb{R} \times L^2(\bar{\Omega}, \mathbb{R}^m)$ which satisfies

$$\ell(\lambda^*, u^*) = 0 \quad (3a)$$

$$\frac{\partial}{\partial u} \ell(\lambda^*, u^*) \phi^* \neq 0. \quad (3b)$$

Equation (1) is then augmented by the equation

$$\ell(\lambda(s), u(s)) = s. \quad (1a)$$

Analogously, an equation of the form

$$\ell_h(\lambda_h(s), u_h(s)) = s$$

has to be added to the FD-approximation of (1). ℓ_h can be considered as a FD-approximation of ℓ .

For sets $M, M' \subset \mathbb{R}^n$, $u: M \rightarrow \mathbb{R}^m$ let $\chi(M')u: M' \rightarrow \mathbb{R}^m$ be given by

$$(\chi(M')u)(x) := \begin{cases} u(x) & \text{for } x \in M \\ 0 & \text{for } x \in M' \setminus M. \end{cases}$$

The following theorem provides an asymptotic expansion in h for the error of the approximation of a solution curve of the continuous problem which may contain a turning point.

Theorem 1

Assumptions:

- (i) Smoothness conditions: Let $2 \leq k \in \mathbb{N}$, $0 < \alpha < 1$, $f \in C^{k,\alpha}$ and $f \in C^{k,\alpha}(\mathbb{R} \times \bar{\Omega} \times \mathbb{R}^m, \mathbb{R}^m)$. Further let ℓ be a three times continuously differentiable functional on $\mathbb{R} \times L^2(\Omega, \mathbb{R}^m)$.
- (ii) Existence of a solution curve for the continuous problem:
Let $\varepsilon > 0$, $J := [-\varepsilon, \varepsilon]$ and $(\lambda(\cdot), u(\cdot)) \in C^1(J, \mathbb{R} \times C^{k,\alpha}(\bar{\Omega}))$ be

a curve of (classical) solutions of (1), (1a). It is assumed, that $u(s)$ is isolated for $s \neq 0$ and either isolated or a turning point for $s=0$. If $u(0)$ is non-isolated, (3a), (3b) is assumed to hold for ℓ .

- (iii) Conditions on the discretization: Assume $1 \leq k \leq 4$, where k denotes the degree of extrapolation near Γ . Let $\Lambda_0 := (h_n)_{n \in \mathbb{N}} \rightarrow 0$. For $\mu \in \mathbb{R}$, $w \in C^{k,\alpha}(\bar{\Omega})$ and $h \in \Lambda_0$ let $\ell_h(\mu, \chi(\bar{\Omega}_h)w) - \ell(\mu, w) = O(h^\gamma)$, $\gamma > 0$ and $\ell_h(\mu, \chi(\bar{\Omega}_h)w) - \ell(\mu, w) \rightarrow 0$, $h \rightarrow 0$.
- (iv) Definition of the first expansion coefficients: Let $s \in J$, $y_s(x) := (\lambda(s), x, u(s, x))$. For $k \geq 4$ let $(\lambda_s^{(1)}, e_s^{(1)})$ be the (unique) solution of

$$\begin{aligned} -\Delta e_{s,j}^{(1)}(x) - \left(\frac{\partial}{\partial u} f(y_s(x)) e_s^{(1)}(x) \right)_j \\ = \sum_{i=1}^n \frac{2}{4!} \frac{\partial}{\partial x_i^4} u_j(s, x) + \frac{\partial}{\partial \lambda} f_j(y_s(x)) \lambda_s^{(1)}, \quad 1 \leq j \leq m, \quad x \in \Omega \end{aligned} \quad (4a)$$

$$e_{s,j}^{(1)}(x) = 0 \quad 1 \leq j \leq m, \quad x \in \Gamma \quad (4b)$$

$$\frac{\partial}{\partial \lambda} \ell(\lambda(s), u(s)) \lambda_s^{(1)} + \frac{\partial}{\partial u} \ell(\lambda(s), u(s)) e_s^{(1)} = 0. \quad (4c)$$

For $k < 4$ define $\lambda_s^{(1)} := 0$ and $e_s^{(1)} := 0$.

- (v) Definition of the second expansion coefficients: Let $s \in J$, $y_s(x) := (\lambda(s), x, u(s, x))$ and $z_s := (\lambda(s), u(s))$. For $k \geq 6$ let $(\lambda_s^{(2)}, e_s^{(2)})$ be the (unique) solution of

$$\begin{aligned} -\Delta e_{s,j}^{(2)}(x) - \left(\frac{\partial}{\partial u} f(y_s(x)) e_s^{(2)}(x) \right)_j \\ = \sum_{i=1}^n \left(\frac{2}{6!} \frac{\partial^6}{\partial x_i^6} u_j(s, x) + \frac{2}{4!} \frac{\partial^4}{\partial x_i^4} e_{s,j}^{(1)}(x) \right) + \frac{\partial}{\partial \lambda} f_j(y_s(x)) \lambda_s^{(2)} \end{aligned} \quad (5a)$$

$$\begin{aligned} &+ \frac{1}{2} \left(\frac{\partial^2}{\partial \lambda^2} f(y_s(x)) \lambda_s^{(1)} \lambda_s^{(1)} + 2 \frac{\partial^2}{\partial \lambda \partial u} f(y_s(x)) \lambda_s^{(1)} e_s^{(1)}(x) \right. \\ &\left. + \frac{\partial^2}{\partial u^2} f(y_s(x)) e_s^{(1)}(x) e_s^{(1)}(x) \right)_j, \quad 1 \leq j \leq m, \quad x \in \Omega \end{aligned}$$

$$e_{s,j}^{(2)}(x) = 0, \quad 1 \leq j \leq m, \quad x \in \Gamma \quad (5b)$$

$$\begin{aligned} & \frac{\partial}{\partial \lambda} \ell(z_s) \lambda_s^{(2)} + \frac{\partial}{\partial u} \ell(z_s) e_s^{(2)} \\ &= -\frac{1}{2} \left(\frac{\partial^2}{\partial \lambda^2} \ell(z_s) \lambda_s^{(1)} \lambda_s^{(1)} + 2 \frac{\partial^2}{\partial \lambda \partial u} \ell(z_s) \lambda_s^{(1)} e_s^{(1)} + \frac{\partial^2}{\partial u^2} \ell(z_s) e_s^{(1)} e_s^{(1)} \right) \end{aligned} \quad (5c)$$

For $\kappa > 6$ define $\lambda_s^{(2)} := 0$ and $e_s^{(2)} := 0$.

Assertion:

There exist constants $C, h_0, \varepsilon_0 > 0$ such that for $h \in \Lambda_0$, $h \leq h_0$, there is a continuous solution arc $(\lambda_h(s), u_h(s))$, $s \in J_0 := [-\varepsilon_0, \varepsilon_0]$, of the FD-approximation of (1), (1a), which satisfies:

$$\max_{s \in J_0} \max_{x \in \Omega_h} \|u_h(s, x) - u(s, x) - h^2 e_s^{(1)}(x) - h^4 e_s^{(2)}(x)\|_\infty \leq Ch^\sigma, \quad (6a)$$

$$\max_{s \in J_0} |\lambda_h(s) - \lambda(s) - h^2 \lambda_s^{(1)} - h^4 \lambda_s^{(2)}| \leq Ch^\sigma, \quad (6b)$$

where $\sigma := \min(\kappa + \alpha - 2, \kappa + 1, \gamma)$.

Proof: A proof for regular solution arcs is essentially contained in [5]. For a detailed proof of the general case, which is quite lengthy, it is referred to a forthcoming paper. However, it should be mentioned, that for $n \geq 2$ the abstract theory of the approximation of solution curves through turning points (cf. [2], [3]) is not applicable directly, as it is not possible to prove the spectral stability of the linearization of the FD-operators respectively the discrete norm-convergence of the inverses of the discretizations of the Laplacian to the solution operator of the continuous Poisson equation with respect to the L_∞ -Norm. This is connected to the fact, that no discrete analogues of the classical Schauder estimates, uniform in h , are known for general Ω .

Remarks: (i) Theorem 1 is stated in a way, which shows the basic facts and avoids undue technicalities. The smoothness assumptions may be weakened slightly. At the same time, some of the results, especially those concerning the dependence of $\lambda^{(1)}$, $\lambda^{(2)}$, $e^{(1)}$, $e^{(2)}$ and u_h on s could be strengthened.

(ii) If $u(0)$ is isolated, we may choose $\ell(\lambda, u) := \lambda$.

Approximation of Simple Turning Points

Theorem 1 does not assert, that turning points of the continuous problem are approximated by turning points of its FD-approximation. On the contrary, if (1) has a turning point (λ^*, u^*) , its FD-approximation does not necessarily have one, even if h is chosen very small.

The situation is different for simple turning points, which are inherited by sufficiently good FD-approximations.

Definition: A turning point (λ^*, u^*) of (1) is simple iff equation (2) is not solvable for $g = \frac{\partial^2}{\partial u^2} f(\lambda^*, \cdot, u^*(\cdot)) \phi^*(\cdot) \phi^*(\cdot)$.

Equation (1) may be augmented in such a way that simple turning points of (1) are precisely the isolated solutions of the enlarged system (cf.[3]). In general, this leads to a semilinear elliptic system with twice as many components as (1), supplemented by a scalar normalization equation.

Simple turning points of (1) are approximated by simple turning points of its FD-approximation in the following sense:

Theorem 2

- (i) Let $\Lambda_0 := (h_n)_{n \in \mathbb{N}} \downarrow 0$, $1 \leq k \leq 4$, $\kappa \geq 2$, $0 < \alpha < 1$ and $f \in C^{k,\alpha}$.
 - (ii) Let $(\lambda^*, u^*) \in \mathbb{R} \times C^{k,\alpha}(\bar{\Omega})$ be a simple turning point of (1).
 - (iii) Let $\varepsilon < 0$, $J_0 := [\lambda^* - \varepsilon, \lambda^* + \varepsilon]$ and $f \in C^{k-2,\alpha}(J_0 \times \bar{\Omega} \times \mathbb{R}^m, \mathbb{R}^m)$
- Then there are constants C , $h_0 > 0$, reals $\lambda^{(1)}, \lambda^{(2)}$ and functions $e^{(1)} \in C^{k-2,\alpha}(\bar{\Omega})$, $e^{(2)} \in C^{\bar{k},\alpha}(\bar{\Omega})$, $\bar{k} := \max(0, \kappa - 4)$, such that for $h \in \Lambda_0$ with $h < h_0$ there is a simple turning point (λ_h^*, u_h^*) of the FD-approximation of (1) which satisfies

$$\max_{x \in \Omega_h} \|u_h^*(x) - u^*(x) - h^2 e^{(1)}(x) - h^4 e^{(2)}(x)\|_\infty \leq Ch^\sigma, \quad (7a)$$

$$|\lambda_h^* - \lambda^* - h^2 \lambda^{(1)} - h^4 \lambda^{(2)}| \leq Ch^\sigma, \quad (7b)$$

where $\sigma := \min(\kappa - 2 + \alpha, \kappa + 1)$.

The proof, which is similar to that of theorem 1, will be published elsewhere.

Remarks: (i) Remark (i) following theorem 1 holds for the present theorem, too.

(ii) Similarly to theorem 1, $(\lambda^{(1)}, e^{(1)})$ and $(\lambda^{(2)}, e^{(2)})$ are given as solutions of certain semilinear elliptic systems, which have to be supplemented by a scalar normalization equation. These defining elliptic systems for $(\lambda^{(i)}, e^{(i)})$, $i=1,2$, are obtained in the standard way (cf.[6]): One applies the FD-approximation-formula for (1) to the Ansatz $\bar{u}_h := u^0 + h^2 e^{(1)} + h^4 e^{(2)}$ and $\bar{\lambda}_h := \lambda^0 + h^2 \lambda^{(1)} + h^4 \lambda^{(2)}$ and determines $(\lambda^{(i)}, e^{(i)})$, $i=1,2$, such that the residual is of maximal order in h . The equations obtained are too lengthy to be given here.

Concluding Remarks

(i) It is quite obvious from the proofs in [5] and their adaptations to theorem 1 and theorem 2, that similar results can be obtained for other types of FD-approximations of (1). Apart from the consistency of the schemes and some rather technical prerequisites it is basic to the proofs, that the restriction of the discretization of the Laplacian to $\overset{\circ}{\Omega}_h$ satisfies a discrete maximum principle, whereas the restriction of the same operator to Γ_h^x is strictly and uniformly diagonally dominant.

(ii) Numerical experiments in which the asymptotic expansions are used to obtain numerical results of high accuracy have been carried out for the regular case. They are reported in [5]. Similar experiments for the turning point case are in preparation.

Literatur

- [1] Böhmer, K.: Asymptotic expansion for the discretization error in linear elliptic boundary value problems on general regions. *Math. Z.* 177, 235-255 (1981)
- [2] Brezzi, F.; Rappaz, J.; Raviart, P.A.: Finite dimensional approximation of nonlinear problems. Part II: Limit points. *Numer. Math.* 37, 1-28 (1981)
- [3] Moore, G.; Spence, A.: The convergence of operator approximations at turning points. *IMA J. Num. Anal.* 1, 23-38 (1981)
- [4] Munz, H.: Uniform expansions for a class of finite difference schemes for elliptic boundary value problems. *Math. Comp.* 36, 155-170 (1981)
- [5] Munz, H.: Asymptotisches Verhalten des Fehlers eines Differenzenverfahrens für semilineare Systeme partieller Differentialgleichungen. Dissertation, Universität Tübingen, 1983
- [6] Pereyra, V.; Proskurowski, W.; Widlund, O.: High order fast Laplace solvers for the Dirichlet problem on general regions. *Math. Comp.* 31, 1-16 (1977)
- [7] Starius, G.: Asymptotic expansions for a class of finite difference schemes. *Math. Comp.* 37, 321-326 (1981)

Harry Munz
Lehrstuhl für Biomathematik
Universität Tübingen
Auf der Morgenstelle 28
D-7400 Tübingen
West-Germany

GLOBAL ASPECTS OF NEWTON'S METHOD FOR
NONLINEAR BOUNDARY VALUE PROBLEMS*

Heinz-Otto Peitgen and Michael Prüfer
Forschungsschwerpunkt "Dynamische Systeme", Fachbereich
Mathematik, Universität Bremen, D-2800 Bremen 33

ABSTRACT. Using Newton's method to compute solutions of a discrete boundary value problem amounts to iterating a certain map in \mathbb{R}^N , and solutions appear as attractors of the dynamical system thus defined. This note is an experimental study of the global properties of the basins of attraction for these attractors. Particular interest is in a comparison with fundamental properties of Julia sets for rational functions in the complex plane.

1. INTRODUCTION

In 1879 A. CAYLEY [3] posed the problem to characterize the global basins of attraction for Newton's method applied to polynomial equations $p(z) = 0$ in the complex plane, see [14]. In a subsequent paper [4] he gave a first result for the special case $p(z) = z^2 - 1$. Cases of higher degree polynomials remained unsettled for quite some time for reasons which became clear by the beautiful and fundamental works of P. FATOU [9, 10] and G. JULIA [11]. Indeed, this innocent looking problem turns out to be one of the most difficult problems in dynamical systems theory as well as complex analysis and though many basic facts are known by now due to the pioneering works of FATOU and JULIA, the major questions have been open now for more than a century.

Recently, the subject of iterating functions in the complex

*) Research was partially supported by 'Stiftung Volkswagenwerk'

plane has become very popular again and has attracted mathematicians from various fields and even physicists. Publications are mushrooming and we list only a few: A. DOUADY and J. H. HUBBARD [8], D. SULLIVAN [18], M. WIDOM, D. BENSIKON, L. P. KADANOFF and S. J. SHENKER [19], D. RUELLE [16], P. CVITANOVIĆ and J. MYRHEIM [6], N. S. MANTON and M. NAUENBERG [13], J.H. CURRY, L. GARNETT and D. SULLIVAN [5]. It is remarkable that most of these papers are based on tremendous computer and computer graphical experiments. Some, e. g. [5], are entirely experimental in nature and this may very well be the origin of a new mathematical discipline: experimental mathematics.

Our interest in this note is to extend CAYLEY'S problem to systems of nonlinear equations in some \mathbb{R}^N , as they are obtained, for example, by a discretized boundary value problem. Our experiments and results here are preliminary. They were motivated by our experiments in [14]. Details and a more rigorous discussion will be published elsewhere.

Before we discuss some of our experiments we give a short review of some of the fundamental facts for Julia sets, which can be found, e. g., in [2, 7, 9, 10, 11]. Let $\Sigma = \mathbb{C} \cup \{\infty\}$ be the Riemannian sphere and let $R : \Sigma \rightarrow \Sigma$ be a rational function of degree greater than or equal two. A fixed point $a = R(a)$ is called *attractive* (a is an *attractor*), provided that $|R'(a)| < 1$. The *basin of attraction* of a is the set

$$(1.1) \quad A(a) = \{z \in \Sigma : R^n(z) \rightarrow a \text{ as } n \rightarrow \infty\} .$$

A fixed point $r = R(r)$ is called *repulsive* (r is a *repeller*), provided $|R'(r)| > 1$. If $z = R^k(z)$, i.e. z is a *periodic point* of R or a *cycle* of order $k \in \mathbb{N}$, then we extend our definitions simply by working with the k -th iterate R^k . Note that if p is a polynomial, then the roots of p are attractors for

$$(1.2) \quad R(z) = z - \frac{p(z)}{p'(z)} .$$

For the investigation of the basins of attraction the following set - the *Julia set* of R - is of fundamental importance:

$$(1.3) \quad J = \{z \in \Sigma : R \text{ is not normal in } z\}.$$

Recall that R is not normal in z , provided that there exists a neighborhood U of z such that the family $\{R^n|_U\}_{n \in \mathbb{N}}$ of mappings from U to Σ is not equicontinuous.

The following is a collection of facts about J :

THEOREM 1.1 (FATOU, JULIA)

Let R be any rational function of Σ of degree greater than or equal to two.

- (1) $J \neq \emptyset$
- (2) $R(J) = J$ and $R^{-1}(J) = J$
(i. e. J is *completely invariant*)
- (3) If a is an attractor for R (fixed point or periodic point), then

$$J = \partial A(a)$$
(i ∂ denotes the boundary).
- (4) Let U be an open set and $J \cap U \neq \emptyset$.
Then there exists $k \in \mathbb{N}$ such that

$$J = R^k(U \cap J)$$
.
(i.e. J is *self-similar*)
- (5) Let $z_0 \in J$, then $J = \text{cl} \{z \in \Sigma : R^k(z) = z_0 \text{ for some } k \in \mathbb{N}\}$
(cl denotes the closure).
- (6) $J = \text{cl} \{r \in \Sigma : r \text{ is a repulsive cycle for } R\}.$

There are many more striking properties of J , such as $R|_J$ is a chaotic dynamical system or J is typically a fractal set, i.e. the Hausdorff dimension of J is greater than 1 (see [16]). We note that property (3) has a surprising consequence: Let p be a polynomial with roots $p(z_i) = 0$, $i = 1, \dots, \deg(p)$, and let R be the Newton map (1.2). Then

$$(1.4) \quad \partial A(z_i) = \partial A(z_j)$$

for any pair of roots z_i, z_j . As a result the boundaries of the basins of attraction must have a very complex structure, if p has more than two distinct roots.

2. NEWTON'S METHOD IN \mathbb{R}^N

As a model problem we investigate the boundary value problem

$$(2.1) \quad \begin{cases} u'' + \lambda f(u) = 0 \\ u(0) = 0 = u(1), \end{cases}$$

where $f(s)$ is a polynomial with $f(0) = 0$ and $f'(0) = 1$. It is well known that for any $\lambda_k = (k\pi)^2$, $k \in \mathbb{N}$, there is a bifurcation of nontrivial solutions for (2.1) from $u \equiv 0$ [15]. Moreover, if $C_k \subset C[0,1] \times \mathbb{R}$ denotes the branch bifurcating from $(0, \lambda_k)$, then C_k is characterized by those solutions of (2.1) which have precisely $(k-1)$ internal zeros. In this note we restrict ourselves to the case

$$(2.2) \quad f(s) = s - s^2.$$

A more general discussion will be given elsewhere. Figure 1 gives a good picture of C_1 and C_2 near $u \equiv 0$ and $\lambda > 0$.

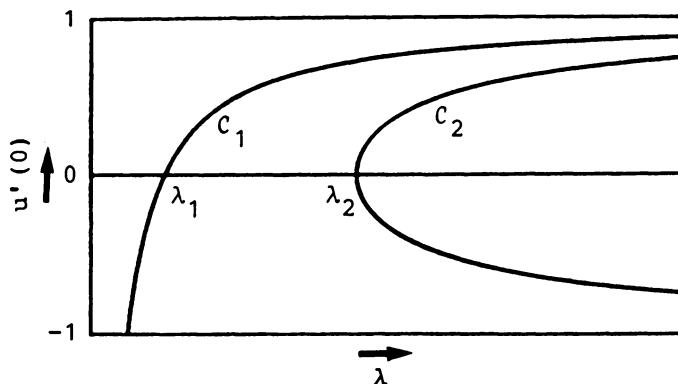


Figure 1: Bifurcation diagram for (2.1) with $f(s) = s - s^2$

Using a Sturm-Liouville comparison theorem and elementary symmetry considerations one shows easily that indeed C_1 bifurcates transcritically, whereas C_2 bifurcates supercritically.

As a discretization we choose for simplicity finite differences, i.e.

$$\begin{aligned} t_i &= i \cdot h, \quad h = 1/(N+1), \\ x_i &= u(t_i), \quad x = (x_1, \dots, x_N)^T. \end{aligned}$$

Thus, the discretized form of (2.1) is obtained by

$$(2.3) \quad G_\mu(x) := Ax - \mu F(x) = 0,$$

where $\mu = \lambda h^2$, $F(x) = (f(x_1), \dots, f(x_N))^T$, and A is the familiar matrix

$$\begin{pmatrix} 2 & -1 & & & & \mathbf{0} \\ -1 & \ddots & \ddots & & & \\ & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & -1 \\ \mathbf{0} & & & \ddots & \ddots & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

Now Newton's method for the computation of the discrete solutions reads

$$(2.4) \quad \begin{cases} R_\mu(x) = x - DG_\mu^{-1}(x) G_\mu(x), \\ x \in D \subset \mathbb{R}^N. \end{cases}$$

Unlike (1.2) the Newton map (2.4) is not defined in all of \mathbb{R}^N . What, however, is a proper domain of definition for R_μ ? Naturally, one has to exclude the set

$$(2.5) \quad S_\mu = \{x \in \mathbb{R}^N : \det(DG_\mu(x)) = 0\}.$$

In the complex case the Julia set J is the set of all points where Newton's method fails to converge. What is the analogous notion and definition here? A natural choice seems to be the following:

$$(2.6) \quad J_\mu = \text{cl} \{x \in \mathbb{R}^N : R_\mu^k(x) \in S_\mu \text{ for some } k \in \mathbb{N}\} .$$

Almost by definition it follows that

$$(2.7) \quad R_\mu(J_\mu) \subset J_\mu .$$

Motivated by theorem 1.1 we would like to understand $R_\mu^{-1}(J_\mu)$, i.e. complete invariance. It is clear, however, that dealing with a map which is constituted by real polynomials we may very well have points $y \in J_\mu$ such that $R_\mu^{-1}(y)$ is empty. For a more detailed analysis we will now restrict ourselves to the case $N = 2$, because there computer graphical experiments may be used to enhance our intuition.

Let now $t_1 = 1/3$, $t_2 = 2/3$, $z = (x, y)$ and $\mu = \lambda/9$. It is easy to analyze the bifurcation problem $G_\mu(z) = 0$. One obtains precisely two points of bifurcation from the trivial solution $z \equiv 0$ at

$$(2.8) \quad \mu_1 = 1 \quad \text{and} \quad \mu_2 = 3 .$$

The corresponding branches \tilde{C}_1 and $\tilde{C}_2 \subset \mathbb{R}^2 \times \mathbb{R}$ can be determined explicitly:

$$(2.9) \quad \tilde{C}_1 = \{((\frac{\mu-1}{\mu}, \frac{\mu-1}{\mu}), \mu) : \mu > 0\} .$$

We omit the tedious calculations for \tilde{C}_2 here. We note, however, that if $(z, \mu) \in \tilde{C}_2$ and $z = (x, y)$, then $\text{sign}(x) \neq \text{sign}(y)$.

In the present situation it is elementary to calculate S_μ :

$$(2.10) \quad S_\mu = \{ (x, y) : y(4\mu - 2\mu^2 + 4\mu^2 x) = -3 + 4\mu - \mu^2 - (4\mu - 2\mu^2)x \} .$$

Moreover, one Newton step

$$(2.11) \quad \begin{pmatrix} \bar{x} \\ \bar{y} \end{pmatrix} = R_\mu \begin{pmatrix} x \\ y \end{pmatrix}$$

is carried out by solving the equations

$$(2.12) \quad \begin{cases} \mu[f(x) - xf'(x)] = 2\bar{x} - \bar{y} - \mu\bar{x}f'(x) , \\ \mu[f(y) - yf'(y)] = 2\bar{y} - \bar{x} - \mu\bar{y}f'(y) , \end{cases}$$

which are, of course, linear in \bar{x} and \bar{y} . In view of (2.6) we are, however, also interested in finding preimages, i.e. given $(\bar{x}, \bar{y})^T$, find $(x, y)^T$, such that (2.11) holds. Thus, we see from (2.12) that finding preimages amounts to solving a nonlinear system of two decoupled equations. Since f is a polynomial of degree 2 each of these equations has either two or no solutions. Note that if we had chosen for f a polynomial of odd degree then $R_\mu^{-1}(z)$ would contain at least one element for any z . These elementary observations are easily extended to arbitrary dimensions $N > 2$.

Our computer graphical experiments in figures 2, 3, 4, 5 are a first attempt to investigate the basins of attraction, their boundaries and the set J (see (2.6)). For a visualization of the basins of attraction and the dynamics in the basins we have chosen a decomposition into *level sets* and *isochrones*: Let a be an attractive fixed point of R_μ and let $D_0(a, \varepsilon)$ denote the closed Euclidean disc with center a and radius ε . We define

$$(2.13) \quad \begin{cases} \tilde{D}_k(a, \varepsilon) := \{ z \in \mathbb{R}^2 : R_\mu^k(z) \in D_0(a, \varepsilon) \} , \\ D_k(a, \varepsilon) := \tilde{D}_k(a, \varepsilon) \setminus \tilde{D}_{k-1}(a, \varepsilon) , \\ k = 1, 2, 3, \dots . \end{cases}$$

Thus, $D_k(a, \varepsilon)$ collects all those points which after k Newton iterations arrive in $D_0(a, \varepsilon)$.

Our first set of experiments in figure 2 shows the situation for $\mu = 0.5$ (before the first point of bifurcation) and for $\mu = 2$ (between the two points of bifurcation). For both values of μ we have two solutions of (2.3)

$$\left\{ \begin{array}{ll} z_0 = (0,0)^T & \text{trivial solution} \\ z_1 = (-1,-1)^T & \text{negative solution} \end{array} \right\} \mu = 0.5 ,$$

$$\left\{ \begin{array}{ll} z_0 = (0,0)^T & \text{trivial solution} \\ z_2 = (0.5,0.5)^T & \text{positive solution} \end{array} \right\} \mu = 2 .$$

Figures 2a - 2d show the window $[-5,5] \times [-5,5]$. In figure 2a and 2b we see the level sets $D_k(z_i, \varepsilon)$ for $\varepsilon = 0.01$ and $i = 0, 1, 2$. In both figures $D_k(z_i, \varepsilon)$, $i = 0, 1, 2$, is colored black for k even. In figure 2a ($\mu=0.5$) the sets $D_k(z_1, \varepsilon)$, k odd, are shown in white, while the sets $D_k(z_0, \varepsilon)$, k odd, are shaded. Likewise, in figure 2b ($\mu=2$) the sets $D_k(z_0, \varepsilon)$, k odd, are shown in white, while the sets $D_k(z_2, \varepsilon)$, k odd, are shaded.

For our particular choices of μ we obtain the explicit hyperbolas S_μ :

$$(2.14) \quad S_\mu = \left\{ \begin{array}{ll} \{(x,y) : y(x+1.5) = -1.5x - 1.25\} & , \mu = 0.5 \\ \{(x,y) : y(16x) = 1\} & , \mu = 2 . \end{array} \right.$$

Figures 2a and 2b seem to display essentially the same patterns up to a deformation. This is noteworthy, because between $\mu=0.5$ and $\mu=2$ there is a bifurcation. In both figures the positive branch of the hyperbola S_μ , which we denote by S_μ^+ , plays a particular role. One can show that

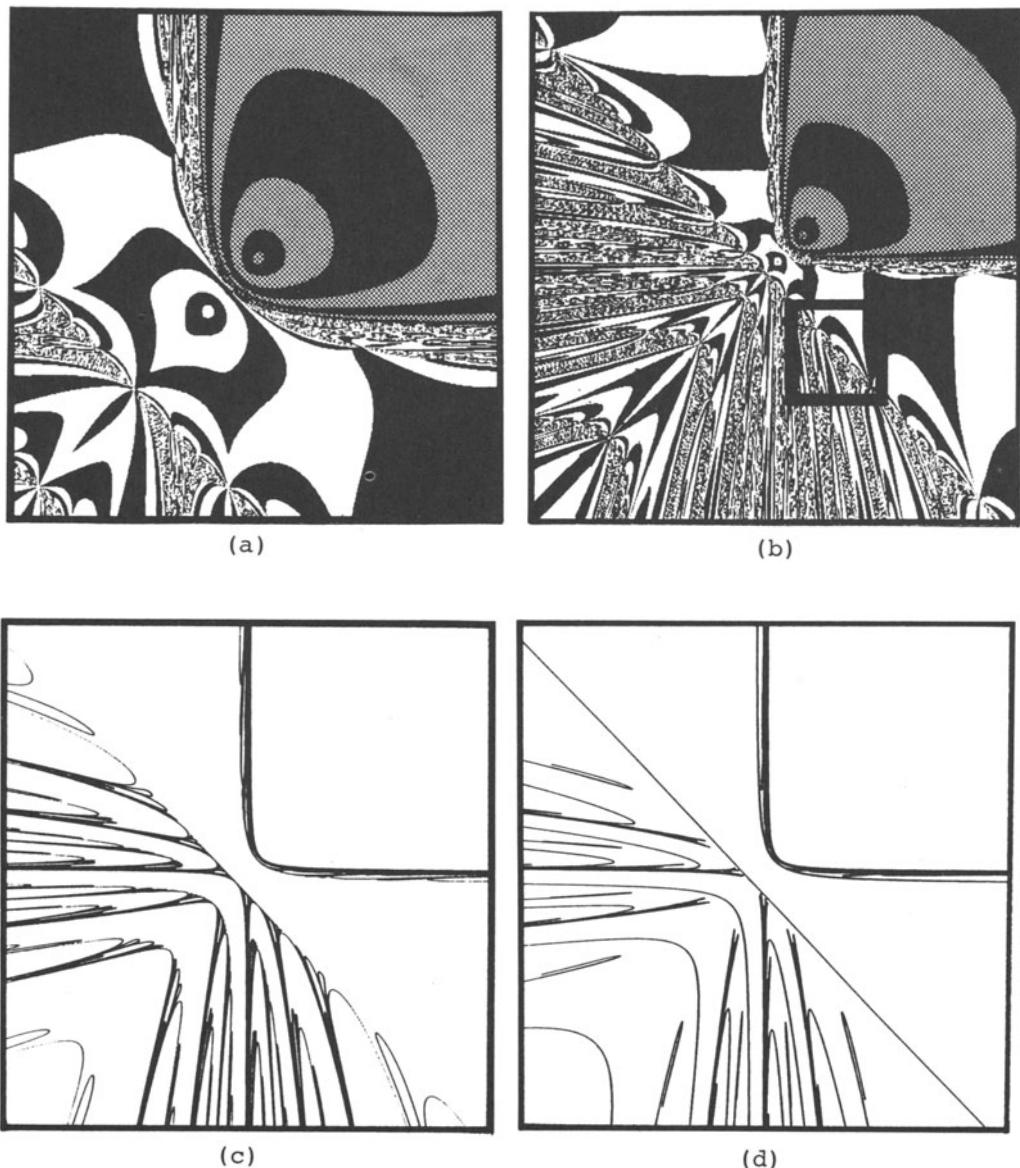


Figure 2: Newton iteration for problem (2.3), $N=2$, with
 $f(s)=s-s^2$, $\mu=0.5$ (figure 2a), $\mu=2$ (figures 2b-2f)

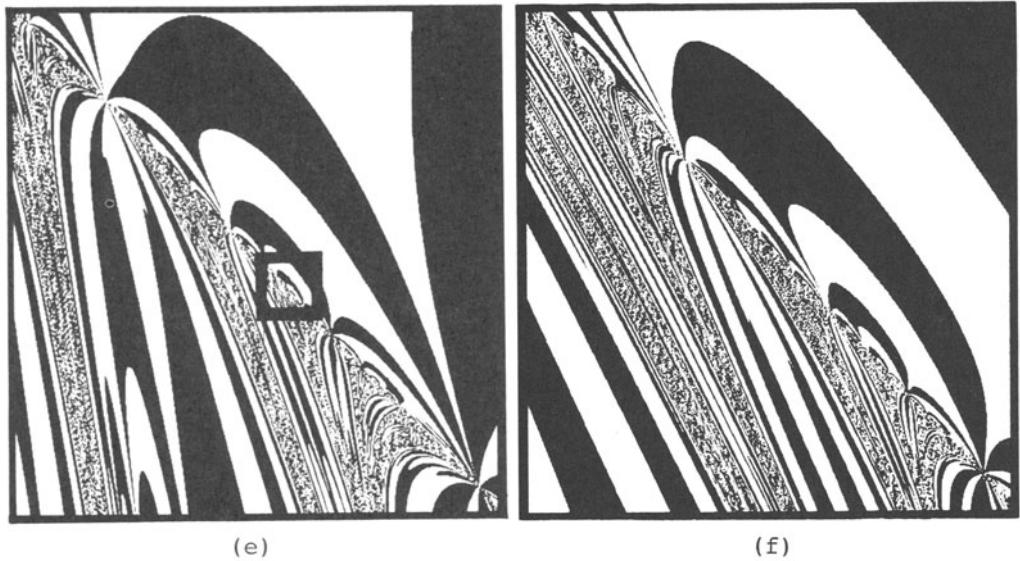


Figure 2 (continued)

$$(2.15) \quad \left\{ \begin{array}{l} A(z_0) = \{(x,y) : x > -1.5 \text{ and } y > (-1.5x-1.25)/(x+1.5)\}, \mu = 0.5 , \\ A(z_2) = \{(x,y) : x > 0 \text{ and } y > 1/16x\}, \mu = 2 . \end{array} \right.$$

Thus, these basins are simply connected and

$$(2.16) \quad \left\{ \begin{array}{l} S_\mu^+ = \partial A(z_0) , \quad \mu = 0.5 , \\ S_\mu^+ = \partial A(z_2) , \quad \mu = 2 . \end{array} \right.$$

The situation for the other roots z_1 (for $\mu = 0.5$) and z_0 (for $\mu = 2$) is much different. Their basins are far from being simply connected. In fact they are not even connected and decompose into an infinity of components. In view of theorem 1.1 (3) and (2.16) we ask: what is their boundary?

Motivated by theorem 1.1 (3) we have investigated the set J_μ according to (2.6). If $\mu = 2$ we have that

$$S_\mu = S_\mu^+ \cup S_\mu^- , \text{ where}$$

$$(2.17) \quad \begin{cases} S_\mu^+ = \{(x, y) : y = 1/16x, x > 0\} \\ S_\mu^- = \{(x, y) : y = 1/16x, x < 0\} . \end{cases}$$

It turns out that

$$(2.18) \quad R_\mu^{-1}(S_\mu^+) = \emptyset ,$$

while $R_\mu^{-1}(S_\mu^-) \neq \emptyset$. Figure 2c shows $R_\mu^{-k}(S_\mu^-)$ for $k = 1, \dots, 7$ and $\mu=2$. We see a remarkably complex structure which apparently fits into the components of $A(z_0)$ in figure 2b. Looking closely at this structure we begin to understand which points have to be added, if we take the closure of all preimages of S_μ . There is an apparent straight line G_μ :

$$(2.19) \quad G_\mu := \{(x, y) : y = -x - 1/2\} .$$

Note that $S_\mu^- \cap G_\mu = \{(-1/4, -1/4)\} = : P_\mu$, i.e. the Newton map R_μ is defined on $G_\mu \setminus \{P_\mu\}$ and in fact it turns out that

$$(2.20) \quad R_\mu(G_\mu \setminus \{P_\mu\}) \subset G_\mu .$$

The dynamics on this interesting subset will be discussed in figure 3. Figure 2d shows $R_\mu^{-k}(G_\mu)$, $k = 1, 2, \dots, 7$. Surprisingly, figure 2d fits very well into figure 2c, i.e. points in $R_\mu^{-k}(G_\mu)$ belong to J_μ . Figure 2c and 2d together with S_μ give a fairly complete picture of J_μ . In view of theorem 1.1 (3) we note the remarkable fact that in our example

$$\partial A(z_0) \neq \partial A(z_2) , \quad \mu = 2 .$$

We conjecture, however, that

$$\partial A(z_0) \cup \partial A(z_2) = J_\mu , \quad \mu = 2 .$$

Figures 2e and 2f are devoted to the question of self-similarity according to theorem 1.1 (4). They show two successive close-ups from figure 2b. (see indicated squares) showing that the level sets $D_k(z_0, \varepsilon)$ accumulate in G_μ with increasing k .

Since $G_\mu \setminus \{P_\mu\}$ is invariant under R_μ the one-dimensional dynamics on G_μ should be revealed by investigating the graph of R_μ when restricted to $G_\mu \setminus \{P_\mu\}$:

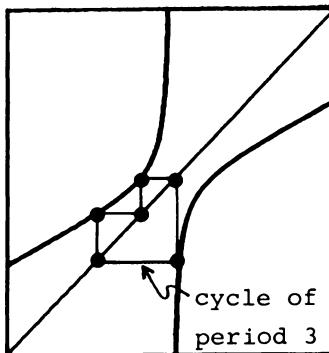


Figure 3: Graph of R_μ restricted to $G_\mu \setminus \{P_\mu\}$

One can show for R_μ restricted to $G_\mu \setminus \{P_\mu\}$:

- (2.21) $\left\{ \begin{array}{l} \text{(a) There are periodic points of any period } k > 2 . \\ \text{(b) There is an uncountable subset of aperiodic} \\ \text{points in the sense of LI-YORKE [12].} \end{array} \right.$

The first assertion follows from an explicit construction, while the second can be shown as an application of lemma 1.2 in [17]. Thus, R_μ is chaotic in the sense of LI-YORKE on $G_\mu \setminus \{P_\mu\}$.

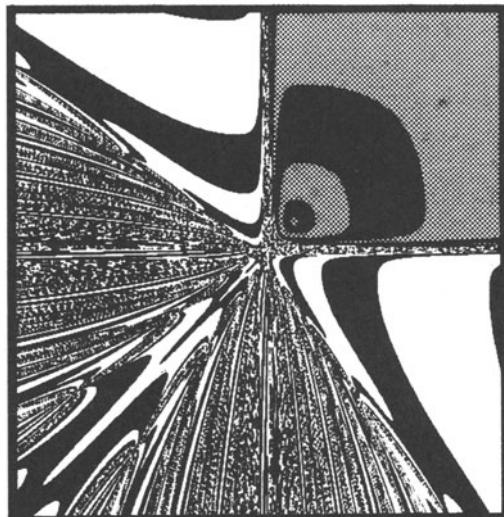
Figure 4 shows results of experiments for $\mu = 3.1$, i.e. after the second bifurcation point, where problem (2.3) has solutions $z_0 = (0,0)^T$, $z_3 \in \tilde{C}_1$, $z_4, z_5 \in \tilde{C}_2$. In figure 4a the level sets (2.13) have been coloured according to figures 2a, 2b: The sets $D_k(z_i, \varepsilon)$, $i = 0, 4, 5$, are white for k odd, while the sets $D_k(z_3, \varepsilon)$ are shaded for k odd. Figures 4b,c show in solid black the basins $A(z_4)$, resp. $A(z_0)$ (the basin $A(z_5)$ is obtained by reflecting $A(z_4)$ at the diagonal $x = y$). As in previous cases one finds that the set S_μ (2.5) is constituted by a hyperbola with two branches S_μ^+, S_μ^- (recall (2.17)). Again one finds $\partial A(z_3) = S_\mu^+$ (cf. figure 4a). The appearance of the two new solutions $z_4, z_5 \in \tilde{C}_2$ obviously has a tremendous impact upon the global dynamics of (2.4): In particular, we conjecture from figures 4b, c that

$$(2.22) \quad \partial A(z_0) = \partial A(z_4) = \partial A(z_5) ,$$

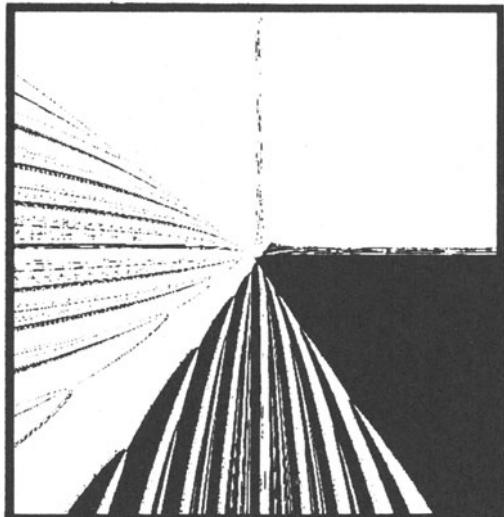
which would provide a remarkable analogue to the result (1.4) for polynomials on the Riemannian sphere.

Figure 4d shows the inverse iterates $R_\mu^{-k}(S_\mu)$, $k = 1, \dots, 7$, and again we find that the set J_μ fits very well into the complex pattern of figures 4a-c. In particular we conjecture that J_μ is equal to the common boundary (2.22) of $A(z_0)$, $A(z_4)$, $A(z_5)$.

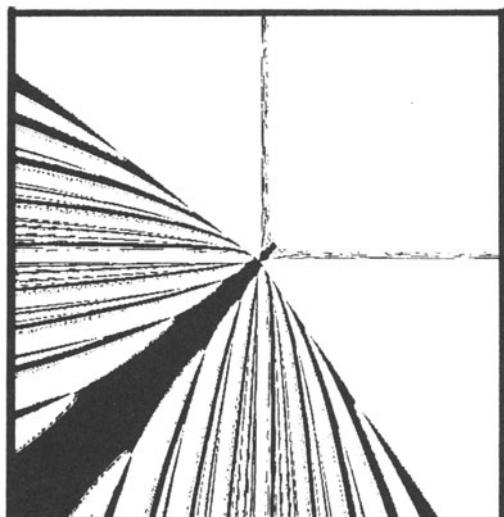
Unlike in figure 2c for $\mu = 2$ we find in figure 4d for $\mu = 3.1$ that J_μ intersects the diagonal $x = y$ only in two points belonging to S_μ^+ , resp. S_μ^- . This is to the effect that $A(z_0)$ now contains a stripe parallel to the diagonal that is not disturbed by J_μ . With increasing μ the influence of J_μ gets restricted even further, as figures 5 a,b show for $\mu = 4$. Figure 5a shows four basins of attraction coloured as in figure 4a, while figure 5b shows the inverse iterates $R_\mu^{-k}(S_\mu)$, $k = 1, \dots, 7$.



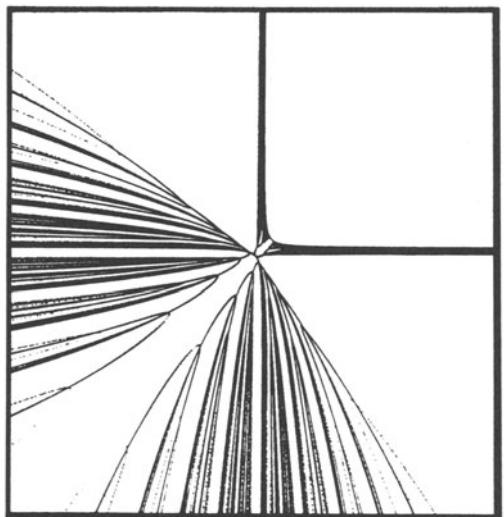
(a)



(b)



(c)



(d)

Figure 4: Newton iteration for problem (2.3), N=2, with
 $f(s)=s-s^2$, $\mu = 3.1$

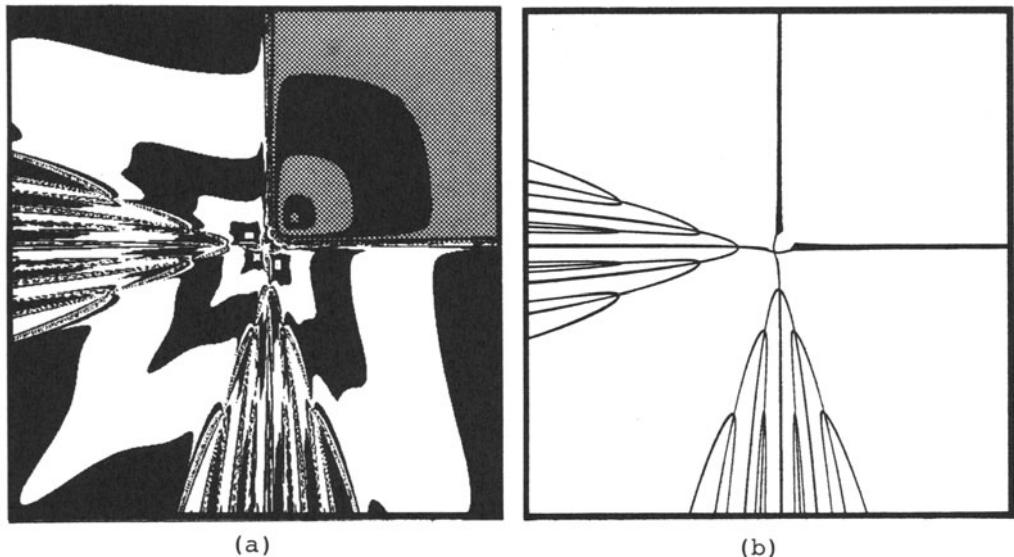


Figure 5: Newton iteration for problem (2.3), $N=2$, with
 $f(s) = s - s^2$, $\mu = 4$

CONCLUDING REMARKS:

Summarizing, we see that our examples for Newton's method in \mathbb{R}^N show both striking similarities as well as differences to Newton's method for rational functions in the complex plane. In particular, it seems noteworthy that Newton's method for discretized boundary value problems apparently gives rise to chaotic subdynamics and basins of attraction with infinitely many components.

For another approach to these questions we refer to the stimulating paper of D. BRAESS [1].

REFERENCES

- [1] BRAESS, D.: Über die Einzugsbereiche der Nullstellen von Polynomen beim Newton-Verfahren
Numer. Math. 29 (1977), 123-132
- [2] BROLIN, H.: Invariant sets under iteration of rational functions
Ark. Math. 6 (1965), 103-144
- [3] CAYLEY, A.: The Newton-Fourier imaginary problem
Am. J. Math. II (1879), 97
- [4] CAYLEY, A.: Sur les racines d'une équation algébrique
CRAS 110 (1890), 215-218
- [5] CURRY, J. H., GARNETT, L., and SULLIVAN, D.: On the iteration of a rational function: Computer experiments with Newton's method
Commun. Math. Phys. 91 (1983), 267-277
- [6] CVITANOVIĆ, P. and MYRHEIM, J.: Universality for period-n-couplings in complex mappings
Physics Letters 94A 8 (1983), 329-333
- [7] DOUADY, A.: Systèmes dynamiques holomorphes
Séminaire Bourbaki 1982/83, No. 599
- [8] DOUADY, A. and HUBBARD, J. H.: Itération des polynômes quadratiques complexes
C. R. Acad. Sc. Paris, t. 294, série I, 123-126
- [9] FATOU, P.: Sur les équations fonctionnelles
Bulletin Soc. Math. Fr. 47 (1919), 161-271
- [10] FATOU, P.: Sur les équations fonctionnelles
Bulletin Soc. Math. Fr. 48 (1920), 33-94, 208-314
- [11] JULIA, G.: Mémoire sur l'itération des fonctions rationnelles
J. Math. Pures et Appl., sér. 8.1 (1918), 47-245
- [12] LI, T.Y. and YORKE, J. A.: Period three implies chaos
Amer. Math. Monthly 82 (1975), 985-992
- [13] MANTON, N. S. and NAUENBERG, M.: Universal scaling behaviour for iterated maps in the complex plane
Commun. Math. Phys. 89 (1983), 555-570

- [14] PEITGEN, H.-O., SAUPE, D., and v. HAESELER, F.: Cayley's Problem and Julia sets
to appear: Mathematical Intelligencer 6 (1984)
- [15] RABINOWITZ, P. H.: Some aspects of nonlinear eigenvalue problems
Rocky Mountain J. Math. 3 (1973), 162-202
- [16] RUELLE, D.: Repellers for real analytic maps
preprint , IHES Bures-sur Yvette 1981
- [17] SIEGBERG, H. W.: Chaotic difference equations: Generic aspects
Trans. Amer. Math. Soc. 279 1 (1983), 205-213
- [18] D. SULLIVAN: Itération des fonctions analytiques complexes
C. R. Acad. Sc. Paris, t. 294, série I, 301-303
- [19] WIDOM, M., BENSIMON, D., KADANOFF, L. P., and SHENKER, S. J.: Strange objects in the plane
preprint, University of Chicago 1983

FINITE DIMENSIONAL APPROXIMATION
OF SOME BIFURCATION PROBLEMS IN PRESENCE OF SYMMETRIES

by

Giuseppe GEYMONAT and Geneviève RAUGEL

0 - Introduction

We give here a brief summary of the results of [4], [14]. The aim of our work was not only the finite dimensional approximation of the specific problem studied in [3], but we also wanted to show how to adapt to discrete problems and how to use, in this case, some classical mathematical tools (e.g. the splitting lemma of Gromoll-Meyer-Magnus, the singularity theory, etc...). Such considerations are developed in [5]. In particular, we prove in [5] that in most cases (e.g. in the examples given in [6], [7], [9], [10]), the discrete bifurcation problem behaves as an imperfect bifurcation problem and we improve the methods of [12]. Another way to use universal unfoldings in the approximation of bifurcation problems is suggested by [2].

In [3], the authors study the qualitative behaviour of spatial buckled states of naturally straight, uniform, nonlinearly hyperelastic rods subjected to terminal load, where the cross-section is invariant under the dihedral group D_n , $n \geq 3$, D_n being the group of symmetries and rotations of the plane that leave invariant a regular polygon with n edges. In Section 1, we recall their model. Section 2 is devoted to the formulation of the exact and discrete problems and to their reduction to two-dimensional ones. Finally, in Section 3, we use the singularity theory for describing the solutions and we give the bifurcation diagrams in the case $D_n = D_3$. For the main definitions and results of the imperfect bifurcation theory, we refer to the classical papers [7] and [8].

1 - The model

The model adopted in [3] is a directory theory based on the Kirchhoff kinetic analogy and on the invariance properties of the cross-sections (see, e.g. [1]).

1.1 - Geometrical assumptions

The main geometrical assumptions can be summed up as follows : (I) the rod is considered as a continuum of plane cross-section slices, that is one supposes that there exists a neutral (unstressed) curve \mathcal{C} coincident with the curve of cross-section centroids and inextensible. \mathcal{C} is parametrized by arc-length $s \in J \equiv [0,1]$. Cross-sections T_s (with centroids s) remain plane and orthogonal to the line \mathcal{C} ; they have no shearing strain. T_s varies homothetically and smoothly with s , is simply-connected and invariant under D_n .

In the Kirchoff kinetic analogy, the deformation is expressed in terms of the rotations mapping a reference cross-section to the slice with centroids s . For this, two orthonormal reference frames of unit vectors are used :

- a) $\vec{e}_1, \vec{e}_2, \vec{e}_3$, for the whole space with \vec{e}_3 pointing in the direction along which the force P is applied ;
- b) $\vec{a}_1(s), \vec{a}_2(s), \vec{a}_3(s)$ local, with \vec{a}_1, \vec{a}_2 , fixed flat to T_s , \vec{a}_2 along the direction of one of the axes of symmetry, with the origin at the centroid of T_s , and $\vec{a}_3 = \vec{a}_1 \times \vec{a}_2$ (see figures 1.1 et 1.2).

If $\vec{r}(s)$ denotes the position in a deformed configuration of the material point at the centroid s , the assumptions (I) mean that $\vec{r}'(s) = \vec{a}_3(s)$. Therefore, if $\vec{r}(s) = \sum_{i=1}^3 x_i(s) \vec{e}_i$, we obtain the conditions of inextensibility $x_1'^2(s) + x_2'^2(s) + x_3'^2(s) = 1$.

The triads $\{\vec{a}_1, \vec{a}_2, \vec{a}_3\}$ are the directors of the theory ; as they are orthonormal, there exists a vector $\vec{u}(s)$, denoting the deformation, such that :

$$(1.1) \quad \vec{a}_i' = \vec{u} \times \vec{a}_i.$$

The components u_i , $1 \leq i \leq 3$, of the deformation \vec{u} , with respect to $\{\vec{a}_i\}$ are the strains. Here, u_1 and u_2 measure the flexure along the axes \vec{a}_1 and \vec{a}_2 , while u_3 measures the twist. Using (1.1) and Frenet formulae, we prove that the flexure $\vec{u}_f = u_1 \vec{a}_1 + u_2 \vec{a}_2$ is a rotation with the binormal to \mathcal{C} as the axis ; we also prove that the twist $u_3 \vec{a}_3$ is a rotation with the tangent vector to \mathcal{C} as the axis.

Usually, the u_i are computed in terms of Euler angles θ, ψ, ϕ ; however, these angles are coordinates on $SO(3)$ which degenerate around the identity, so that it is more convenient to introduce the new coordinate

$\alpha = \frac{\pi}{2} - \phi - \psi$. Thus, one can express (u_1, u_2, u_3) in terms of x_1, x_2, α :

$$(1.2) \quad \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} \cos \alpha & -\sin \alpha & 0 \\ \sin \alpha & \cos \alpha & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1'' - \frac{x_1' x_3''}{1+x_3'} \\ x_2'' - \frac{x_2' x_3''}{1+x_3'} \\ -\alpha' - \frac{x_1' x_2'' - x_1'' x_2'}{1+x_3'} \end{bmatrix}$$

where x_3' and x_3'' are given by

$$(1.3) \quad x_3' = (1 - x_1'^2 - x_2'^2)^{\frac{1}{2}}$$

and

$$(1.4) \quad x_3'' = \frac{x_1' x_1'' + x_2' x_2''}{(1 - x_1'^2 - x_2'^2)^{\frac{1}{2}}}.$$

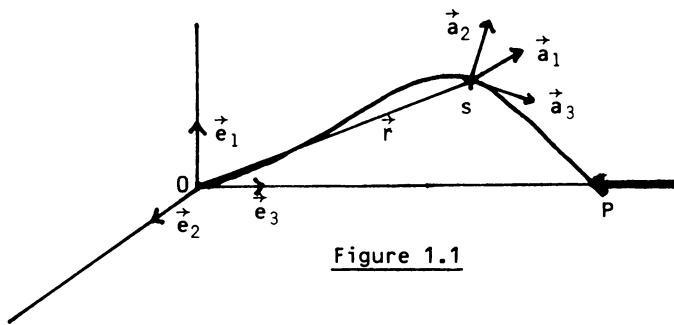


Figure 1.1

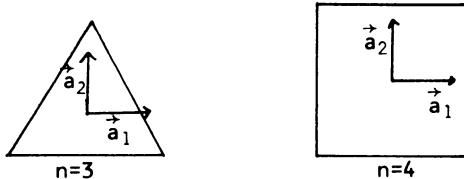


Figure 1.2

1.2 - Assumptions on the deformation energy

Following the usual director approach to rod theory, it is assumed in [3] that : (II) the deformation energy is given by

$$(1.5) \quad \int_0^1 W(\vec{u}, s) ds,$$

where $W(\vec{u}, s)$ is a \mathcal{C}^∞ function in the variables \vec{u} and s (this means that one considers only hyperelastic materials).

Let us define the action \mathcal{R}_n of $D_n \oplus \mathbb{Z}_2$ on \vec{u} by

$$\mathcal{R}_{n,(\gamma,\varepsilon)}\vec{u} = \begin{bmatrix} Y_{11} & Y_{12} & 0 \\ Y_{21} & Y_{22} & 0 \\ 0 & 0 & \varepsilon \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix},$$

where $\gamma = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \in D_n \subset O(2)$ and $\varepsilon = \pm 1$. With the choice of coordinates made in (1.1), it is generated by

$$(1.6) \quad \mathcal{R}_{n,r} = \begin{bmatrix} \cos 2\pi/n & -\sin 2\pi/n & 0 \\ \sin 2\pi/n & \cos 2\pi/n & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{R}_{n,s} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{R}_{n,t} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

In order to distinguish between D_n -invariant cross-sections and $O(2)$ -invariant cross-sections, the usual Kirchhoff symmetry assumption is relaxed, in [3], in the following one : (III) we have

$$(1.7) \quad W(\mathcal{R}_{n,(\gamma,\varepsilon)}\vec{u}, s) = W(\vec{u}, s), \quad \forall (\gamma, \varepsilon) \in D_n \oplus \mathbb{Z}_2.$$

Thanks to a deep theorem of G. Schwarz, it is proven in [3] that there exists a C^∞ function $H : \mathbb{R}^3 \times \mathbb{R} \rightarrow \mathbb{R}$ such that

$$(1.8)_a \quad W(\vec{u}, s) = H(\tau, s),$$

where $\tau = (\tau_1, \tau_2, \tau_3)$ and

$$(1.8)_b \quad \tau_1 = u_1^2 + u_2^2, \quad \tau_2 = \operatorname{Re}(u_1 + i u_2)^n, \quad \tau_3 = u_3^2.$$

Let $r = (r_1, r_2, r_3) \in \mathbb{N}^3$, $|r| = r_1 + r_2 + r_3$ and $\tau^r = \tau_1^{r_1} \tau_2^{r_2} \tau_3^{r_3}$.

Then the expansion of H to order K is

$$(1.9) \quad H(\tau, s) = \sum_{|r| \leq K} H_r(s) \tau^r + O(|\tau|^{K+1}).$$

Hereafter, one assumes : (IV)

$$(1.10) \quad H_{000}(s) \equiv 0,$$

$$(1.11) \quad H_{100}(s) > 0, \quad H_{001}(s) > 0, \quad \forall s \in J.$$

Such hypotheses can be justified, if one expresses the elastic potential $W(\vec{u}, s)$ in terms of the three-dimensional nonlinear elasticity theory (see [3], Section I, paragraph 1.5). For instance, $H_{001}(s) = \frac{1}{2} C(s)$, where $C(s)$ is the torsional rigidity, $H_{100}(s) = \frac{E_0 I_{10}}{2}$ where $E_0 > 0$ is the Young's modulus and $I_{10}(s)$

is the principal moment of inertia of T_s .

2 - Formulation of the exact and discrete problems and their reduction to two-dimensional problems

2.1 - The exact problem

From the model of Section 1, one deduces the energy functional f presumed to be (locally) minimized by the rod. This consists of the deformation energy minus the work done by the *terminal load force P*, in moving along its line of action. The end of the rod $s=0$ is held fixed, while the end $s=1$ moves :

$$1 - x_3(1) = \int_0^1 (1 - x_3'(s)) ds \quad (x_3(0) = 0),$$

so that, from (1.2), (1.5) and (1.8), we obtain :

$$(2.1) \quad f(x_1, x_2, x_3, P) = \int_0^1 H(\tau, s) - P(1 - x_3'(s)) ds,$$

where H is given by (1.8)_a, τ by (1.8)_b and x_3' by (1.3).

Many sets of boundary conditions can be considered :

(s.s.) rod simply supported at both ends ;

(c.s.) rod clamped at the first end and simply supported at the other ;

(c.c.) rod clamped at both ends.

The different kinds of boundary conditions are written in terms of the displacements x_1, x_2, α in the following way :

(i) at a simply supported end (e.g. $s=0$), we must have :

$$(2.2)_a \quad x_1(0) = x_2(0) = 0$$

and

$$(2.2)_b \quad \frac{\partial W}{\partial u_1} \Big|_{s=0} = \frac{\partial W}{\partial u_2} \Big|_{s=0} = \frac{\partial W}{\partial u_3} \Big|_{s=0} = 0.$$

(The conditions (2.2)_b are natural boundary conditions coming from the minimization of f);

(ii) at a clamped end (e.g. $s=0$), we must have :

$$(2.3) \quad x_1(0) = x_2(0) = x_1'(0) = x_2'(0) = \alpha(0) = 0.$$

For the sake of simplicity, we only consider the case (c.s.) here.

In order to apply some results of differential geometry, that are proved only for C^∞ functions, one must choose an adequate functional framework.

To this end, we set

$$x = x^1 \times x^2 \times x^3$$

where

$$x^i = \left\{ x_i \in C^2(J) ; x_i(0) = x'_i(0) = x''_i(1) = 0 \right\}, \quad i = 1, 2$$

and

$$x^3 = \left\{ \alpha \in C^1(J) ; \alpha(0) = 0 \right\}.$$

We shall denote by $x = (x_1, x_2, \alpha)$ the elements of X , and by Ω the set $\{(x, P) \in X \times \mathbb{R} ; \max_{s \in J} \{x_1'^2(s) + x_2'^2(s)\} < 1\}$. Since (1.3) takes sense for $x_1'^2(s) + x_2'^2(s) \leq 1$, f is in $C^\infty(\Omega; \mathbb{R})$. We set

$$H_*^2(J) = \left\{ y \in H^2(J) \cap H_0^1(J) ; y'(0) = 0 \right\},$$

where $H^2(J)$ and $H_0^1(J)$ are the usual Sobolev spaces.

In [3], we have the following result :

PROPOSITION 2.1 : For each $P \in \mathbb{R}$, the unstressed configuration $(x_1, x_2, \alpha) = (0, 0, 0)$ is a critical point of $f : \Omega \rightarrow \mathbb{R}$. For

$$(2.4) \quad P < P_0 = \min_{y \in H_*^2(J)} \frac{\int_0^1 2 H_{100}(s) y''^2 ds}{\int_0^1 y'^2(s) ds},$$

the configuration $(0, 0, 0)$ corresponds to a strict (local) minimum of f and hence to a stable equilibrium. Moreover, the quadratic form on X , $D_{xx}^2 f(0, P_0)[x, x]$ is degenerate. It follows that $(0, 0, 0, P_0)$ is a possible bifurcation point for the minima of $f : \Omega \rightarrow \mathbb{R}$.

Therefore, we can state the following variational bifurcation problem :

$$\begin{cases} \text{(VBP)} \quad \text{Find the minima of the functional } f : \Omega \rightarrow \mathbb{R} \\ \text{for } (x, P) \in \Omega, \text{ near } (0, P_0). \end{cases}$$

Remark 2.1 : One has :

$$D_{xx}^2 f(0, P)[x, x] = \int_0^1 \left\{ 2 H_{100}(s) (x_1''^2 + x_2''^2) + 2 H_{001}(s) \alpha'^2 - P(x_1'^2 + x_2'^2) \right\} ds.$$

Remark 2.2 : Since $H_{100} = \frac{E_0 I_{10}}{2}$, P_0 equals the well-known critical load for the buckling of an elastic rod, i.e. the first eigenvalue of

$$(2.5) \quad \begin{cases} (2 H_{100}(s) \varphi'')'' + P \varphi'' = 0, \\ \varphi(0) = \varphi'(0) = 0, \quad \varphi(1) = \varphi''(1) = 0. \end{cases}$$

If H_{100} is constant (i.e. if the unstressed rod has a constant cross-section), the first eigenvalue P_0 is simple. Hereafter, we assume that P_0 is simple and $\varphi_0(s)$ denotes the relevant eigenfunction (suitably normalized).

Consider now the action $R_n : D_n \rightarrow \text{Aut}(X)$ given by

$$R_{n\gamma} x = \begin{bmatrix} Y_{11} & Y_{12} & 0 \\ Y_{21} & Y_{22} & 0 \\ 0 & 0 & \det Y \end{bmatrix} \begin{bmatrix} x_1(s) \\ x_2(s) \\ x_3(s) \end{bmatrix},$$

where $\gamma = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix} \in D_n \subset O(2)$, and generated by

$$R_{n_r} = \begin{bmatrix} \cos 2\pi/n & -\sin 2\pi/n & 0 \\ \sin 2\pi/n & \cos 2\pi/n & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ and } R_{n_s} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}.$$

One can prove the

PROPOSITION 2.2 : $f : \Omega \rightarrow \mathbb{R}$ is R_n -invariant, i.e.

$$(2.6)_i \quad (R_{n\gamma} \times \text{id}_{\mathbb{R}}) \Omega \subset \Omega, \forall \gamma \in D_n;$$

$$(2.6)_{ii} \quad f(R_{n\gamma} x, P) = f(x, P), \forall \gamma \in D_n, \forall x \in X.$$

2.2 - The discrete problem

Let us next study the finite-dimensional approximation of problem (VBP). For each value of a real parameter $h > 0$ which will tend to zero, we introduce a finite-dimensional subspace X_h^i of X^i , $i = 1, 2, 3$, with $X_h^1 = X_h^2$. We set : $X_h = X_h^1 \times X_h^2 \times X_h^3$. We denote by $x_h = (x_{1h}, x_{2h}, x_{3h})$ the elements of X_h and by Ω_h the set $\{(x_h, P) \in X_h \times \mathbb{R} ; \max_{s \in J} \{x_{1h}^{i2}(s) + x_{2h}^{i2}(s)\} < 1\}$. Finally, we set $f_h = f|_{X_h}$.

We can now state the approximate variational bifurcation problem :

$$(VBP_h) \quad \left\{ \begin{array}{l} \text{Find the minima of the functional } f_h : \Omega_h \rightarrow \mathbb{R} \\ \text{for } (x_h, P) \in \Omega_h \text{ near } (0, P_0). \end{array} \right.$$

Clearly, one has the :

PROPOSITION 2.3 : $f_h : \Omega_h \rightarrow \mathbb{R}$ is R_n -invariant, i.e.

$$(2.7)_i \quad (R_{n\gamma} \times \text{id}_{\mathbb{R}}) \Omega_h \subset \Omega_h, \forall \gamma \in D_n;$$

$$(2.7)_{ii} \quad f_h(R_{n\gamma} x_h, P) = f_h(x_h, P), \forall \gamma \in D_n, \forall x_h \in X_h.$$

Let us now define the following norms :

$$\|g\|_{k,\infty} = \max_{s \in J, 0 \leq i \leq k} \{ |g^{(i)}(s)| \} \quad \text{if } g \in W^{k,\infty}(J)$$

and

$$\|g\|_{k,2} = \left(\sum_{i=0}^k \int_0^1 |g^{(i)}(s)|^2 ds \right)^{\frac{1}{2}} \quad \text{if } g \in H^k(J).$$

Now, we make some assumptions on the method of approximation :

(a) There exists an integer $k \geq 2$ (resp. $k \geq 1$) and an operator $r_h^i \in L(x_i^i, x_h^i)$, $i = 1, 2$ (resp. r_h^3), such that, for all $x_i \in X^i$, $i = 1, 2$ (resp. $\alpha \in X^3$), for all ℓ with $2 \leq \ell \leq k$ (resp. $1 \leq \ell \leq k$), one has

$$(2.8)_i \quad \|x_i - r_h^i x_i\|_{0,j} + h \|x_i - r_h^i x_i\|_{1,j} + h^2 \|x_i - r_h^i x_i\|_{2,j} \leq c h^{\ell+1} \|x_i\|_{\ell+1,j}; \quad (i=1,2)$$

$$(2.8)_{ii} \quad \|\alpha - r_h^3 \alpha\|_{0,j} + h \|\alpha - r_h^3 \alpha\|_{1,j} \leq c h^{\ell+1} \|\alpha\|_{\ell+1,j} \quad (j=2 \text{ or } \infty)$$

(b) There exists a constant $c > 0$ such that, for all $x_h^i \in X_h^i$, $i = 1, 2$, (resp. $\alpha_h \in X_h^3$), one has

$$(2.9) \quad \|x_h^i\|_{0,\infty} \leq c h^{-\frac{1}{2}} \|x_h^i\|_{0,2}; \quad \|\alpha_h\|_{0,\infty} \leq c h^{-\frac{1}{2}} \|\alpha_h\|_{0,2}.$$

(c) If $X^{i''} = \{x_i'' ; x_i \in X^i\}$, $i = 1, 2$, and $X^{3'} = \{\alpha' ; \alpha \in X^3\}$ ($X_h^{i''}$ and $X_h^{3'}$ are defined likewise) and if \mathcal{P}_h^i , $i = 1, 2$ (resp. \mathcal{P}_h^3) denotes the L^2 -projection from $X^{i''}$ onto $X_h^{i''}$ (resp. from $X^{3'}$ onto $X_h^{3'}$), there exists a constant $c > 0$ such that

$$(2.10) \quad \|\mathcal{P}_h^i y\|_{0,\infty} \leq c \|y\|_{0,\infty}; \quad \forall y \in X_h^{i''}, i = 1, 2 \text{ or } \forall y \in X_h^{3'}.$$

Using (a) and (b), we prove ([14]) that there exists a sequence $(P_{0h})_h$ converging to P_0 and a sequence of suitably normalized vectors $(\varphi_{0h})_h$ converging to φ_0 in the norm of $\mathcal{C}^2(J)$ such that

$$(2.11) \quad \int_0^1 (2 H_{100} \varphi_{0h}'' + P_{0h} \varphi_{0h}') \varphi_h'' ds = 0, \quad \forall \varphi_h \in X_h^i, i = 1 \text{ or } 2.$$

2.3 - The reduction to two-dimensional problems

Now we want to reduce the problems (VBP) and (VBP_h) to two-dimensional ones. To this end, we use the splitting lemma of Magnus [11], which is a generalization of Morse's lemma and its discretized form [4]. With respect to the Liapunov-Schmidt procedure, the splitting lemma allows a simpler and more rigorous

study of the stability.

In [4], [14], we prove the following results :

. Let $V = \left\{ v = (\eta_1 \varphi_0, \eta_2 \varphi_0, 0) \in X ; \eta_1, \eta_2 \in \mathbb{R} \right\}$ (resp. $V_h = \left\{ v_h = (\eta_1 \varphi_{0h}, \eta_2 \varphi_{0h}, 0) \in X_h ; \eta_1, \eta_2 \in \mathbb{R} \right\}$) and $Z = \left\{ z = (z_1, z_2, \alpha) \in X ; \int_0^1 \varphi'' z_i'' ds = 0, i = 1, 2 \right\}$ (resp. $Z_h = \left\{ z_h = (z_{1h}, z_{2h}, \alpha_h) \in X_h ; \int_0^1 \varphi_h'' z_{ih}'' ds = 0, i = 1, 2 \right\}$, so that $X = V \oplus Z$ (resp. $X_h = V_h \oplus Z_h$).

. Then, there exist

(i) a local \mathcal{C}^∞ -diffeomorphism Φ (resp. Φ_h) defined on a neighbourhood $U \subset \Omega$ (resp. $U_h \subset \Omega_h$) of $(0, P_0)$ and preserving $(0, P_0)$ (resp. $(0, P_{0h})$) :

$$\Phi : (v \oplus z, P) \in (V \oplus Z) \times \mathbb{R} \rightarrow (v \oplus \varphi(v \oplus z, P), P)$$

$$(\text{resp. } \Phi_h : (v_h \oplus z_h, P) \in (V_h \oplus Z_h) \times \mathbb{R} \rightarrow (v_h \oplus \varphi_h(v_h \oplus z_h, P), P)),$$

where $\varphi(v \oplus z, P) \in Z$ (resp. $\varphi_h(v_h \oplus z_h, P) \in Z_h$). Moreover, the diameter of U_h does not depend on h .

(ii) a \mathcal{C}^∞ mapping $L : U' \rightarrow Z$ (resp. $L_h : U'_h \rightarrow Z_h$), where $U' = U \cap (V \times \mathbb{R})$ (resp. $U'_h = U_h \cap (V_h \times \mathbb{R})$), which satisfies $L(0, P) \equiv 0$ for $(0, P) \in U'$ (resp. $L_h(0, P) \equiv 0$ for $(0, P) \in U'_h$), and is equivariant under the action R_n of D_n , such that

$$(2.12) \quad f(\Phi(v \oplus z, P)) = \sum_{j=1}^2 \frac{1}{2} \int_0^1 [2 H_{100} z_j''^2 - P_0 z_j'^2] ds + \frac{1}{2} \int_0^1 2 H_{001} \alpha'^2 ds + \tilde{f}(n_1, n_2, \lambda),$$

where $v = (\eta_1 \varphi_0, \eta_2 \varphi_0, 0) \in V$, $z = (z_1, z_2, \alpha) \in Z$, $\lambda = P - P_0$ and

$$(2.13) \quad \tilde{f}(n_1, n_2, \lambda) = f(v \oplus L(v, P), P)$$

(resp.

$$(2.14) \quad f_h(\Phi_h(v_h \oplus z_h, P)) = \sum_{j=1}^2 \frac{1}{2} \int_0^1 [2 H_{100} z_{jh}''^2 - P_{0h} z_{jh}'^2] ds + \frac{1}{2} \int_0^1 2 H_{001} \alpha_h'^2 ds + \tilde{f}_h(n_1, n_2, \lambda),$$

where $v_h = (\eta_1 \varphi_{0h}, \eta_2 \varphi_{0h}, 0) \in V_h$, $z_h = (z_{1h}, z_{2h}, \alpha_h) \in Z_h$ and

$$(2.15) \quad \tilde{f}_h(n_1, n_2, \lambda) = f_h(v_h \oplus L_h(v_h, P), P).$$

The so-called reduced functionals \tilde{f} and \tilde{f}_h are invariant under the action ρ_n of D_n on $\mathbb{R}^2 \simeq \mathbb{C}$, generated by

$$(2.16) \quad \rho_{n\underline{r}} \zeta = e^{2i\pi/n} \zeta \quad \text{and} \quad \rho_{n\underline{s}} \zeta = \bar{\zeta}$$

where $\zeta = \eta_1 + i \eta_2$.

. Moreover, one has

$$(2.17) \quad D_n \tilde{f}(0,0,\lambda) = D_n \tilde{f}_h(0,0,\lambda) = 0, \text{ for } \lambda \text{ near } 0 \\ \text{and}$$

$$(2.18) \quad D_{nn}^2 \tilde{f}(0,0,0) = D_{nn}^2 \tilde{f}_h(0,0,\lambda_0) = 0, \text{ where } \lambda_0 = P_{0h} - P_0.$$

. Furthermore, the functions \tilde{f}_h converge to \tilde{f} , together with all their derivatives at any order. The functions L_h (resp. Ψ_h) converge to L (resp. Ψ) in some sense that will not be precised here. We also obtain optimal error estimates (see [4], [14]).

. Finally, there exists a one-to-one and onto correspondence between the critical points (resp. minima) of $f(x,P)$ on U (resp. of $f_h(x_h,P)$ on U_h) and those of \tilde{f} (resp. \tilde{f}_h) on U' (resp. U'_h). This correspondence is given by

$$x = v + L(v,P) \quad (\text{resp. } x_h = v_h + L_h(v_h,P)),$$

where $v = (n_1\varphi_0, n_2\varphi_0, 0)$, $P = P_0 + \lambda$, $n = (n_1, n_2)$ is a critical point of \tilde{f} (resp. $v_h = (n_1\varphi_{0h}, n_2\varphi_{0h}, 0)$, $n = (n_1, n_2)$ is a critical point of \tilde{f}_h).

If one defines the isotropy subgroups Σ_v of v (resp. Σ_{v_h} of v_h) and Σ_x of x (resp. Σ_{x_h} of x_h), as the subgroups of D_n that leave invariant v and x (resp. v_h and x_h), then the isotropy subgroups Σ_x and Σ_v (resp. Σ_{x_h} and Σ_{v_h}) are equal. Therefore, the VBP (resp. VBP_h) is equivalent to seeking the minima of \tilde{f} (resp. of \tilde{f}_h) for (n, λ) near $(0, 0, 0)$.

3 - Use of singularity theory and bifurcation diagrams

3.1 -

So we have to solve the equations :

$$(3.1) \quad D_n \tilde{f}(n, \lambda) = 0$$

and

$$(3.2) \quad D_n \tilde{f}_h(n, \lambda) = 0,$$

and to study the sign of the eigenvalues of $D_{nn}^2 \tilde{f}(n, \lambda)$ (resp. $D_{nn}^2 \tilde{f}_h(n, \lambda)$) on the solutions of (3.1) (resp. (3.2)) for (n, λ) near $(0, 0, 0)$. Let

$$\begin{cases} G(\zeta, \lambda) = G_1(\zeta, \lambda) + i G_2(\zeta, \lambda), \\ G_h(\zeta, \lambda) = G_{1h}(\zeta, \lambda) + i G_{2h}(\zeta, \lambda), \end{cases}$$

where

$$G_j(\zeta, \lambda) = \frac{\partial \tilde{f}}{\partial n_j}(\zeta, \lambda), \quad G_{jh}(\zeta, \lambda) = \frac{\partial \tilde{f}_h}{\partial n_j}(\zeta, \lambda), \quad j = 1 \text{ or } 2.$$

Let us remark that, since \tilde{f} and \tilde{f}_h are ρ_n -invariant (i.e. $\tilde{f}(\rho_{n\gamma}\zeta, \lambda) = \tilde{f}(\zeta, \lambda)$, $\forall \gamma \in D_n$, $\zeta \in \mathbb{C}$), G and G_h are ρ_n -equivariant (i.e. $G(\rho_{n\gamma}\zeta, \lambda) = \rho_{n\gamma}G(\zeta, \lambda)$, $\forall \zeta \in \mathbb{C}$, $\forall \gamma \in D_n$). We set $\sigma = (\sigma_1, \sigma_2)$ with $\sigma_1 = \zeta\bar{\zeta}$ and $\sigma_2 = \operatorname{Re} \zeta^n$.

PROPOSITION 3.1 : For all real C^∞ -mapping $G : \mathbb{C} \times \mathbb{R} \rightarrow \mathbb{C}$ that is ρ_n -equivariant, there exists a pair (p, q) of C^∞ -functions from $\mathbb{R}^2 \times \mathbb{R}$ into \mathbb{R} such that

$$(3.3) \quad G(\zeta, \lambda) = p(\sigma(\zeta), \lambda)\zeta + q(\sigma(\zeta), \lambda)\zeta^{n-1}.$$

If G_h are ρ_n -equivariant C^∞ -mappings converging to G together with all their derivatives at any order, there exists a sequence of pairs $(p_h, q_h)_h$ of C^∞ -functions from $\mathbb{R}^2 \times \mathbb{R}$ into \mathbb{R} such that

$$(3.4) \quad G_h(\zeta, \lambda) = p_h(\sigma(\zeta), \lambda)\zeta + q_h(\sigma(\zeta), \lambda)\zeta^{n-1},$$

and the functions p_h (resp. q_h) converge to p (resp. q) together with all their derivatives at any order.

We obtain also error estimates which are not given here. Hereafter, we shall write $G(\zeta, \lambda)$ and $G_h(\zeta, \lambda)$ in the following way :

$$G(\zeta, \lambda) \equiv (p(\sigma, \lambda), q(\sigma, \lambda)), \quad G_h(\zeta, \lambda) \equiv (p_h(\sigma, \lambda), q_h(\sigma, \lambda)).$$

3.2 - How to solve the bifurcation equations

If $\tilde{\zeta} \in \mathbb{C}$ is a solution of (3.1) (resp. (3.2)), $\rho_{n\gamma}\tilde{\zeta}$ is also a solution of (3.1) (resp. (3.2)). Hence, G (resp. G_h) vanishes on orbits of the action of ρ_n and it is sufficient to describe a unique representative for each zero-orbit of $G(\zeta, \lambda) = 0$ (resp. $G_h(\zeta, \lambda) = 0$).

THEOREM 3.1 : A unique representative for each zero-orbit of $G(\zeta, \lambda) = 0$ (resp. $G_h(\zeta, \lambda) = 0$) is found by solving the following system of equations :

	Equation	Isotropy group
(0)	$\eta_1 = \eta_2 = 0$	D_n
(1)	$p(\eta_1^2, \eta_1^n, \lambda) + q(\eta_1^2, \eta_1^n, \lambda)\eta_1^{n-2} = 0$	\mathbb{Z}_2 (reflection through the axis $\eta_1 = 0$)
(1) _h	$p_h(\eta_1^2, \eta_1^n, \lambda) + q_h(\eta_1^2, \eta_1^n, \lambda)\eta_1^{n-2} = 0$ $\eta_2 = 0 ; \eta_1 \neq 0$ if n is odd ; $\eta_1 > 0$ if n is even.	

$$\begin{aligned}
 (2) \quad p(\tilde{\eta}_1^2, -\tilde{\eta}_1^n, \lambda) - q(\tilde{\eta}_1^2, -\tilde{\eta}_1^n, \lambda) \tilde{\eta}_1^{n-2} &= 0 & \text{Z}_2 \quad (\text{reflection through the} \\
 (\text{resp.}) \quad p_h(\tilde{\eta}_1^2, -\tilde{\eta}_1^n, \lambda) - q_h(\tilde{\eta}_1^2, -\tilde{\eta}_1^n, \lambda) \tilde{\eta}_1^{n-2} &= 0 & \text{the axis } \eta_1 = (\tan \pi/n) \tilde{\eta}_1 \\
 (2)_h \quad \text{with } \eta_2 = \sin(\pi/n) \tilde{\eta}_1, \tilde{\eta}_1 = (\cos \pi/n)^{-1} \eta_1 &> 0. & \\
 (2) \text{ and } (2)_h \text{ are valid for } n \text{ even only} \\
 (3) \quad p = q = 0 & \\
 (\text{resp.}) \quad p_h = q_h = 0 & \quad \{1\} . \\
 (3)_h \quad \operatorname{Im} \zeta^n \neq 0. &
 \end{aligned}$$

We do not give any detail or result about the stability here.

3.3 - Use of the singularity theory (I)

Now we set

$$G = (p, q), \quad G_h = (p_h, q_h)$$

where

$$\begin{cases} p = A_1 \lambda + A_2 \sigma_1 + A_3 \sigma_2 + \tilde{p}, \\ q = B_0 + B_1 \lambda + B_2 \sigma_1 + B_3 \sigma_2 + \tilde{q}; \end{cases}$$

$$\begin{cases} p_h = A_{0h} + A_{1h} \lambda + A_{2h} \sigma_1 + A_{3h} \sigma_2 + \tilde{p}_h, \\ q_h = B_{0h} + B_{1h} \lambda + B_{2h} \sigma_1 + B_{3h} \sigma_2 + \tilde{q}_h; \end{cases}$$

with $\tilde{p} \in \mathcal{M}^2$, $\tilde{q} \in \mathcal{M}^2$, $\tilde{p}_h \in \mathcal{M}^2$, $\tilde{q}_h \in \mathcal{M}^2$. Here, \mathcal{M} denotes the ideal of $\mathcal{C}^\infty(\mathbb{R}^2 \times \mathbb{R}, \mathbb{R})$ generated by $\lambda, \sigma_1, \sigma_2$. From Proposition 3.1, \tilde{p}_h (resp. \tilde{q}_h) converges to \tilde{p} (resp. \tilde{q}) together with all their derivatives at any order. Moreover, one has :

$$\lim_{h \rightarrow 0} A_{0h} = 0; \quad \lim_{h \rightarrow 0} A_{jh} = A_j; \quad \lim_{h \rightarrow 0} B_{jh} = B_j; \quad j = 0, 1, 2, 3.$$

As in [13], one proves the

PROPOSITION 3.2 : If $A_1 \neq 0$, there exist a neighbourhood I of 0 in \mathbb{R} , functions λ, λ_h in $\mathcal{C}^\infty(I, \mathbb{R})$ (resp. $\tilde{\lambda}, \tilde{\lambda}_h$ in $\mathcal{C}^\infty(I, \mathbb{R})$) such that the solutions of (1) and (1)_h (resp. (2) and (2)_h) are given by $(\eta_1, 0, \lambda(\eta_1))$ and $(\eta_1, 0, \lambda_h(\eta_1))$ (resp. $((\cos \pi/n)\tilde{\eta}_1, (\sin \pi/n)\tilde{\eta}_1, \tilde{\lambda}(\tilde{\eta}_1))$ and $((\cos \pi/n)\tilde{\eta}_1, (\sin \pi/n)\tilde{\eta}_1, \tilde{\lambda}_h(\tilde{\eta}_1))$). Moreover, one has, for all $\eta_1, \tilde{\eta}_1$ in I :

$$\begin{aligned}
 (3.5) \quad |\lambda(\eta_1) - \lambda_h(\eta_1)| &\leq K \max_{s \in I} (|p(s^2, s^n, \lambda) - p_h(s^2, s^n, \lambda)| + |q(s^2, s^n, \lambda) - q_h(s^2, s^n, \lambda)|) \\
 (\text{resp.}) \quad (3.6) \quad |\tilde{\lambda}(\tilde{\eta}_1) - \tilde{\lambda}_h(\tilde{\eta}_1)| &\leq K \max_{s \in I} (|p(s^2, -s^n, \lambda) - p_h(s^2, -s^n, \lambda)| + |q(s^2, -s^n, \lambda) - q_h(s^2, -s^n, \lambda)|) \\
 \text{and} \quad (3.7) \quad \lambda(0) = \tilde{\lambda}(0) = 0, \quad \lambda_h(0) = \tilde{\lambda}_h(0) = -\frac{A_{0h}}{A_{1h}} = \lambda_{0h} \quad (\lambda_{0h} \text{ is given by (2.18)}).
 \end{aligned}$$

Let $\tilde{G}_h(\zeta, \lambda)$ be the mapping $G_h(\zeta, \lambda - \lambda_{0h})$. From the singularity theory point of view, we obtain the (see [4]) :

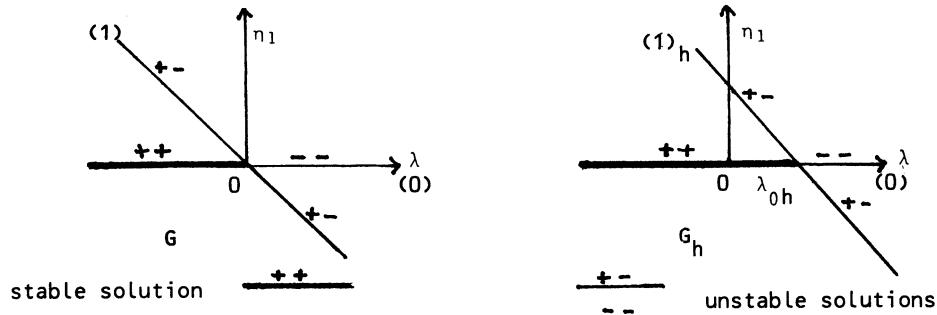
PROPOSITION 3.3 : Assume $A_1 \neq 0$ and $B_0 \neq 0$, throughout

- (i) For $n=3$, G (resp. \tilde{G}_h) is D_3 -equivalent to $(k_1\lambda, \varepsilon_0)$ with $k_1 = (\text{sgn } A_1)1$, $\varepsilon_0 = (\text{sgn } B_0)1$. Moreover, $\text{codim } G = \text{codim } \tilde{G}_h = 0$.
- (ii) For $n=4$, if $A_2^2 \neq B_0^2$, G (resp. \tilde{G}_h) is D_4 -equivalent to $(k_1\lambda + a\sigma_1, \varepsilon_0)$ (resp. $(k_1\lambda + a_h\sigma_1, \varepsilon_0)$) with $\text{sgn } (a \pm \varepsilon_0) = \text{sgn } (a_h \pm \varepsilon_0) = \text{sgn } (A_2 \pm B_0)$, $a^2 \neq 1$, $a_h^2 \neq 1$ and $\lim_{h \rightarrow 0} a_h = a$. Moreover, $\text{codim } G = \text{codim } \tilde{G}_h = 1$, a universal unfolding of G is given by $(k_1\lambda + \tilde{a}\sigma_1, \varepsilon_0)$ with \tilde{a} near a .
- (iii) For $n \geq 5$, if $A_2 \neq 0$, G (resp. \tilde{G}_h) is D_n -equivalent to $(k_1\lambda + k_2\sigma_1, \varepsilon_0)$ with $k_2 = (\text{sgn } A_2)1$.

In the three cases above, there exists a neighbourhood N of $(0, 0, 0)$ independent of h , such that equations (3) and (3)_h have no solution in N .

Remark 3.1 : The cases $A_1 \neq 1$, $B_0 \neq 0$ (and in particular those described in the Proposition 3.3) arise in the boundary problems (s.s.) and (c.s.).

We draw below the diagrams of the solutions of $G=0$ and $G_h=0$ in the case D_3 , with $k_1=-1$, $\varepsilon_0=-1$. Of course, we give an orbit-representative of each branch (the other ones can be obtained by rotations of $2\pi/3$) :



3.4 - Use of the singularity theory (II)

If the unstressed rod has a constant cross-section, one can prove that, in the case $D_n = D_3$, one has, for the boundary problem (c.c.) :

$$(3.8) \quad B_0 = 0, A_1 \neq 0, A_2 \neq 0, A_1 B_2 - B_1 A_2 = 0, A_2 B_3 - A_3 B_2 \neq 0.$$

Thus, $\text{codim } G \neq 0$, and this case gives rise to more interesting problems.

In [3] and [4], one proves the

PROPOSITION 3.4 : Assume that (3.8) holds. Then

(i) G is D_3 -equivalent to

$$(3.9) \quad (k_1 \lambda + k_2 \sigma_1 + a \sigma_2, \varepsilon_1 \sigma_2)$$

where $k_1 = (\text{sgn } A_1)$, $k_2 = (\text{sgn } A_2)$, $\varepsilon_1 = \text{sgn } (A_2(A_2B_3 - B_2A_3))$, $\text{codim } G = 3$.

(ii) \tilde{G}_h is D_3 -equivalent to

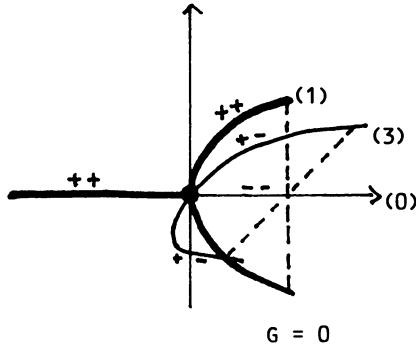
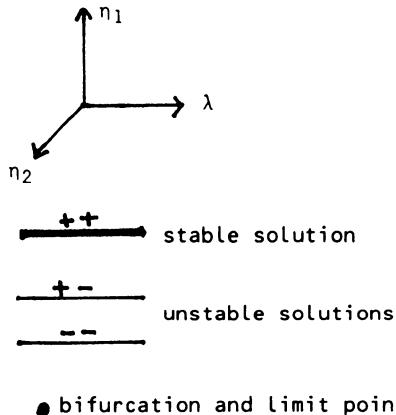
$$(3.10) \quad (k_1 \lambda + k_2 \sigma_1 + a_h \sigma_2, b_{0h} + b_{1h} \sigma_1 + \varepsilon_2 \sigma_2)$$

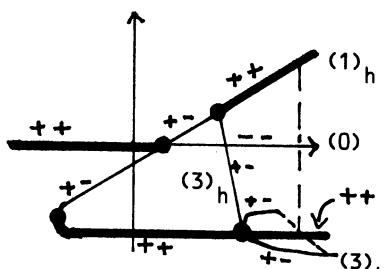
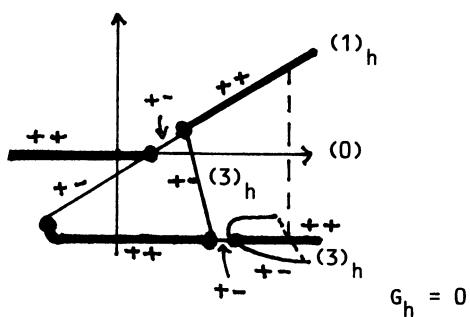
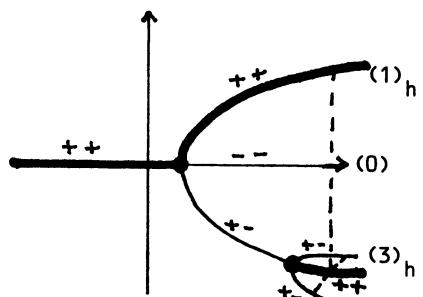
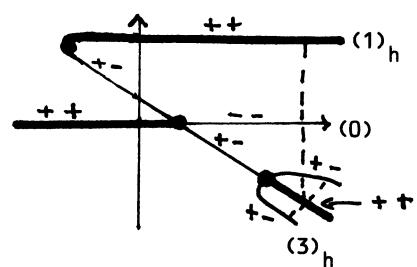
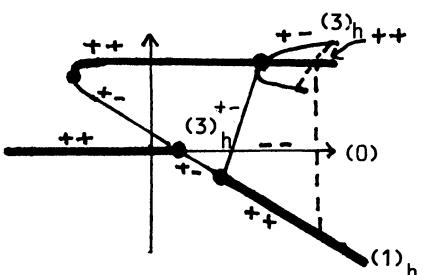
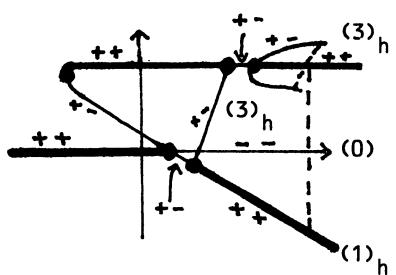
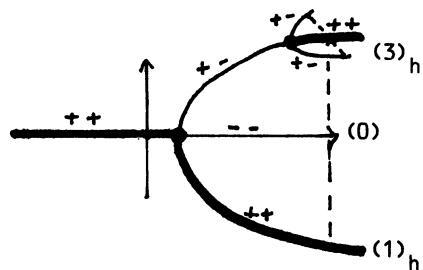
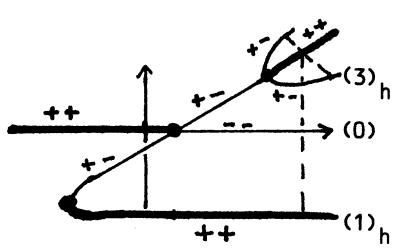
where $\lim_{h \rightarrow 0} a_h = a$, $\lim_{h \rightarrow 0} b_{0h} = \lim_{h \rightarrow 0} b_{1h} = 0$.

Therefore, if (3.8) holds, equations (3) and $(3)_h$ have solutions in a neighbourhood of $(0,0,0)$. The solutions of $(3)_h$ converge to those of (3); but the error estimate between the solutions of (3) and those of $(3)_h$ is no longer as good as estimates (3.5) (or (3.6)). In fact, it is a $O(|p-p_h|^{\frac{2}{3}} + |q-q_h|^{\frac{2}{3}})$.

Remark 3.2 : The intermediate case $B_0 = 0$, $A_1 \neq 0$, $A_2 \neq 0$, $A_1 B_2 - B_1 A_2 \neq 0$ has been studied in [10] and [4].

We obtain the qualitative behaviour of the solutions of $G = 0$ and $G_h = 0$ by considering the behaviour of the zeroes of (3.9) and (3.10). We draw below the solutions of $G = 0$ and $G_h = 0$ in the case D_3 , with $k_1 = -1$, $k_2 = 1$, $\varepsilon_2 = -1$. As in section 3.3, we only give one orbit representative of each branch.





REFERENCES

- [1] ANTMAN, S.S., KENNEY, C.S., Large buckled states of nonlinearly elastic rods under torsion, thrust and gravity, *Arch. Rat. Mech. Anal.*, 76 (1981), 339-354.
- [2] BREZZI, F., FUJII, H., Numerical imperfections and perturbations in the approximation of nonlinear problems, in "The Mathematics of Finite Elements and Applications IV", MAFELAP 1981, Academic Press, New York (1982), 431-452.
- [3] BUZANO, E., GEYMONAT, G., POSTON, T., Post-buckling behaviour of a nonlinearly hyperelastic thin rod with cross-section invariant under the dihedral group D_n , to appear.
- [4] GEYMONAT, G., RAUGEL, G., Finite dimensional approximation of some bifurcation problems arising in the buckling of rods with symmetries, to appear.
- [5] GEYMONAT, G., RAUGEL, G., Imperfection theory in numerical approximation of bifurcation problems, to appear.
- [6] GOLUBITSKY, M., KEYFITZ, B.L., SCHAEFFER, D., A singularity theory analysis of a thermal-chainbranching model for the explosion peninsula, *Comm. Pure Appl. Math.*, 34 (1981), 433-463.
- [7] GOLUBITSKY, M., SCHAEFFER, D., A theory for imperfect bifurcation via singularity theory, *Comm. Pure Appl. Math.*, 32 (1979), 21-98.
- [8] GOLUBITSKY, M., SCHAEFFER, D., Imperfect bifurcation in the presence of symmetry, *Comm. Math. Phys.*, 67 (1979), 205-232.
- [9] GOLUBITSKY, M., SCHAEFFER, D., Boundary conditions and mode jumping in the buckling of a rectangular plate, *Comm. Math. Phys.*, 69 (1979), 209-236.
- [10] GOLUBITSKY, M., SCHAEFFER, D., Bifurcation with $O(3)$ symmetry including applications to the Bénard problem, *Comm. Pure Appl. Math.*, 35 (1982), 81-111.
- [11] MAGNUS, R., A splitting lemma for non-reflexive Banach spaces, *Math. Scan.*, 46 (1980), 118-128.
- [12] RAPPAZ, J., RAUGEL, G., Finite-dimensional approximation of bifurcation problems at a multiple eigenvalue, *Rapport Interne n° 71* (1981), Centre de Mathématiques Appliquées, Ecole Polytechnique, Palaiseau, France.
- [13] RAUGEL, G., Finite dimensional approximation of bifurcation problems in presence of symmetries, *Rapport Interne n° 81* (1982), Centre de Mathématiques Appliquées, Ecole Polytechnique, Palaiseau, France.
- [14] RAUGEL, G., Thèse d'Etat, Université de Rennes-Beaulieu, Rennes, France (in preparation).

Giuseppe GEYMONAT
 Dipartimento di Matematica
 Politecnico di Torino
 Corso Duca Degli Abruzzi 24
 10129 TORINO - ITALIA

Geneviève RAUGEL
 Laboratoire d'Analyse Numérique
 Université de Rennes (IRMAR)
 Rennes-Beaulieu
 35042 RENNES CEDEX - FRANCE

COMPUTATION OF GENERALIZED TURNING POINTS AND
TWO-POINT BOUNDARY VALUE PROBLEMS

A. Griewank
G. W. Reddien

Department of Mathematics
Southern Methodist University
Dallas, Texas 75275

1. Introduction

The accurate location of turning and bifurcation points for ordinary differential equations is a problem that has received much attention recently. See, for example, the conference proceedings edited by Mittelmann and Weber [12], the papers [2], [10], [16], and several of the articles in the proceedings of the conference at which this paper was presented. The procedures we will describe here are not curve tracing methods, but rather they take the basic equation and imbed it in a larger system so that the full system is nonsingular and can be solved directly by Newton's method or a quasi-Newton method. The efficient characterization of these points is important, in addition to its own sake, for multiparameter problems where branches of such points are computed. See in addition to articles in these proceedings the paper of Doedel [5].

This paper will survey some recent results of the authors on the computation and characterization of generalized turning points and then will describe several applications of these results in the context of two-point boundary problems. Related results have been obtained by several authors including Abbott [1], Beyn [2], Moore and Spence [13], Poenisch and Schwetlick [14], Jepson and Spence [9] among others. We will also look at discretization error in the context of boundary value problems. Related results here have been obtained by Brezzi, Rappaz and Raviart [3] and also Fink and Rheinboldt [6].

2. Generalized Turning Points

We first define and describe our characterization procedure. We consider

$$F(z, \lambda) = 0 \quad (2.1)$$

where F is always Fredholm and a C^r -mapping, $r \geq 2$, from an open set D in $X \times R$ into Y where X and Y are Banach spaces. In the finite dimensional case X would be R^n and Y would be R^m with $m \leq n$. We assume the codimension of $R(F_z)$ is no more than one on D and that F_λ is not in $R(F_z)$ when $\text{codim } R(F_z) = 1$. We refer to $M = F^{-1}(0)$ as the solution manifold and $S = \{(z, \lambda) \in D : \text{codim } R(F_z) = 1\}$ as the singular manifold. In the finite dimensional setting, it follows under appropriate regularity assumptions [7] that M is a $p = n+1-m$ dimensional manifold and S is a $n+1-p$ dimensional manifold. The general problem we have considered is the efficient characterization of S and then the numerical computation of intersections of S and M . We call these intersections generalized turning points. At such a point F_λ will not be in the range of F_z . Thus the tangent space to M will not have a λ -component and so can be considered orthogonal to the λ -axis.

In the simple turning point case [13], one has $p = 1$. However, other problems can be unfolded and put into our framework. For example, the regular singular point problem $F(z) = 0$ where F has a one-dimensional null space at the solution z_0 should be written in the form $F(z) + \lambda r = 0$ where r is not in $R(F_z(z_0))$. The problem of determining z_0 becomes that of finding a generalized turning point for the unfolded problem. This unfolding was treated in [18]. Simple bifurcation problems have F_z with a one-dimensional null space but F_λ in the range of F_z at the bifurcation point. Unfolding these problems as $F(z, \lambda) + \gamma r = 0$ where r is not in the range of the total derivative (F_z, F_λ) at the bifurcation point converts them to that of determining a generalized turning point for the unfolded problems. In this case $p = 2$. Perturbed bifurcation problems as considered in [10] have the form $F(z, \lambda, t) = 0$ and so generalized turning points without any modification. As in the simple bifurcation case, $p = 2$.

Optimization problems can be considered also as generalized turning point problems. For the problem $\min \phi(z)$ subject to the constraints $c(z) = 0$, it is classical that at a stationary point there exists a vector of Lagrange multipliers u such that $\phi_z + u^T c_z = 0$. Thus, if one defines $F(z, \lambda) = (c(z), \phi(z) - \lambda)^T$, $F(z, \lambda)$ will be rank deficient at a stationary point for ϕ and so such a point will be a generalized turning point for $F(z, \lambda)$.

We next describe our procedure for characterizing the singular manifold S and then computing the intersection of S and M . First choose T^*

to consist of p -linear functionals in X^* and r in Y so that the matrix operator

$$A = \begin{pmatrix} F_z & r \\ T^* & 0 \end{pmatrix}$$

is nonsingular in D as a mapping from $X \times R$ into $Y \times R^p$. In the finite dimensional case, T^* has p rows, r is in R^m and A is an $(n+1) \times (n+1)$ matrix. If r is chosen so that it is not in $R(F_z^0)$ at the generalized turning point (superscript zero denotes evaluation at (z_0, λ_0)) and if T^* is constant and satisfies $T^*(\phi_1, \phi_2, \dots, \phi_p) = I_p$ where $\text{span}\{\phi_1, \dots, \phi_p\} = \text{null}(F_z)$ and I_p is the $p \times p$ identity, then the matrix A can be seen to be nonsingular on a neighborhood of (z_0, λ_0) . We next solve the systems

$$A \begin{pmatrix} v \\ -g \end{pmatrix} = \begin{pmatrix} 0 \\ I_p \end{pmatrix}, \quad (u^*, -g^T)A = (0, 1) \quad (2.2)$$

for V , u and g . Here V will have p -columns and g will be a p -vector. It can be shown that the same g appears in the solution to the two problems and that they are both uniquely solvable if A is invertible.

Lemma. (a) $g = u^* F_z^* V$

$$(b) \quad g' = u^* F_z'^* V + (u^* r' g^T + g^T T^* V),$$

where the prime denotes differentiation in z or λ .

The proof of this lemma is straightforward and can be found in [7]. In computations, we normally choose r and T^* to be constants and so the differentiation formula simplifies to $g_z = u^* F_{zz}^* V$ and $g_\lambda = u^* F_{z\lambda}^* V$. We note also that the manifold defined by solving $g = 0$ is S .

Our determining system for finding generalized turning points consists of augmenting (1.2) with $g = 0$, i.e.,

$$F(z, \lambda) = 0$$

$$g(z, \lambda) = 0.$$

Note that $g = 0$ consists of p -scalar equations.

We make the following remarks regarding our determining equations. First, as noted in the differentiation formulas, T^* and r may be functions of z and λ . Normally, however, we choose them locally, at least, to be constants. The functions u , V and g will be smooth, and will actually have one less order of differentiability than F . One could consider the solution of our determining equations by either Newton's method or a quasi-Newton procedure. In the latter case, it is important that we have augmented (1.2) with as few equations as possible. Assuming the gradient of F can be computed exactly, this reduces the size of the additional gradient that has to be approximated. If Newton's method is used, then the gradient of g will be required. This can be done conveniently using a difference approximation based on the exact formula for the derivative of g . In particular, the derivative of the i^{th} component of g in z can be approximated by

$$g_z^i \approx u^*(F(z + \epsilon v_i, \lambda) - F(z, \lambda)) / \epsilon$$

where v_i is the i^{th} column of V . A similar formula holds for the derivative of g in λ . Thus, for example, in the simple turning point case the evaluation of the complete gradient of g only requires two additional evaluations of the gradient of F . In general, p additional evaluations of the total derivative of F are required.

We note also that it is possible in our framework to obtain information about the structure of level sets in M with respect to λ , i.e., $\{z : F(z_0, \lambda) = 0\}$. At (z_0, λ_0) in $S \cap M$ it follows easily [7] that $H = g_z V = u^* F_{zz} V V$ will be a $p \times p$ symmetric matrix. The solution (z_0, λ_0) will be isolated if and only if H is nonsingular. In the bifurcation case ($p = 2$) it follows, for example, that H indefinite is equivalent to the usual Crandall-Rabinowitz condition [4]. In the optimization case, the function ϕ will have a maximum or a minimum if H is positive or negative definite respectively.

Our framework generalizes and improves much of what has already appeared on the characterization of turning or bifurcation points. First, we have given a single method that includes problems with a null space of dimension more than 2, but a simple rank drop. We do this through the addition of scalar equations which keeps the size of the extended system as small as possible. We point out, however, that although other characterizations of the singular manifold may be vector equations, the linear algebra

required to solve the extended system by Newton's method can in special cases be reduced. See, for example, Moore and Spence [13]. In our case g may be thought of as an approximate singular value with u and V as approximate singular vectors, but g is not a singular value. It is a smooth function of z and λ even in the bifurcation case. However, if r and T are allowed to vary by choosing $T = V$ and $r = u$, then g will be a singular value. These singular values are differentiable only when $p = 1$.

These results can be extended to the case of semi-simple problems. For example, let F map $R^n \times R$ into R^m and suppose that at the bifurcation point, F has a two-dimensional null space and that the codimension of the range is two-dimensional also. Letting $x = (z, \lambda)$, our system to define g will now have the form

$$\begin{aligned} F_x V - R g^T &= 0 \\ T^* V &= I_3 \end{aligned}$$

where V has three columns, R has two columns, g^T is two by three, and T^* has three rows. Then the singular manifold is characterized by six scalar equations. If the system $F(z, \lambda) = 0$ is partially unfolded by adding vectors out of the range of F_x at the bifurcation point, one can then write down the system

$$F(z, \lambda) + \gamma_1 r_1 + \gamma_2 r_2 = 0$$

$$g(z, \lambda) = 0$$

to characterize the generalized turning point which in this case is a possible multiple bifurcation point. This system will be overdetermined, since it consists of $n+6$ equations in $n+3$ unknowns. However, it can be shown under the conditions of McCleod and Sattinger [11] that the system will have full rank at the bifurcation point. It could then be solved numerically by a Gauss-Newton method. In order to have a square system, one needs to have three more unfolding or control parameters in the sense of Beyn [2] in F . One could, however, simply drop three equations from F or g . This would correspond to an unfolding, although it is difficult to prove that it is sufficient.

The basic framework that we have given in this section was described in the finite dimensional case in [7]. Extensions of this general

framework to a Banach space setting will appear elsewhere. See also [8].

3. Two-Point Boundary Value Problems

We consider the application of the results in the previous section to the case of two-point boundary value problems having the form

$$F(z, \lambda) = \begin{cases} z'' + f(z, \lambda) = 0 & 0 < x < 1 \\ z(0) = z(1) = 0 \end{cases} \quad (3.1)$$

This simple problem will be sufficient to illustrate the ideas. The theory we have developed applies to general first order vector systems with multi-point boundary conditions. We will also limit ourselves in this section to simple turning point problems.

In the setting of (3.1), the equations required to find g have the form

$$\begin{aligned} v'' + f(z, \lambda)v - gr &= 0, & v(0) &= v(1) = 1 \\ v(\frac{1}{2}) &= 1 \end{aligned} \quad (3.2)$$

where the linear functional $T^*v = v(\frac{1}{2})$ provides the normalization condition. Our system then to determine a simple turning point for (2.1) takes the form as before

$$F(z, \lambda) = 0$$

$$g(z, \lambda) = 0 .$$

Define the operator L by $Lv = v'' + f(z, \lambda)v$. The system required to find the analogue to u in (2.2) is given by

$$L^*u^* - gT^* = 0$$

$$u^*r = 1$$

where now L^* represents the operator adjoint of L . One could consider L as a mapping from $C^2[0,1] \cap \{u : u(0) = u(1) = 0\}$ into $C[0,1]$ with the usual norms. Thus u^* in this case is a linear functional.

We carry out the solution of the above equations after they have been discretized. We consider here discretization through the use of projection methods [15]. Let $\{\phi_i\}$ be a set of C^2 -functions satisfying the boundary conditions, let $\{\lambda_i\}$ be a set of continuous linear functionals on $C[0,1]$, and let the set of functions $\{\psi_i\}$ be defined by $\psi_i'' = \phi_i$. A popular

choice for the functions $\{\phi_i\}$ is piecewise polynomial splines. We assume the matrix $\{\lambda_i(\psi_i)\}$ is nonsingular, and then define the projection operator P_n by $P_n f = s$ if and only if $s = \sum \alpha_j \psi_j$ and $\lambda_i(s) = \lambda_i(f)$, $i = 1, \dots, n$. It follows that P_n is a linear projection, i.e., $P_n^2 = P_n$. If one chooses the linear functionals to be point interpolatory, then the projection approximation to the equation $v'' + qv = f$ would take the form $v_n'' + P_n q v_n = P_n f$ where $v_n = P_n v$ and would be collocation. Writing $v_n = \sum \alpha_i \phi_i$, the projection method discretization can be expressed as a matrix problem in the coefficients of the basis representation as

$$(\phi_j(x_i) + q_i \phi_j(x_i))\{\alpha_j\} = \{f(x_i)\}$$

if the point functionals are given by $\lambda_i f = f(x_i)$.

We approximate g and v by solving the system

$$L_n v_n - g_n P_n r = 0$$

$$v_n(\zeta) = 1$$

where $L_n = P_n L$. The normalization condition given is only an example; others are of course possible. In order to carry out the solution of the system $P_n F(z, \lambda) = 0$ augmented with the equation $g_n^* = 0$ by Newton's method, it is necessary to find the analogue to u^* in this finite dimensional setting. If a quasi-Newton procedure is used to approximate g_z^* , then u^* is not needed. Formally, we have to solve the system

$$L_n^* u_n^* - g_n^* P_n^* r = 0$$

$$u_n^*(P_n r) = 1.$$

As before, it will be the case that $g_n = u_n^* L_n v_n$ and $g_{n,z} = u_n^* L_{zz} v_n$ where we are using the natural duality product on the space. We next show how one can compute u_n^* . Write $v_n = \sum \alpha_i \phi_i$ and express the equations for v_n and g_n as

$$A \begin{pmatrix} \{\alpha_i\} \\ -g_n \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

i.e., detach the coefficients. Next consider the solution to the matrix problem

$$A^T \begin{pmatrix} \{\beta_i\} \\ -g_n \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Then we have the following result.

Lemma. $P_n^{*} u_n^* = \sum \beta_i \lambda_i$.

For the purposes of computation, this will be sufficient for the evaluation of g_n . Indeed, using duality product notation one sees that

$$g_n = \langle u_n^*, L_n v_n \rangle = \langle u_n^*, P_n L v_n \rangle = \langle P_n^{*} u_n^*, L_n v_n \rangle$$

Using collocation, g_n is then simply a weighted residual for $P_n L v_n$ at the collocation points. Using the general theory of projection methods, it is possible to establish error estimates. We give a theorem here for the simple turning point case only.

Theorem 3.1. Let $P_n \rightarrow I$ on $C[0,1]$. Let (z_0, λ_0) be a simple turning point [13] for (3.1), and let v_0 solve (3.2) at (z_0, λ_0) and $g(z_0, \lambda_0) = 0$. Then there exists an integer N and a neighborhood Ω of (z_0, λ_0) in $C^2 \times R$ such that solutions (z_n, λ_n) to $P_n F = 0$, $g_n = 0$ exist and are unique in Ω for all $n \geq N$. Moreover, there exists a constant k such that

$$\| (z_n, \lambda_n) - (z_0, \lambda_0) \| \leq k (\| P_n z_0'' - z_0'' \| + \| P_n v_0'' - v_0'' \|), \quad (3.3)$$

where the norm on the left in (3.3) is taken in $C^2 \times R$ and on the right in C .

We remark that results of this type in the simple turning point case are obtainable by other means. See, for example, the papers of Fink and Rheinboldt [6] and Brezzi, Rappaz and Raviart [3].

Two-point boundary value problems can also be discretized through multiple shooting, and we want to show what our characterization equations generate in that case. We consider a more general first order vector system with y in R^n given by

$$\begin{aligned} dy/dx + f(x, y, \lambda) &= 0, \quad a \leq x \leq b, \\ B_a y(a) + B_b y(b) &= 0. \end{aligned}$$

Discretizing this problem by using multiple shooting over m intervals, we solve the problems

$$\begin{aligned} y_j + f(x, y_j, \lambda) &= 0, & x_j \leq x \leq x_{j+1} \\ y_j(x_j) &= s_j & j = 0, 1, \dots, m-1. \end{aligned} \quad (3.4)$$

Now define $F_j = y(x_{j+1}, s_j, \lambda) - s_{j+1}$, $j = 0, \dots, m-2$ and $F_{m-1} = B_a s_0 + B_b y_{m-1}(x_m, s_{m-1}, \lambda)$. Then the boundary value problem will be equivalent to the system

$$F(s, \lambda) = 0, \quad F = \{F_j\}, \quad s = \{s_j\}.$$

We next interpret v , g and u in this setting. Define the fundamental matrices Y_j by

$$Y_j' + f_y Y_j = 0, \quad x_j \leq x \leq x_{j+1}, \quad Y_j(x_j) = I, \quad j = 0, \dots, m-1.$$

Then it follows directly that the system for defining v and g in the simple turning point case has the form

$$\left| \begin{array}{cc|c} Y_0(x_i) & -I & 0 \\ & \cdot & \cdot \\ & -I & 0 \\ \hline B_a & B_b Y_{m-1}(x_m) & r \\ t_1^T & 0 & 0 \end{array} \right| \begin{pmatrix} v \\ -g \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

where we have selected r so that it only results in the relaxation of the boundary conditions and t_1^* so that it gives a normalization through the addition of an additional boundary condition. In component form, the equations defining v and g become

$$(a) \quad Y_i(x_{i+1})v_i - v_{i+1} = 0, \quad i = 0, \dots, m-2$$

$$(b) \quad B_a v_0 + B_b Y_{m-1}(x_m)v_{m-1} = rg$$

$$(c) \quad t_1^T v_0 = 1.$$

Define $v(x) = Y_i(x)v_i$ on $[x_i, x_{i+1}]$. Then it follows that v solves $v' + f_x(x, y, \lambda)v = 0$ plus the boundary conditions. We next consider the equations in this case that define u , namely

$$\begin{aligned} F_y^T u - g^T &= 0 \\ u^T r &= 1. \end{aligned}$$

In component form these equations are

$$\begin{aligned} u_0^T Y_0(x_1) + u_{m-1}^T B_a - g t_1^T &= 0 \\ -u_i^T + u_{i+1}^T Y_{i+1}(x_{i+2}) &= 0, \quad i = 0, \dots, m-3, \\ -u_{m-2}^T + u_{m-1}^T B_b Y_{m-1}(x_m) &= 0 \\ u_{m-1}^T r &= 1. \end{aligned}$$

Let Z_i be the fundamental matrices for the adjoint problems for (3.4), i.e.,

$$-dZ_i/dx + f_y^T Z_i = 0, \quad x_i \leq x \leq x_{i+1},$$

$$Z_i(x_{i+1}) = I.$$

It is known that $Z_i(x_i) = Y^T(x_{i+1})$. If we define $\psi(x) = Z_i(x)u_i$ on $[x_i, x_{i+1}]$, $i = 0, \dots, m-2$, and $\psi(x) = Z_{m-1}(x)B_b^T u_{m-1}$ on $[x_{m-1}, x_m]$, then it follows that

$$\begin{aligned} -\psi' + f^T \psi &= 0 \\ B_a^* \psi(a) + B_b^* \psi(x_m) &= g B_a^* t_1 \end{aligned}$$

and

$$g = u_{m-1}^T (B_a v(a) + B_b v(b)),$$

where B_a^* , B_b^* give the usual adjoint boundary conditions.

The linear algebra required to solve the equations characterizing the turning point is simple since one has a bordered block bi-diagonal system of order $(mn+1) \times (mn+1)$. The computational effort is spent in finding the fundamental matrices, but this can be done accurately and so highly accurate results can be obtained for approximations to the turning points.

4. Non-Simple Turning Points.

We next describe the extension of our procedure to the determination of cusp points and the following of fold curves. This terminology is

explained in [17]. Consider once again our determining system

$$\begin{aligned} F(z, \lambda) &= 0 \\ g(z, \lambda) &= 0 \end{aligned} \tag{4.1}$$

where g has only a single component. Let (z_0, λ_0) be a turning point for F with ϕ_0 a basis vector for the one-dimensional null space of F_z^0 and ψ_0 a basis for the one-dimensional null space of $(F_z^0)^*$. Then we have the following result.

Theorem 4.1. The turning point (z_0, λ_0) is an isolated solution of (4.1) if and only if $\psi_0^{*F_z^0}\phi_0\phi_0^* \neq 0$.

Similar results to this and others in this section for other determining systems have been obtained in [9] and [17]. Given a two-parameter problem, $F(z, \lambda, \mu) = 0$, define

$$\tilde{F}(x, \mu) = \begin{cases} F(z, \lambda, \mu) \\ g(z, \lambda, \mu) \end{cases}$$

where $x = (z, \lambda)$. Now \tilde{F} will be nonsingular from the previous theorem at simple turning points with μ fixed, and so such turning points can be continued in μ . We also have the next result with F in C^3 .

Theorem 4.2. Let $\psi_0^{*F_z^0}\phi_0\phi_0^* = 0$ and $\psi_0^{*(3F_{zz}^0\phi_0v_0 + F_{zzz}^0\phi_0^3)} \neq 0$ where $F_z^0v_0 = -F_{zz}^0\phi_0\phi_0^*$, $T^*\phi_0 = 0$. Let $\eta_0^{*F_\lambda^0} = -\psi_0^{*F_{z\lambda}^0}\phi_0$, $\eta_0^{*F_z^0} = -\psi_0^{*F_{zz}^0}\phi_0$ and assume $\eta_0^{*F_z^0} = -\eta_0^{*F_\mu^0} + \phi_0^{*F_{z\mu}^0}\phi_0^* \neq 0$. Then a double turning point (z_0, λ_0, μ_0) of F with respect to (z, λ) and $\mu = \mu_0$ fixed corresponds to a simple turning point (z_0, λ_0, μ_0) of \tilde{F} with respect to (x, μ) .

Thus one can repeat the process defining a \tilde{g} to correspond to \tilde{F} in the usual way. Then one obtains that the system

$$\begin{aligned} F(z, \lambda, \mu) &= 0 \\ g(z, \lambda, \mu) &= 0 \\ \tilde{g}(z, \lambda, \mu) &= 0 \end{aligned} \tag{4.2}$$

will be a nonsingular characterization of a cusp point for F . Cusp points are characterized by the hypotheses of Theorem 4.2 [17]. We next consider what is required to carry out the solution of system (4.2) using Newton's method.

The basic derivatives of F are easy to compute if F is the result of the discretization of (3.1) by finite differences or projection methods using splines. For example, the usual central difference discretization of (3.1) in the two-parameter case with uniform mesh $x_i = ih$, $h = 1/n$, produces

$$(z_i - 2z_{i+1} + z_{i+2}) + h^2 f(z_i, \lambda, \mu) = 0$$

as the basic set of equations. Writing these equations in the form $F(z, \lambda, \mu) = 0$ where $z = (z_1, \dots, z_{n-1})$ in \mathbb{R}^{n-1} , it follows that F_z is the tri-diagonal matrix $(0, \dots, 0, 1, -2 + h^2 f_z(z_i, \lambda, \mu), 1, 0, \dots, 0)$. Letting $v = (v_1, \dots, v_{n-1})$, it follows that $F_{zz}v$ is the diagonal matrix $\text{diag}(h^2 f_{zz}(z_i, \lambda, \mu)v_i)$, and also that $F_{z\lambda}$ is the diagonal matrix $\text{diag}(h^2 f_{z\lambda}(z_i, \lambda, \mu))$. Other derivatives of F are equally easy to compute. Let $\tilde{v} = (w, \rho)$ solve

$$\begin{pmatrix} F_z & F_\lambda \\ g_z & g_\lambda \end{pmatrix} \begin{pmatrix} w \\ \rho \end{pmatrix} - gr_1 = 0, \quad T_1^* \begin{pmatrix} w \\ \rho \end{pmatrix} = 1.$$

Analogously define $\tilde{u}^* = (u_1^*, \delta)$. Then

$$\begin{aligned} \tilde{g}_z &= u_1^*(F_{zz}w + F_{x\lambda}\rho) + \delta(g_{zz}w + g_{z\lambda}\rho) \\ \tilde{g}_\lambda &= u_1^*(F_{z\lambda}w + F_{\lambda\lambda}\rho) + \delta(g_{z\lambda}w + g_{\lambda\lambda}\rho) \\ \tilde{g}_\mu &= u_1^*(F_{z\mu}w + F_{\lambda\mu}\rho) + \delta(g_{z\mu}w + g_{\lambda\mu}\rho). \end{aligned} \tag{4.3}$$

Now one can derive the following formulas:

$$(a) \quad g_{zz}w = u^* F_{zzz}vw + 2u^* F_{zz}v_z w$$

where

$$F_z v_z w - r g_z w = -F_{zz} v w \tag{4.4}$$

$$T^*(g_z w) = 0$$

$$(b) \quad g_{z,\lambda}\rho \approx (g_z(z, \lambda + \epsilon\rho) - g_z(z, \lambda))/\epsilon$$

$$(c) \quad g_{z,\lambda}w = u^* F_{zz\lambda}vw + u^* F_{zz}v_\lambda w + u^* F_{z\lambda}v_z w$$

where

$$F_z r_\lambda - g_\lambda r = - F_{z\lambda} v$$

$$T^* v_\lambda = 0.$$

The formulas (4.4) are evaluated using the basic system of linear equations that are used to define v and g . Then they can be substituted into (4.3). We recommend using the finite difference formula (4.4)(b) to approximate $g_{z,\lambda}^0$. This completes the evaluation of the derivatives required to carry out the solution of (4.2) using Newton's method.

5. Numerical Experiments

We report in this section on the results of a few of the numerical experiments we have performed to illustrate some of the results given here.

With an exact Jacobian available, which is generally the case for discretizing boundary value problems, one could use a quasi-Newton method (Broyden update) to maintain an approximation to the Jacobian of g . Superlinear convergence (one-step) can be shown. For the optimization problem, this places our procedure intermediate to those which update the full Hessian of ϕ or the projected Hessian H . In our case, u may be considered an approximate Lagrange multiplier and z is not constrained to lie on the manifold. The numerical work involved in carrying out a step amounts to solving $p+3$ linear systems with our basic matrix

$$\begin{pmatrix} F_z & r \\ T^* & 0 \end{pmatrix}.$$

As an example, we solved the two-dimensional problem $\Delta u + \lambda e^u = 0$ over the unit rectangle with zero boundary conditions. The discretization used was the usual central difference approximation to the Laplacian. The quasi-Newton method exhibited superlinear convergence, with the number of steps taken being independent of the mesh size. For example, after nine iterations the relative reduction in the value of g was 10-13, 10-9, and 10-13 for meshes with spacings $h = 1/8, 1/16$, and $1/32$, respectively.

We solved the one-dimensional version of the preceding problem also using central differences. Recall that the equations defining v and g in this case take the form

$$v'' + \lambda e^u v - gr = 0, \quad v(0) = v(1) = 0,$$

$$v(\frac{1}{2}) = 1.$$

We experimented with other normalizations such as $\int_0^1 v(x)dx = 1$ and noticed no significant difference in the results. We observed that the error $\lambda - \lambda_h$ was .01874, .00457 and .00114 for meshes with spacings $h = 1/10, 1/20$, and $1/40$, respectively. These numbers are $O(h^2)$ as is predicted by Theorem 3.1.

We also solved the same problem using collocation at Gauss points with quintic splines as the discretization. The value of λ at the first simple turning point is approximately 3.513830719. We computed values of 3.513706567, 3.513830590 and 3.513830719 with meshes of size $h = 1/2, 1/4$, and $1/8$. This is a highly accurate method with the basic discretization having order $O(h^6)$ for its accuracy.

In order to illustrate the results on shooting methods, we consider the bifurcation problem (off the trivial branch)

$$u'_1 = -\lambda u_2$$

$$u'_2 = \sin u_1 \quad u_2(0) = u_2(1) = u_3(0) = 0$$

$$u'_3 = \cos u_1 - 1.$$

Using simple shooting, one can simply set $u_1(0) = s$ and reduce the differential equation to the single scalar equation

$$u_2(1, s, \lambda) = 0.$$

Now choosing $r = 1$ and $T = I_2$ ($p = 2$), the basic system used to determine V and g becomes

$$\begin{pmatrix} \frac{\partial u_2}{\partial s} & \frac{\partial u_2}{\partial \lambda} & -1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} v_{11} & v_{21} \\ v_{12} & v_{22} \\ g_1 & g_2 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

where $g_1 = \partial u_2 / \partial s$ and $g_2 = \partial u_2 / \partial \lambda$. Thus our system for characterizing the bifurcation point in this case simply becomes the assertion that $u_2(1, s, \lambda)$ is stationary. With $h = .05$ and using a fourth order Runge-Kutta method

from a starting point of $s = 1$ and $\lambda = 8$, the first bifurcation point π^2 was computed to seven digits in five steps. The results of further experiments will be reported elsewhere.

References

- [1] Abbott, J. P., An efficient algorithm for the determination of certain bifurcation points, *J. Comp. Appl. Math.* 4 (1978), 19-27.
- [2] Beyn, W. J., Numerical analysis of singularities in a diffusion reaction model, to appear.
- [3] Brezzi, F., Rappaz, T. and Raviart, P., Finite dimensional approximation of nonlinear problems.
Part I. Branches of nonsingular solutions, *Numer. Math.* 36 (1980), 1-25.
Part II. Limit points, *Numer. Math.* 37 (1981), 1-28.
Part III. Simple bifurcation points, *Numer. Math.* 38 (1981), 1-30.
- [4] Crandall, M. G. and Rabinowitz, P. H., Bifurcation from simple eigenvalues, *J. Functional Anal.* 9 (1971), 321-340.
- [5] Doedel, E. J., Auto: A program for the automatic bifurcation analysis of autonomous systems, *Congressus Numerantium* 30 (1981), 265-284.
- [6] Fink, J. P. and Rheinboldt, W. C., On the discretization error of parameterized nonlinear equations, Univ. of Pittsburgh, Inst. for Comp. Math. and Applications, Tech. Report ICMA-83-59, June, 1983.
- [7] Griewank, A. and Reddien, G. W., Characterization and computation of generalized turning points, *SIAM J. Numer. Anal.*, to appear.
- [8] Griewank, A. and Reddien, G. W., Computation of turning and bifurcation points for two-point boundary value problems, in *Proceedings of Seventh Annual Lecture Series in the Mathematical Sciences*, University of Arkansas, Fayetteville, Arkansas (1983).
- [9] Jepson, A. and Spence, A., Folds in solutions of two-parameter systems:
Part I, Tech. Report NA-82-02, Computer Science Dept., Stanford University, (1982).
- [10] Keener, J. P. and Keller, H. B., Perturbed bifurcation theory, *Arch. Rat. Mech. Anal.* 50 (1973), 159-175.
- [11] McCleod, J. B. and Sattinger, D. H., Loss of stability and bifurcation at a double eigenvalue, *Journal of Functional Analysis* 14 (1973), 62-84.
- [12] Mittelman, H. D. and Weber, H., *Bifurcation Problems and Their Numerical Solution*, ISNM 54, Birkhauser, Basel (1980).

- [13] Moore, G. and Spence, A., The calculations of turning points of nonlinear equations, SIAM J. Numer. Anal. 17 (1980), 567-576.
- [14] Pönisch, G. and Schwetlick, M., Computing turning points of curves implicitly defined by nonlinear equations depending on a parameter, Computing 26 (1981), 107-121.
- [15] Reddien, G. W., Projection methods for two-point boundary value problems, SIAM Review 22 (1980), 156-171.
- [16] Seydel, R., Numerical computation of branch points in ordinary differential equations, Numer. Math. 32 (1979), 51-68.
- [17] Spence, A. and Werner, B., Non-simple turning points and cusps, IMA Journal of Numerical Analysis 2 (1982), 413-427.
- [18] Weber, H. and Werner, W., On the accurate determination of non-isolated solutions of nonlinear equations, Computing 26 (1981), 315-326.

On Some Methods for the Computational Analysis of Manifolds¹⁾

by

Werner C. Rheinboldt²⁾

1. Introduction

For nearly a century now the concept of a manifold has played a fundamental role throughout mathematics. In the global view of modern mechanics, pioneered by H. Poincaré, the phase-space of a dynamical system constitutes a differentiable manifold. Accordingly, manifolds have become essential tools wherever a global study of nonlinear phenomena is undertaken. A list of such areas would be extensive. It would include, for example, bifurcation theory and the study of stability and chaos in dynamical systems, gravitational studies and other work involving modern field theories in physics, as well as many other problems concerning nonlinear operator equations.

In line with this it is hardly surprising that recent years have brought increasing interest in the development of numerical methods for the computational analysis of manifolds. Among the principal aims of such methods is the computational approximation of specific paths on a given manifold and the detection of particular features along such paths. In essence, this emphasis on path-tracing derives once again from Poincaré's global view of mechanics in which the qualitative theory of dynamical systems is based on the properties of the phase portrait; that is, the family of solution curves which fill up the entire phase space, (see eg. [1]).

For different problems the paths to be computed on a manifold are specified in different ways. We consider here two types of such path specifi-

-
- 1) This work was in part supported by the National Science Foundation under Grant MCS-78-05299.
 - 2) Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, PA 15261.

cations arising frequently in applications. The first one relates to the mentioned study of phase portraits and concerns the case when the paths constitute the solution curves of a vector field on the manifold. A class of problems of this kind consists of systems formed by a mixture of differential equations and algebraic equations.

The second type of path specification arises in the study of equilibrium problems subjected to certain parameter changes. Usually this involves a parameterized nonlinear operator equation where the set of all its regular solutions -- in the space of the combined state and parameter variables -- constitutes our manifold. In practice the typical approach is to study the variation of these solutions under parameter changes with one degree of freedom. It turns out that any such one-dimensional parameter combination defines a so-called 1-distribution on the manifold. The numerical task then involves the computation of the one-dimensional integral manifolds of such a 1-distribution. Any segment of one of them, together with an appropriate choice of coordinate, once again represents a path on our manifold.

The aim of this paper is to develop some of the theoretical results underlying these two types of path-specifications, as far as they are relevant to the numerical methods, and to point out the principal differences and similarities. More specifically, in Section 2 we recall some basic relations between nonlinear equations and manifolds. Section 3 then turns to the case of vector fields on manifolds and gives an overview of recent existence and uniqueness results for differential/algebraic equations. In Section 4 we consider the mentioned 1-distributions on solution manifolds of parameterized equations and the connections with augmented equations. Finally, Section 5 addresses certain computational aspects and points to some new approaches for solving differential/algebraic equations.

2. Nonlinear Equations and Manifolds

Many of the manifolds occurring in computational applications are defined (at least locally) as solution sets of nonlinear equations. In the simplest, finite-dimensional case, these are equations of the form

$$F(z, \lambda) = y_0 \quad (2.1)$$

where $F: S \subset R^n \times R^m \rightarrow R^n$ is a C^r -map ($r \geq 1$) on an open set S of $R^n \times R^m$ and $y_0 \in R^n$ a given vector. Here $z \in R^n$ signifies a state variable and $\lambda \in R^m$, $m \geq 1$, is a parameter-vector. For ease of notation, we shall often write x instead of (z, λ) .

A point $x \in S$ is regular if the derivative $DF(x)$ has (maximal) rank n . Let $R(F, S) \subset S$ denote the set of all regular points of F in S . Then, for any $y_0 \in F(S)$ the regular solution set

$$M = \{x \in R(F, S); F(x) = y_0\} \quad (2.2)$$

is an m -dimensional C^r -manifold in R^{n+m} without boundary. The proof is straightforward and, in fact, a local form of this result is often used as the basis of the definition of sub-manifolds of R^d (see eg. [8]).

As a simple example we consider the spring system of Figure 1 discussed in [19]. After a suitable scaling, the equilibrium equation is

$$F(x) = \begin{pmatrix} -2(1-p) + 2\lambda \cos q + v \sin q \\ 4\gamma q - 2\lambda p \sin q + vp \cos q \end{pmatrix} = 0 \quad (2.3)$$

where p, q are the state variables and the parameters λ, v, γ represent

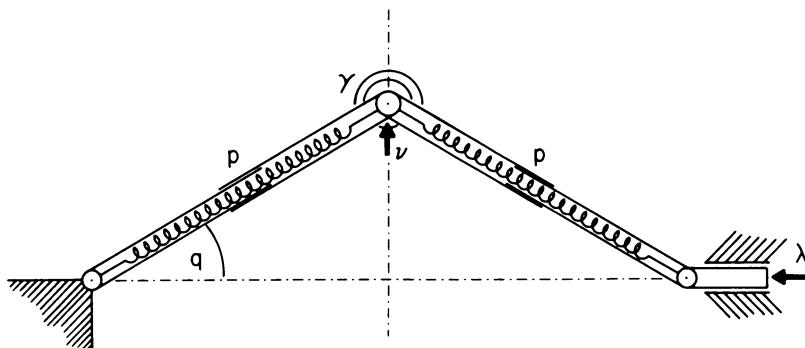


Figure 1

applied forces and a spring constant, respectively. The spring constant γ is intrinsically positive and the mapping $F: S \subset \mathbb{R}^5 \rightarrow \mathbb{R}^2$ is of class C^∞ on $S = \{x = (p, q, \lambda, v, \gamma)^T \in \mathbb{R}^5; \gamma > 0\}$. It is easily verified that all points of S are regular for F and that $0 \in F(S)$. Hence, the equilibria of the system correspond to points on the three-dimensional manifold $M = \{x \in S; F(x) = 0\}$.

As long as we restrict attention to mappings F between finite-dimensional spaces, the null-space, $\ker DF(x)$, of the derivative $DF(x)$ has dimension m for any $x \in R(F, S)$. In fact, for any $n \times (n+m)$ matrix A with adjoint $A^* = A^T$ the index

$$\text{ind } A = \dim \ker A - \dim \ker A^* \quad (2.4)$$

always equals m , and for $x \in R(F, S)$ and $A = DF(x)$ the dimension of the co-kernel, $\ker A^*$, is zero.

It is this property which does not carry over to a mapping F between Banach spaces X and Y since then the index (2.4) is not necessarily finite for all bounded linear operators $A \in L(X, Y)$. The class of operators with finite index includes the well-known Fredholm operators. Recall (see eg. [24]) that these are the operators $A \in L(X, Y)$ for which both $\dim \ker A$ and $\dim \ker A^*$ (and hence also the index (2.4)) are finite and the range space AX is closed in Y . A nonlinear mapping $F: S \subset X \rightarrow Y$ of class C^1 on the open set S then is a Fredholm map on S if $DF(x)$ is a Fredholm operator for each $x \in S$. On any connected component of S the index of $DF(x)$ is constant and hence is called the index of F .

For Fredholm mappings the mentioned results about the regularity set and the solution manifolds remain valid also in the infinite dimensional case (see eg. [6]).

Theorem 1: Let $F: S \subset X \rightarrow Y$ be a Fredholm mapping of class C^r , $r \geq 1$, and index $m \geq 1$ from the open set S of the Banach space X into the Banach space Y . Then the regularity set

$$R(F, S) = \{x \in S; DF(x)X = Y\} \quad (2.5)$$

is open in X and for any $y_0 \in F(S)$ the regular solution set M of (2.2) is an m -dimensional C^r -manifold in X without boundary.

Fredholm mappings occur frequently in applications. In fact, many nonlinear Dirichlet problems have this property. A generic example, based on the theory in [2], is outlined in [7].

3. Differential Equations on Manifolds

As a first class of computational problems on a given manifold M , we consider the classical situation when a vector field is prescribed on M . In applications, M usually represents the phase space of a mechanical system and the vector field defines its dynamics. Then the numerical task is to compute a solution path which is tangent at each of its points to the vector given at that point.

As in Section 2, we restrict the attention to the frequent case when the manifold M is the regular solution set of some nonlinear equation. Let the vector field be given by an ordinary differential equation, then, we have here a mixed system of algebraic and differential equations (DAE's). Such systems occur not only in mechanics. For instance, in the simulation of electronic circuits they are often called semistate equations (see eg. [16]) and a different example arises in the modeling of electrophoretic separation processes (see eg. [4]).

A simple example from point-mechanics will illustrate the situation. Suppose that a mass-point is constrained to move on the parabolic surface

$$f(x) \equiv \alpha(x_1^2 + x_2^2) - x_3 = 0, \quad y = (x_1, x_2, x_3)^T \in \mathbb{R}^3, \quad \alpha > 0, \quad (3.1)$$

(see Figure 2). If gravity is the only external force then the equation of motion becomes

$$m \frac{d^2 x}{dt^2} = q - \xi f'(x)^T \quad (3.2)$$

where $q = (0, 0, mg)^T$ and $\xi f'(x)^T$ represents the constraining force orthogonal to the surface. With the dimensionless phase-space variables

$$y_i = \alpha x_i, \quad y_{3+i} = \sqrt{\frac{\alpha}{g}} \dot{x}_i, \quad i = 1, 2, 3, \quad y_7 = \frac{1}{mg} \xi, \quad \tau = \sqrt{g\alpha} t \quad (3.3)$$

the system (3.1), (3.2) assumes the form

$$y_3 - y_1^2 - y_2^2 = 0 \quad (3.4)$$

$$\begin{aligned} \dot{y}_1 &= y_4 & \dot{y}_4 &= 2y_1 y_7 \\ \dot{y}_2 &= y_5 & \dot{y}_5 &= 2y_2 y_7 \\ \dot{y}_3 &= y_6 & \dot{y}_6 &= 1 - y_7 \end{aligned} \quad (3.5)$$

where dots represent derivatives with respect to τ . The first equation specifies a six-dimensional manifold in \mathbb{R}^7 and the differential equations represent the dynamical system on it.

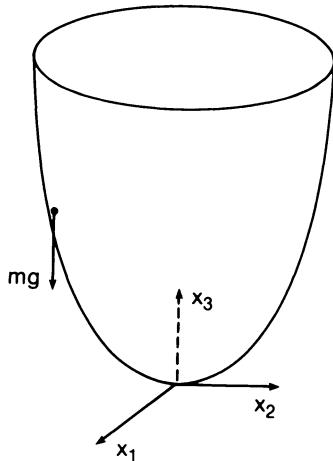


Figure 2

Differential-algebraic systems of this type have many properties in common with systems of ordinary differential equations. But there are also differences especially from a numerical point of view (see eg. [10], [18]). By considering DAE's in the setting used here, namely as differential equations on manifolds, it becomes possible to develop an existence and uniqueness theory which provides insight into the properties of these

problems. Such a theory was presented in [21] and we sketch here only some of the results especially as they are relevant to our example (3.4), (3.5).

We begin by summarizing some basic facts about vector fields and dynamical systems on manifolds (see eg. [14], [26]). Let M be a finite-dimensional Hausdorff manifold of class C^r , $r \geq 2$, modelled on R^m . The tangent space M at $y \in M$ is denoted by $T_y M$ and TM is the tangent bundle. A vector field on M of class C^p , $0 \leq p \leq r$, is a C^p -mapping $v: M \rightarrow TM$ such that $v(y) \in T_y M$ for each $y \in M$. For any $y_0 \in M$ an integral curve of the vector field through y_0 is a mapping $\eta: J \rightarrow M$ of class C^p on some open interval J of R^1 containing 0 such that

$$\eta'(t) = v(\eta(t)), \quad \forall t \in J, \quad \eta(0) = y_0. \quad (3.6)$$

Such integral curves exist locally on M :

Theorem 2: Under the stated conditions about M and v , there exists for any $y_0 \in M$ an integral curve of class C^r through y_0 . Moreover, for any two integral curves $\eta_i: J_i \rightarrow M$, $i = 1, 2$, through y_0 we have $\eta_1(t) = \eta_2(t)$ for all $t \in J_1 \cap J_2$.

This theorem shows that the union of the domains of all integral curves of v through a given point $y \in M$ is an open interval J_y of R^1 . Then the following global result holds:

Theorem 3: Under the cited assumptions about M and v , the set

$$\mathcal{D}(v) = \{(t, y) \in R^1 \times M; \quad t \in J_y\}$$

is open in $R^1 \times M$ and contains $\{0\} \times M$. Moreover, there exists a unique C^p -mapping $\eta^*: \mathcal{D}(v) \rightarrow M$ such that for any $y \in M$ the mapping $\eta_y^*: J_y \rightarrow M$, $\eta_y^*(t) = \eta^*(t, y)$ is an integral curve of v through y .

We apply this result to the case of differential-algebraic equations. For simplicity suppose that we have an autonomous system of the

form

$$F(y) = 0 \quad (3.7)$$

$$A(y) \frac{dy}{dt} = G(y)$$

where the mappings

$$\begin{aligned} F: S \rightarrow R^{m_1}, \quad A: S \rightarrow L(R^n, R^{m_2}), \quad G: S \rightarrow R^{m_2} \\ 1 \leq m_1 < n, \quad m_2 \geq 1, \quad m_1 + m_2 \geq n \end{aligned} \quad (3.8)$$

are of class C^r , $r \geq 2$, on some open set $S \subset R^n$. Thus, the manifold under consideration is here the regular solution manifold of F , namely,

$$M = \{y \in R(F, S); F(y) = 0\}. \quad (3.9)$$

Let $y: J \rightarrow M$ denote a C^1 -solution of (3.7) on some open interval $J \subset R^1$; then for any $t \in J$ the tangent vector $\dot{y}(t) = \frac{d}{dt} y(t)$ must satisfy

$$N(y(t))\dot{y}(t) = \begin{pmatrix} 0 \\ G(y(t)) \end{pmatrix}, \quad N(y) = \begin{pmatrix} DF(y) \\ A(y) \end{pmatrix} \quad (3.10)$$

For the solvability of (3.10) it is necessary that the vector on the right side belongs to the range of $N(y)$. This leads to the following result:

Theorem 4: Let the maps (3.8) be of class C^r , $r \geq 2$, on the open set $S \subset R^n$ and

$$S_0 = \{y \in S; \text{rank } N(y) = n, \begin{pmatrix} 0 \\ G(y) \end{pmatrix} \in \text{rge } N(y)\}. \quad (3.11)$$

If there exists a non-empty set $M_0 \subset M \cap S_0$ which is open in M , then for any $y_0 \in M_0$ there exists on M_0 a unique, maximally extended C^{r-1} -solution of (3.7) through y_0 , and the dependence of the solutions of (3.7) upon their

initial points in M_0 is of class C^{r-1} .

Proof: As an open subset of M the set M_0 constitutes a sub-manifold of the same dimension. Hence, for each $y \in M_0$, the vector $v(y) \in T_y M = T_{y_0} M$ specified by

$$N(y)v(y) = \begin{pmatrix} 0 \\ G(y) \end{pmatrix}, \quad y \in M_0. \quad (3.12)$$

is well-defined and introduces a vector field $v: M_0 \rightarrow TM_0$ on M_0 . From the differentiability assumptions it follows readily that v is of class C^{r-1} on M_0 . Clearly, the integral curves of v on M_0 are exactly the solutions of (3.7) in that set and hence the result follows directly from Theorem 3.

In the special case when the system (3.7) is "square"; that is, when $m_1 + m_2 = n$, then $\text{rank } N(y) = n$ implies the non-singularity of the matrix $N(y)$ of (3.10). Hence the set (3.11) is equivalent with

$$S_0 = \{y_0 \in S; \quad N(y) \text{ non-singular}\} \quad (3.13)$$

and the intersection $M_0 = M \cap S$ is necessarily open in M . Therefore, if M_0 is non-empty then the conclusions of the theorem apply on M_0 . In this case the vector field on M_0 is given by

$$v: M_0 \rightarrow TM_0, \quad v(y) = N(y)^{-1} \begin{pmatrix} 0 \\ G(y) \end{pmatrix}, \quad \forall y \in M_0. \quad (3.14)$$

Suppose that this system is non-autonomous:

$$F(y, t) = 0 \quad (3.15)$$

$$A(y, t) \frac{dy}{dt} = G(y, t),$$

where the mappings (3.8) with $m_1 + m_2 = n$ are of class C^r , $r \geq 2$, on some open subset $S \subset R^n \times R^1$. The standard approach here is to add the differential equation $t' = 1$; that is, to restrict consideration to

solution-paths parametrized by t . With this added equation the problem reduces to an autonomous system and the above results apply. On the other hand, if we allow for solutions $y = y(s)$, $t = t(s)$, $s \in J \subset \mathbb{R}^1$, parametrized in terms of some parameter s , then the tangent vector of this path in \mathbb{R}^{n+1} must satisfy

$$K(y(s), t(s)) \begin{pmatrix} \dot{y}(s) \\ \dot{t}(s) \end{pmatrix} = 0, \quad K(y, t) = \begin{pmatrix} D_y F(y, t) & D_t F(y, t) \\ A(y, t) & -G(y, t) \end{pmatrix} \quad (3.16)$$

where dots represent derivatives with respect to s . If $K(y, t)$ has full rank then its one-dimensional null-space defines a one-dimensional subspace of the tangent manifold at (y, t) . As we shall see later, we encounter here a 1-distribution on the given manifold. In order to define a vector field we need to choose a specific direction and normalization of the null-vector of K . This will be addressed in the next section.

We end this brief sketch of the existence theory for DAE's developed in [21] by commenting on the case when in Theorem 4 the set S_0 of (3.11) is empty. An example of this is our problem (3.4), (3.5). In fact, there the seventh column of $N(y)$ is zero and it is quickly verified that the right side of (3.12) is in the range of $N(y)$ exactly if

$$2y_1 y_4 + 2y_2 y_5 - y_6 = 0. \quad (3.17)$$

The validity of this necessary condition for the solvability of (3.4), (3.5) can also be deduced directly from the equations. Thus, instead of the six-dimensional manifold M_1 in \mathbb{R}^7 defined by the single equation (3.4) we have to restrict consideration to the five-dimensional regular solution manifold M_2 specified by the two equations (3.4) and (3.17). In other words, (3.17) has to be added to our equations. But, once again, the matrix $N(y)$ of the new, expanded system has a vanishing seventh column. Now the right side is in the range of $N(y)$ exactly if

$$2y_4^2 + 2y_5^2 + 4y_3 y_7 + y_7 - 1 = 0. \quad (3.18)$$

Again this equation has to be added to the system and hence we have to

restrict attention to the four-dimensional regular solution manifold M_3 on \mathbb{R}^7 defined by the three equations (3.4), (3.17) and (3.18). A simple calculation shows that now the corresponding matrix $N(y)$ has rank seven on $S_0 = \{y \in \mathbb{R}^7; y_3 > -1/4\}$ and that on all of $M_3 \cap S_0$ the right side is in the range of $N(y)$. Thus, the expanded system has the form (3.7) with $n = 7$, $m_1 = 3$, $m_2 = 6$. Moreover, on $M_0 = M_3 \cap S_0$ Theorem 4 applies to this augmented system and hence shows that the original system (3.4), (3.5) has a unique solution for any starting point $y_0 \in \mathbb{R}^7$ which satisfies all three equations (3.4), (3.17) and (3.18).

This example illustrates the general procedure. In [21] differential-algebraic equations were called algebraically incomplete if existence and uniqueness of solution only apply on lower-dimensional sub-manifolds of the manifold defined by the original algebraic part of the system. In [21] it was also shown that the problems with so-called index larger than one considered in [10] are algebraically incomplete. In [10] these higher index problems were identified as the systems for which standard ODE-solvers are expected to fail.

4. Equilibrium Manifolds.

In the previous section we considered dynamical problems for which the solution paths to be computed were defined by a vector field on the given manifold. In the case of equilibrium problems, such as the spring problem of Section 2, the manifold constitutes the set of all equilibrium configurations of the system under study and no vector field is prescribed on it. Typically, the manifold here is the solution set of a nonlinear, parameter-dependent equation and the problem is to determine numerically how the solutions change when the parameters vary.

Since it is difficult to assess the influence of all parameters at once, the usual approach consists in considering only parameter-combinations with one degree of freedom which define one-dimensional sub-manifolds. The numerical task then is to compute segments of such sub-manifolds which, in essence, represent again paths on the given manifold. In other words, even though the setting differs from that of Section 3, we are led once more to a problem of computing certain paths on a manifold.

From a differential-geometric viewpoint it turns out that instead of a global vector we have here a 1-distribution on the given manifold. As before, let M denote any finite-dimensional Hausdorff-manifold of class C^r , $r \geq 1$, and write again $T_x M$ for the tangent space at $x \in M$ and TM for the tangent bundle of M . A 1-distribution Δ of class C^p , $0 \leq p < r$, on an open subset M_0 of M is defined by the properties

- (i) Δ is a mapping $\Delta: M_0 \rightarrow TM$ such that Δ_x is a one-dimensional subspace of $T_x M$ for each $x \in M_0$;
- (ii) for each $x_0 \in M_0$ there exists an open neighborhood $U \subset M_0$ of x_0 in M and a local vector field $v: U \rightarrow TM$ of class C^p on U such that $\Delta_x = \text{span } v(x)$ for all $x \in U$.

A one-dimensional sub-manifold N of M_0 is an integral-manifold of Δ on M_0 if for any $x \in N$ the tangent space $T_x N$ is equal to Δ_x . For further details see eg. [26].

An example of this concept turns out to be given by (3.16). But instead we consider as a simple illustration the spring example of Section 2. Obvious combinations of the parameters λ, v, γ with one degree of freedom are obtained by fixing the values of two of the parameters, say, $v = 0$, $\gamma = \frac{1}{8}$. This is equivalent with the introduction of the augmented equation

$$G(x) \equiv \begin{pmatrix} F(x) \\ v \\ \gamma \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 1/8 \end{pmatrix}. \quad (4.1)$$

Evidently, $G: S \subset \mathbb{R}^5 \rightarrow \mathbb{R}^4$ is again of class C^∞ on S , but we have now $R(G, S) \subset R(F, S)$ where the inclusion is strict. In fact, at $x_0 = (1/2, 0, 1/2, 0, 1/8)^T \in S$ we find that $\text{rank } (DG(x_0)) = 3$ is not maximal while $\text{rank } (DF(x_0)) = 2$ is so. Let M be the regular solution manifold of the equation $F(x) = 0$. Then $M_0 = M \cap R(G, S)$ is an open subset of M and $\Delta_x = \ker DG(x)$, $x \in M_0$, is a mapping which associates with each $x \in M_0$ a one-dimensional subspace of the tangent space $T_x M$. In this case, the second condition of a 1-distribution is easily verified. In fact, for any $x \in M_0$ a vector $v(x) \in T_x M$ is uniquely defined by

$$DG(x)v(x) = 0, \quad \|v(x)\|_2 = 1, \quad \det \begin{pmatrix} DG(x) \\ v(x)^T \end{pmatrix} > 0, \quad (4.2)$$

and the mapping $v: M_0 \rightarrow TM$ is of class C^∞ on M_0 . Hence, Δ is indeed a 1-distribution of class C^∞ on M_0 .

The point $x_0 \in R(F,S) \setminus R(G,S)$ is a singular point of this vector field v . In fact, for $\gamma = 1/8$ the bifurcation set in the λ, v -plane consists of a butterfly and touching dual cusps (see Figure 3 and [19]). The dual cusp point corresponds to x_0 .

For the case of (3.16) the definition (4.2) also provides us with the desired vector field on M_0 and hence allows the application of Theorem 4. But such a global definition of a vector field for a given 1-distribution is generally possible only in the finite-dimensional case. In [7] the relations between one-dimensional parameter combinations, augmented equations, 1-distributions, and local coordinates were considered in general for the infinite-dimensional case. We shall sketch only some of the results.

Let $F: S \subset X \rightarrow Y$ be a mapping between the Banach spaces X and Y which satisfies the conditions of Theorem 1. Thus, for $y_0 \in F(S)$ the regular solution manifold

$$M = \{x \in R(F,S); \quad F(x) = y_0\} \quad (4.3)$$

is well-defined.

The identification of certain of the variables of the problem as natural parameters is equivalent with the availability of an intrinsic splitting $X = Z \oplus \Lambda$ of X into a state space Z and an m -dimensional parameter space Λ . A combination of the natural parameters with one degree of freedom then corresponds to the choice of a one-dimensional subspace $\Lambda_1 \subset \Lambda$ which defines the remaining degree of freedom. We call Λ_1 a reduced parameter space of the problem. For example, in our example leading to (4.1) we used $\Lambda_1 = \text{span}(0,0,1,0,0)^T$.

With any reduced parameter space Λ_1 we may associate many different augmented mappings. In particular, there exist linear operators $L: \Lambda \rightarrow \mathbb{R}^{m-1}$ with $\ker L = \Lambda_1$, and with any such L we may define the augmented mapping

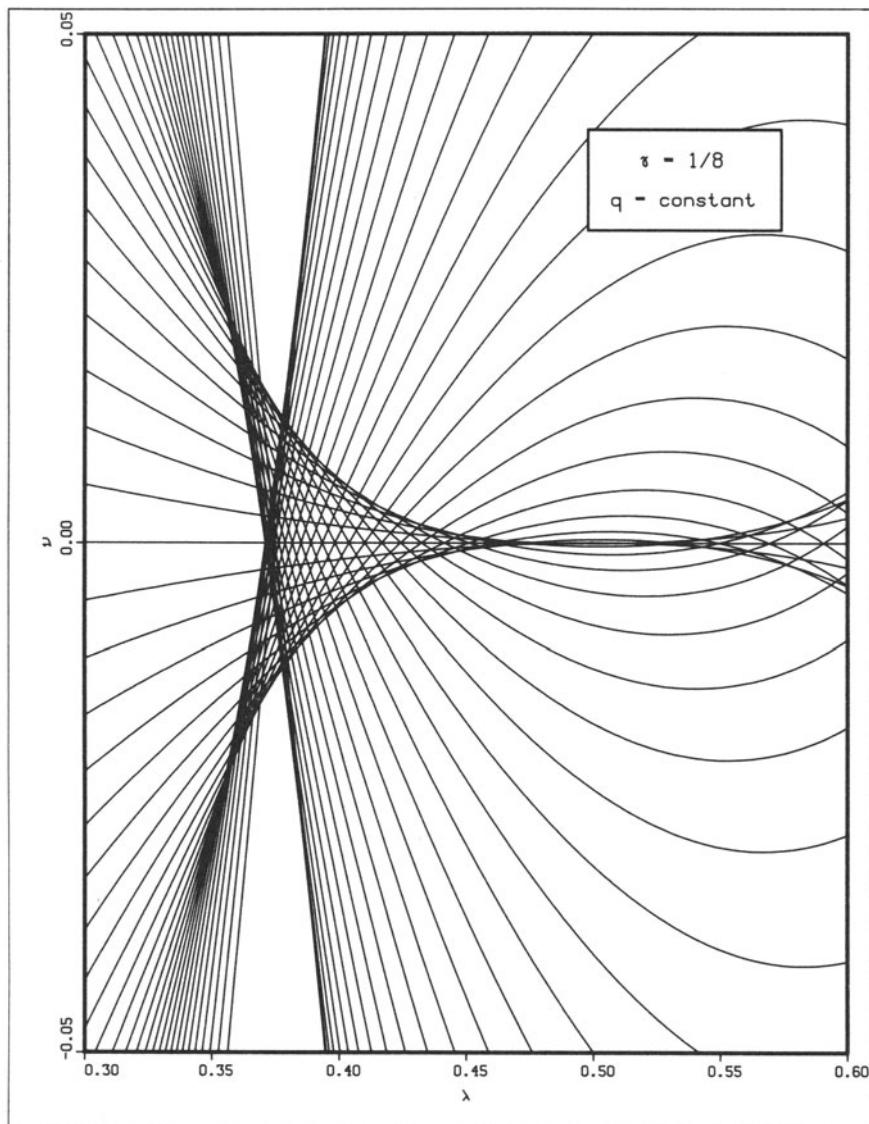


Figure 3

$$G: S \subset X \rightarrow Y \times R^{m-1}, \quad G(x) = (F(x), L\pi x), \quad \forall x \in S \quad (4.4)$$

where π denotes the projection of X onto Λ along Z . Any such choice of G generates the augmented equations

$$G(x) = (y_0, L\pi a_0) \quad (4.5)$$

where y_0 is the vector used in (4.3) and $a_0 \in X$ is arbitrary. For example, in (4.1) we chose

$$L = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad y_0 = 0, \quad a_0 = x_0. \quad (4.6)$$

By Theorem 1 applied to G the subset $M_0 = M \cap R(G, S)$ is open in M . Then the following generalization holds for the earlier example about 1-distributions:

Theorem 5: Under the stated assumptions about the various quantities, the mapping $\Delta: M_0 \rightarrow TM$, $\Delta_x = \ker DG(x)$, $x \in M_0$, is a 1-distribution of class C^{r-1} on M_0 . Moreover, if $r \geq 2$ then for any $a_0 \in M_0$ the regular solution set of the augmented equation (4.5) is an integral manifold of Δ on M_0 .

For the proof we refer to [7].

The 1-distribution of this theorem relates to the given reduced parameter space Δ_1 via the equation

$$\pi\Delta_x = \Delta_1, \quad \forall x \in M_0. \quad (4.7)$$

It turns out that any 1-distribution with this property can be discussed in the context of augmented functions of the form (4.4). More specifically the following result holds:

Theorem 6: Let $\Delta_1 \subset \Delta$ be a given reduced parameter space, and suppose that on some open subset M_0 of M a 1-distribution $\Delta: M_0 \rightarrow TM$ of class C^{r-1}

is defined for which (4.7) holds. Then for any augmented function G of the form (4.4) we have $\Delta_x \subset \ker DG(x)$, for each $x \in M_0$. Hence, any integral manifold N of Δ on M_0 is a solution manifold of an augmented equation (4.5) where a_0 is a given point on N .

Once again we refer to [7] for the proof.

These results show that the one-dimensional parameter combinations represented by a reduced parameter space $\Lambda_1 \subset \Lambda$ are essentially equivalent with 1-distributions on certain open subsets M_0 of M for which (4.7) holds. Our numerical problem now is to compute specific integral manifolds of these 1-distributions.

The open sets M_0 turned out to be the regular solution sets of the augmented mappings (4.4). At the same time, we know that the points of M not in M_0 are always regular for F . These points show a singular behavior only because we considered a particular one-dimensional parameter combination and the corresponding 1-distribution. The type of singularity may be characterized in terms of the augmenting function. We consider here only three commonly occurring cases:

Definition 7: Let $\Lambda_1 \subset \Lambda$ be a reduced parameter space, $W \subset X$, $X = W \oplus \Lambda_1$, a complementary subspace, and G any augmenting function (4.4). A point $x \in M$ is a

- (i) non-singular point of G if $\dim \ker DG(x) = 1$ and $W \cap \ker DG(x) = \{0\}$;
- (ii) limit point of G if $\dim \ker DG(x) = 1$ and $\ker DG(x) \subset W$;
- (iii) simple critical point of G if $\dim \ker DG(x) = 2$ and $\dim(W \cap \ker DG(x)) = 1$.

It turns out that $\ker DG(x)$ depends only on Λ_1 and not on the particular choice of the operator $L: \Lambda_1 \rightarrow \mathbb{R}^{M-1}$. Hence the characterization of the types of points depends only on Λ_1 and W and is otherwise independent of G . Actually, in [7] it was shown that we may characterize completely the nature of a point $x \in M$ in terms of the subspaces $Z \cap T_x^M$,

$\pi_{x_0}^T M$, and Λ_1 .

This discussion shows that a particular singularity usually disappears again when the reduced parameter space Λ_1 is changed. For example, suppose that in our spring example we fix $\lambda = 1/2$, $\gamma = 1/8$ and let v vary. Then it is readily checked that the resulting augmenting mapping is regular at the point x_0 where formerly the dual cusp of Figure 3 occurred. This opens up the possibility of considering numerical methods which avoid a particular singularity entirely by working with other more suitable 1-distribution.

5. Computational Aspects

As we saw, in the numerical analysis of manifolds at least two types of problems lead to the computation of paths on the given manifold. In the first case, these paths were the solution curves of certain vector fields while in the second one they constituted the integral manifolds of specific 1-distributions. In our discussion, the first class of problems was represented by differential-algebraic systems of equations while the second one involved augmented, parametrized equations with a one-dimensional parameter space.

By nature, the numerical methods for solving these two types of problems have the same principal features. In both cases, a predictor phase produces an approximation of a new point further along the path, and then a corrector process improves upon this approximation. But in their details the methods differ, as, for instance, in the controls upon which the steplength algorithms are based, the form of the corrector equations, or the parametrizations used for the paths. It will be useful to recall here briefly the underlying ideas of these methods.

Differential-algebraic systems (DAE's) are usually handled by means of suitable solvers for ordinary-differential-equation (ODE's). This approach has become widely accepted ever since it was proposed by Gear [9]. We sketch here only the ideas upon which such codes as LSODI, [11], or DASSL, [17], are based. For this suppose that the given DAE has the form

$$G(y', y, t) \equiv \begin{pmatrix} F(y, t) \\ H(y', y, t) \end{pmatrix} = 0 \quad (5.1)$$

where y, y' ($= dy/dt$) are n -dimensional and there are m algebraic equations and $n-m$ differential equations. Let $y = y(t)$ denote the desired solution for a specific initial condition $y(t_0) = y_0$. For the next step from the current approximation $y_{k-1} \stackrel{\circ}{=} y(t_{k-1})$, the predictor is based on the interpolating polynomial $q = q(t)$ that passes through the $s_k + 1$ most recently computed points along the path ($s_k \geq 1$). This polynomial is used to evaluate the predicted point $q(t_k)$, $t_k = t_{k-1} + \Delta t_k$, corresponding to a suitably chosen steplength Δt_k . Of course, for $k = 1$ this prediction has to be modified, for instance, by assuming that $y'(t_0)$ is known. For the corrector phase the derivative y' in H is replaced by a backward differentiation formula (BDF) of order s_k . At t_k this produces an equation of the form

$$G(\alpha_k y_k + b_k, y_k, t_k) = 0 \quad (5.2)$$

involving some $\alpha_k \in R^1$ and $b_k \in R^n$, as, for example, $\alpha_k = 1/\Delta t_k$, $b_k = -(1/\Delta t_k)y_{k-1}$, when $s_k = 1$. This equation (5.2) for the new point y_k is solved by means of a Newton- or chord-Newton process started at the predicted point $q(t_k)$.

The second class of problems involving parametrized equations is generally solved by continuation methods. The literature in this area is large (see eg. the surveys [3], [15], [27]) and once again we sketch here only the idea underlying a particular form of such a method ([22], [23]).

Suppose that the given augmented system has the form

$$G(x) \equiv \begin{pmatrix} F(x) \\ H(x) \end{pmatrix} = 0 \quad (5.3)$$

where H maps R^{n+1} into R^n . We wish to compute a segment of the one-dimensional, regular solution manifold $M_1 \subset R^{n+1}$ of (5.3). In this finite-dimensional case we may use the definition (4.2) of the vector field inherent in our 1-distribution to obtain a tangent vector $v(x)$ at $x \in M_1$. In other

words, if x_{k-1} is the last computed point approximating M_1 , then the predictor phase begins by computing the tangent vector $v_{k-1} = v(x_{k-1})$ by (4.2). Then, with an appropriately chosen steplength h_k , the predicted point $q_k = x_{k-1} + h_k v_{k-1}$ is obtained. For the manifold M_1 no intrinsic parametrization is available and various approaches for choosing a local parametrization have been proposed (see eg. [5], [12], [13], [20] and the theoretical discussion in [6], [7]). A simple and yet flexible technique results in the augmented equation

$$\hat{G}(x) \equiv \begin{pmatrix} G(x) \\ (e^i)^T(x - q_k) \end{pmatrix} = 0 \quad (5.4)$$

where $e^i \in R^{n+1}$ is one of the natural basis vectors of R^{n+1} . In essence, the index i is chosen such that $|(e^i)^T v_{k-1}|$ is maximal. Now the corrector phase consists in solving (5.4) for the next point $x_k \in M_1$ by means of the Newton- or chord-Newton process started at q_k .

Clearly, these descriptions of the two types of methods are very sketchy and in both cases there are many details that need to be considered. For this we refer to the relevant literature.

As these brief descriptions already indicate, the tangent vector of the solution path is utilized only in the continuation approach for parametrized equations but not in the ODE approach for DAE's. But the discussion in section 3 shows that the tangent vector is available also in the DAE case at about the same cost as in the continuation method. The availability of the tangent vector, in turn, opens up various new possibilities for solving DAE's.

In order to see this, suppose, as in Theorem 4, that the maps (3.8) are of class C^r , $r \geq 2$, on the open set $S \subset R^n$. Assume further that the set S_0 defined by (3.11) is open in R^n . Then, as in the proof of Theorem 4, we have a well-defined vector field

$$v: S_0 \rightarrow R^n, \quad N(y)v(y) = \begin{pmatrix} 0 \\ G(y) \end{pmatrix}, \quad \forall y \in S_0 \quad (5.5)$$

of class C^{r-1} on the n -dimensional manifold S_0 . Here $N(y)$ denotes again

the matrix of (3.10). Hence, by Theorem 3 there exists for any $y_0 \in S_0$ a unique maximally extended C^{r-1} solution of

$$y' = v(y) \quad (5.6)$$

through y_0 . Since $M_0 = S_0 \cap M$ is open in M and $v(y) \in TM$ for all $y \in S_0$ the restriction of v to M_0 is identical with the vector field defined in the proof of Theorem 4. This implies that for any $y_0 \in M_0$ the solution of (5.6) is identical with the solution of DAE (3.7) through that point on their common interval of definition. In other words, under the stated assumption the problem of solving the DAE may be replaced by that of solving the standard ODE (5.6).

We illustrate this with the example (3.4) (3.5) augmented by the equations (3.17), (3.18). In this case we have three algebraic equations and six differential equations and a seven-dimensional, dependent variable. In other words, the standard ODE-approach sketched earlier cannot be applied to the problem in this form, although we know that Theorem 4 applies. For any $y \in S_0 = \{y \in \mathbb{R}^7; y_3 > -1/4\}$ a simple calculation provides the following explicit expression for the vector field (5.5):

$$v(y) = \begin{pmatrix} y_4 \\ y_5 \\ 2y_1y_4 + 2y_2y_5 \\ -2y_1y_7 \\ -2y_2y_7 \\ 2y_4^2 + 2y_5^2 - 4y_7(y_1^2 + y_2^2) \\ -\frac{8y_7}{1+4y_4}(y_1y_4 + y_2y_5) \end{pmatrix}, \quad \forall y \in S_0. \quad (5.7)$$

The resulting ODE (5.6) was solved numerically with the well-known Runge-Kutta-Fehlberg solver RKF (see eg. [25]). For the starting point

$y_0 = (1, 0, 1, 0.5, 0.8, 1, 0.556)^T \in M \cap S$, Figure 4 shows the projection of a segment of the solution path onto the y_1, y_2 -plane.

Of course, in practical applications we do not have an explicit expression of the vector field (5.5). Instead for any given $y \in S_0$ the vector $v(y)$ has to be computed separately. This is numerically straightforward. In fact, if the DAE has the form (3.7) with $m_1 + m_2 = n$ then S_0 is equivalent with (3.13) and for any $y \in S_0$ the vector $v(y)$ is the unique solution of the $n \times n$ linear system (3.12). For $m_1 + m_2 > n$ this linear system is overdetermined but still has a unique solution which can be obtained with the help of a QR-decomposition of the matrix. In the case of a non-autonomous system (3.15) and variable parametrization of the solution path, the desired tangent vector is the unique null-vector of the $n \times (n+1)$ matrix K of (3.16). Hence its computation may proceed exactly as in the case of the continuation process described in [22]. There is no need to repeat the details here.

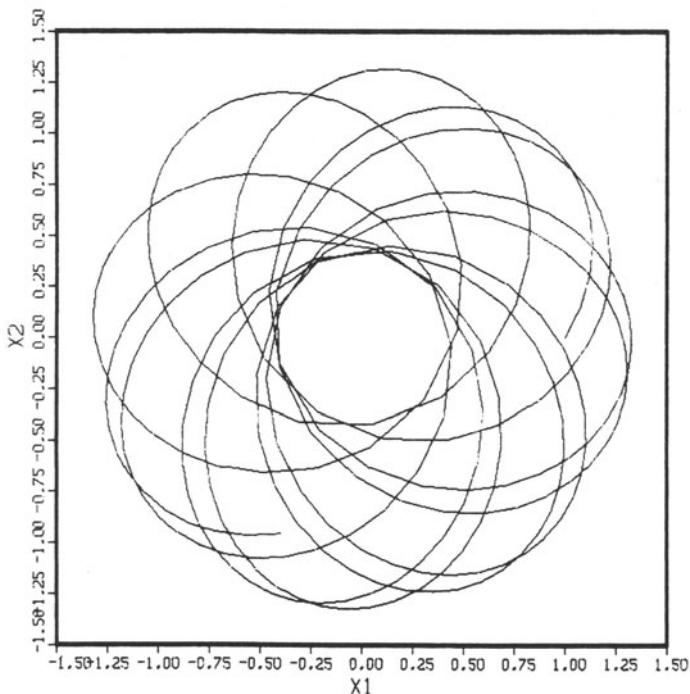


Figure 4

Once a routine for the computation of $v(y)$ for any $y \in S_0$ is available, any standard ODE solver may be applied to the solution of (5.6) for a given starting point $y_0 \in S_0$. A first question in the selection of such a solver is the possible stiffness of (5.6). Problems involving DAE's are generally considered to behave analogously to stiff systems of ODE's although this concept is not well-defined here. It appears that this behavior may be caused by the formulation of the problem as a DAE and that often the corresponding ODE (5.6) is far from being stiff. This is certainly true for our example and has been observed in a number of other cases as well. There is a need for a closer study of this stiffness question.

The ODE solvers applied to (5.6) possess no direct information about the manifold M underlying the problem and, in particular, their corrector equations are not guaranteed to produce solutions on M . Hence, in general, even if the starting point is exactly on M , the computed solution path will drift away from the manifold. For instance, the results of Figure 4 were computed in single precision (36 bit word) on a DEC-1090 (KL10) with the absolute and relative error tolerance in RKF set to 1.0E-5. At the first point of maximum elevation the maximum norm of the residual of the three algebraic equations (3.4), (3.17), (3.18) equaled 2.8E-6. At subsequent, computed points of maximum height this value increased to 7.8E-6, 1.7E-5, ..., 1.3E-4.

If it is essential for the results to be guaranteed to remain close to M , then either a further correction process has to be added or the corrector equations of the ODE-solver itself have to be modified to produce solutions on M . For the first approach one possibility consists in the application of some projection onto the manifold to any computed point that is too far away from it. This is a simple, although not entirely inexpensive, numerical task. In fact, it requires, in essence, the solution, by a chord-Newton method, of the (nonlinear) equations of M augmented by suitable (linear) equations defining the desired projection. Some details will be provided elsewhere together with a study of the error behavior of this overall method.

For the second of the two mentioned possibilities, one choice of a modified corrector equation for the ODE-solver is offered by the method for solving DAE's described at the beginning of this section. In fact, for

problems of the form (5.1) the solutions of the corrector equation (5.2) are guaranteed to lie on M . Any ODE-solver for (5.6) with its corrector based on (5.2) may also be viewed as a modification of the ODE-method for DAE's in which the simple extrapolatory predictor is replaced by one of the predictors commonly used for ODE's, as, for instance, an Adams-Bashforth formula applied to (5.6). This may be expected to provide for closer predictions, fewer corrector iterations, and potentially larger steps. There is certainly a need for a detailed analysis of the error behavior of such combinations of predictors with the corrector induced by (5.2). At the same time, comparisons with the earlier indicated method involving projections onto M would be useful as well.

6. References

- [1] E. Abraham and J. Marsden, Foundations of Mechanics, Second Edition, The Benjamin/Cummings Publ. Co., London 1978.
- [2] S. Agmon, A. Douglis and L. Nirenberg, Estimates Near the Boundary for Solutions of Elliptic Partial Differential Equations Satisfying General Boundary Conditions I, Comm. Pure Appl. Math. 12, 1959, 623-727.
- [3] E. Allgower and K. Georg, Simplicial and Continuation Methods for Approximating Fixed Points and Solutions to Systems of Equations, SIAM Review 22, 1980, 28-85.
- [4] M. Bier, O. A. Palusinski, R. A. Mosher and D. A. Saville, Electrophoresis: Mathematical Modeling and Computer Simulation, Science, Vol. 219, 18 March 1983, No. 4590, 1281-1286.
- [5] M. A. Crisfield, A Fast Incremental/Iterative Solution Procedure that Handles "Snap Through", Comp. and Structures 13, 1981, 55-62.
- [6] J. P. Fink and W. C. Rheinboldt, On the Discretization Error of Parametrized Nonlinear Equations, SIAM J. Num. Anal. 20, 1983, 732-746.
- [7] J. P. Fink and W. C. Rheinboldt, Solution Manifolds and Submanifolds of Parametrized Equations and Their Discretization Error, Univ. of Pittsburgh, Inst. f. Comp. Math. and Appl., Tech. Report ICMA-83-59, 1983.
- [8] W. Fleming, Functions of Several Variables, Second Edition, Springer Verlag, New York, 1977.
- [9] C. W. Gear, Simultaneous Numerical Solution of Differential-Algebraic Equations, IEEE Trans. on Circuit Theory, CT-18, 1971, 89-95.

- [10] C. W. Gear and L. R. Petzold, ODE Methods for the Solution of Differential-Algebraic Systems, Univ. of Illinois at Urbana-Champaign, Dept. of Comp. Science, Tech. Report 82-1103.
- [11] A. C. Hindmarsh, ODE Solvers for Use with the Method of Lines, in "Adv. in Computer Methods for Partial Diff. Eqn. IV", ed. by R. Vichnevetsky and R. S. Stepleman, IMACS, New Brunswick, NJ 1981, 312-316.
- [12] H. B. Keller, Numerical Solution of Bifurcation and Nonlinear Eigenvalue Problems, in "Applications of Bifurcation Theory", ed. by P. Rabinowitz, Academic Press, New York, NY 1977, 359-384.
- [13] H. B. Keller, Global Homotopies and Newton Methods, in "Recent Advances in Numerical Analysis", ed. by C. deBoor, G. H. Golub, Academic Press, New York, NY 1978, 73-94.
- [14] S. Lang, Introduction to Differentiable Manifolds, Wiley and Sons, New York 1962.
- [15] H. D. Mittelmann and H. Weber, A Bibliography on Numerical Methods for Bifurcation Problems, Univ. Dortmund, Angew. Mathem., Bericht 56, 1981.
- [16] R. W. Newcomb, The Semistate Description of Nonlinear Time-Variable Circuits, IEEE Trans. on Circuits and Systems, CAS-28, 1981, 62-71.
- [17] L. R. Petzold, A Description of DASSL: A Differential-Algebraic System Solver, in "Proc. IMACS World Congress 1982", to appear.
- [18] L. R. Petzold, Differential-Algebraic Equations are not ODE's, SIAM J. on Scientific and Statist. Computing 3, 1982, 367-384.
- [19] T. Poston and I. Stewart, Catastrophe Theory and its Applications, Pitman Publ. Ltd., London 1978.
- [20] W. C. Rheinboldt, Solution Fields of Nonlinear Equations and Continuation Methods, SIAM J. Num. Anal. 17, 1980, 221-237.
- [21] W. C. Rheinboldt, Differential-Algebraic Systems as Differential Equations on Manifolds, Univ. of Pittsburgh, Inst. f. Comp. Math. and Appl., Tech. Report ICMA-83-55, 1983.
- [22] W. C. Rheinboldt and J. V. Burkardt, A Locally Parametrized Continuation Process, ACM Trans. on Math. Software 9, 1983, 215-235.
- [23] W. C. Rheinboldt and J. V. Burkardt, Algorithm 596: A Program for a Locally Parametrized Continuation Process, ACM Trans. on Math. Software 9, 1983, 236-241.
- [24] M. Schechter, Principles of Functional Analysis, Academic Press, NY 1971.
- [25] L. F. Shampine and R. C. Allen, Jr., Numerical Computing: An Introduction, W. B. Saunders Co., Philadelphia, PA 1973.

- [26] M. Spivak, A Comprehensive Introduction to Differential Geometry, Vol. I, Second Edition, Publish or Perish, Inc., Berkeley, CA 1979.
- [27] H. J. Wacker (editor), Continuation Methods, Academic Press, New York, NY 1978.

DIRECT METHODS FOR THE COMPUTATION OF A NONSIMPLE TURNING POINT
CORRESPONDING TO A CUSP.

D. Roose and R. Caluwaerts

We present a method for the numerical computation of a double turning point of a nonlinear operator equation depending on two parameters, which corresponds to a cusp catastrophe. We introduce two variants of an augmented system, for which the double turning point is an isolated solution. We discuss the implementation of this method in the case of differential equations and integral equations. Results are given for some chemical engineering problems.

1. Introduction

We consider a nonlinear operator equation

$$(1) \quad G(\lambda, \gamma, u) = 0 \quad G : \mathbb{R} \times \mathbb{R} \times X \rightarrow Y \quad (X \subseteq Y)$$

where X and Y are real Banach spaces and G is a C^3 mapping. Usually, we regard $u(\lambda, \gamma)$ as the solution of (1) depending on two parameters λ and γ . Let

$$(2) \quad F(\lambda, u) = G(\lambda, \gamma^*, u) \quad F : \mathbb{R} \times X \rightarrow Y$$

for a fixed value of $\gamma = \gamma^*$.

Assume that the solution curve $u(\lambda)$ of $F(\lambda, u) = 0$ exhibits a turning point behaviour for $\gamma^* < \gamma^{\circ}$ and that these turning points disappear for $\gamma^* > \gamma^{\circ}$ (see Fig. 1). For $\gamma = \gamma^{\circ}$ the two "simple" turning points coalesce in a "nonsimple" turning point $(\lambda^{\circ}, u^{\circ})$. It can be shown that such a critical point $(\lambda^{\circ}, \gamma^{\circ}, u^{\circ})$ corresponds to a cusp catastrophe point [18].

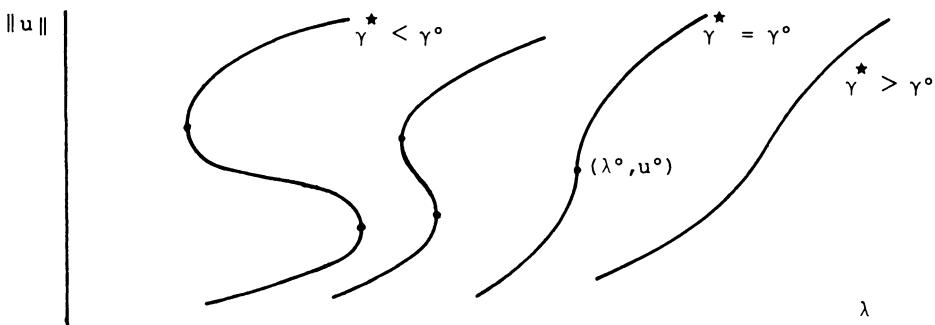


Fig. 1 : The solution of $F(\lambda, u) = G(\lambda, \gamma^*, u) = 0$ (γ^* fixed) exhibits a nonsimple turning point $(\lambda^{\circ}, u^{\circ})$ for $\gamma^* = \gamma^{\circ}$.

Direct methods for the computation of branching points (simple turning points, bifurcation points) are developed by several authors (e.g. [10,14,15, 16,20]). The original problem is embedded in a larger system of equations, which has the branching point as an isolated solution. The augmented system can thus be solved by a standard method.

In this paper we discuss similar methods for the computation of a cusp point $(\lambda^0, \gamma^0, u^0)$. We also indicate how these methods can be implemented in the case of ordinary and partial differential equations and integral equations. Here we emphasize on the efficiency of the numerical procedure and the convergence of discrete approximations.

2. Double turning points and cusps

We recall some definitions and results about turning points in one and two parameter problems [10,18]. We denote by $G_u : X \rightarrow Y$ and $G_u^* : Y^* \rightarrow X^*$ the Fréchet derivative of G with respect to u and its dual, respectively. A solution (λ^t, u^t) of $F(\lambda, u) = G(\lambda, \gamma^*, u) = 0$ (γ^* fixed) is a turning point with respect to λ if the following conditions are satisfied :

- (3a) $G_u^*(\lambda^t, \gamma^*, u^t)$ has a one-dimensional nullspace spanned by $\phi \in X$
- (3b) $G_u^*(\lambda^t, \gamma^*, u^t)$ has a one-dimensional nullspace spanned by $\psi \in Y^*$
- (3c) $R(G_u^*(\lambda^t, \gamma^*, u^t)) = \{y \in Y : \psi(y) = 0\}$ is closed with codim 1
- (3d) $G_\lambda(\lambda^t, \gamma^*, u^t) \notin R(G_u^*(\lambda^t, \gamma^*, u^t))$

In a neighbourhood U of the turning point (λ^t, u^t) the solution curve of $F(\lambda, u) = 0$ can be parametrized such that

$$F^{-1}(0) \cap U = \{(\lambda(s), u(s)) : |s-s^t| \leq \delta\}$$

where $s \in \mathbb{R}$, $\delta > 0$, $\lambda(\cdot)$ and $u(\cdot)$ are C^3 mappings satisfying $\lambda(s^t) = \lambda^t$, $u(s^t) = u^t$ and $\lambda'(s^t) = 0$, $u'(s^t) = \phi$ [18].

Definition [18] :

A turning point (λ^t, u^t) is called simple iff $\lambda''(s^t) \neq 0$; (λ^t, u^t) is double iff $\lambda''(s^t) = 0$, $\lambda'''(s^t) \neq 0$.

Spence and Werner [18] proved the following Lemma

Lemma 1 :

A turning point (λ^t, u^t) of $F(\lambda, u) = G(\lambda, \gamma^*, u) = 0$ is simple iff

$$(4) \quad a_2 = \psi(G_{uu}(\lambda^t, \gamma^*, u^t) \phi \phi) \neq 0.$$

(λ^t, u^t) is a double turning point iff

$$(5a) \quad a_2 = 0$$

$$(5b) \quad a_3 = \psi(G_{uuu}(\lambda^t, \gamma^*, u^t) \phi \phi \phi + 3G_{uu}(\lambda^t, \gamma^*, u^t) \phi v) \neq 0$$

where v is a solution of

$$(6) \quad G_u(\lambda^t, \gamma^*, u^t)v = -G_{uu}(\lambda^t, \gamma^*, u^t)\phi \phi. \quad \square$$

Assume that Eq. (1) has a turning point behaviour as described in section 1. Then we infer from Fig. 1 that the critical point $(\lambda^0, \gamma^0, u^0)$ is a double turning point of $G(\lambda, \gamma^0, u) = 0$ with respect to λ .

If the parameter γ is varied also, the projection of the turning point set on the (λ, γ) -plane shows a cusp-like behaviour.

We now illustrate the connection between double turning points and cusp catastrophe points. The canonical cusp catastrophe is described by [12, 21]

$$(7) \quad x^3 + \gamma x + \lambda = 0$$

which has a cusp point $(\lambda, \gamma) = (0, 0)$. For $\gamma = 0$, Eq. (7) has a double turning point $(\lambda, x) = (0, 0)$. If λ is fixed ($\lambda=0$), Eq. (7) exhibits a pitchfork bifurcation behaviour.

Theorem 1 [18] :

A double turning point $(\lambda^0, \gamma^0, u^0)$ of a two parameter problem $G(\lambda, \gamma, u) = 0$ corresponds to a cusp catastrophe point provided that

$$(8) \quad \zeta(G_\gamma(\lambda^0, \gamma^0, u^0)) + \psi(G_{\gamma u}(\lambda^0, \gamma^0, u^0)\phi) \neq 0$$

with $\zeta \in \gamma^*$ defined by

$$\zeta(G_\lambda(\lambda^0, \gamma^0, u^0)) = -\psi(G_{\lambda u}(\lambda^0, \gamma^0, u^0)\phi)$$

(9) $\zeta(G_u(\lambda^0, \gamma^0, u^0)) = -\psi(G_{uu}(\lambda^0, \gamma^0, u^0)\phi). \quad \square$

Indeed, $G(\lambda, \gamma, u) = 0$ is locally equivalent to the (scalar) bifurcation equation

$$(10) \quad H(\bar{\lambda}, \bar{\gamma}, \bar{u}) = 0$$

with $\bar{\lambda}, \bar{\gamma}$ and $\bar{u} \in \mathbb{R}$, which can be obtained by the Lyapunov-Schmidt method.

If the generic condition

$$(11) \quad D = \begin{vmatrix} H_{\bar{\lambda}}(0) & H_{\bar{\lambda}\bar{u}}(0) \\ H_{\bar{\gamma}}(0) & H_{\bar{\gamma}\bar{u}}(0) \end{vmatrix} \neq 0$$

is satisfied, there exists a smooth change of coordinates such that $H = 0$ has the normal form of the cusp catastrophe given by Eq (7). Spence and Werner [18] proved that this generic condition is satisfied for a double turning point if condition (8) holds.

The correspondence with a cusp catastrophe is also shown by the following theorem [13] :

Theorem 2 :

If condition (8) is satisfied, there exists a linear transformation

$$[\hat{\lambda} \ \hat{\gamma}] = [\lambda \ \gamma] \times T \quad (T \in \mathbb{R}^{2 \times 2})$$

such that the double turning point $(\lambda^0, \gamma^0, u^0)$ of $G(\lambda, \gamma, u) = 0$ is a pitchfork bifurcation point with regard to $\hat{\gamma}$. \square

The augmented systems which we propose in the next section have an isolated solution at a double turning point, only if it corresponds to a cusp point.

3. Direct methods for the computation of cusp points

Method I.

Consider the following augmented system

$$(12) \quad K^{(1)}(y) = 0 \quad K^{(1)} : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \times X \times X \times X^* \rightarrow Y \times Y \times X^* \times \mathbb{R} \times \mathbb{R} \times \mathbb{R}$$

with $y = (\lambda, \mu, \gamma, u, p, q)$

and

$$(13) \quad K^{(1)}_{y}(\lambda, \mu, \gamma, u, p, q) = \begin{cases} G(\lambda, \gamma, u) \\ G_u(\lambda, \gamma, u)p \\ \star G_u^*(\mu, \gamma, u)q \\ \ell_1(p)-1 \\ \ell_2(q)-1 \\ q(G_{uu}(\lambda, \gamma, u)pp) \end{cases}$$

Here ℓ_1 and ℓ_2 are functionals, chosen to scale p and q . If $(\lambda^*, \gamma^*, u^*)$ is a double turning point of $G(\lambda, \gamma, u)$ with regard to λ , then $y^* = (\lambda^*, \mu^*, \gamma^*, u^*, \phi, \psi)$ with $\mu^* = \lambda^*$ is a solution of $K^{(1)}_y(y) = 0$, provided that $\ell_1(\phi) = \ell_2(\psi) = 1$. A double turning point can thus be determined by solving system (12). If the double turning point corresponds to a cusp point, the solution y^* will be isolated and can thus be computed by standard methods, as indicated by the following theorem [13].

Theorem 3 :

Let X be a norm-reflexive Banach space and let $(\lambda^*, \gamma^*, u^*)$ be a double turning point of $G(\lambda, \gamma, u) = 0$ with respect to λ . Then the Fréchet derivative $K^{(1)}_y(y^*)$ is nonsingular if condition (8) is satisfied and if

$$(14) \quad \psi(G_{\lambda u}(\lambda^*, \gamma^*, u^*)\phi) \neq 0. \quad \square$$

Condition (14) holds if the system

$$(15) \quad M(\lambda, p) = \begin{cases} G_u(\lambda, \gamma^*, u^*)p \\ \ell_1(p)-1 \end{cases} = 0$$

has an isolated solution (λ^*, ϕ) , i.e. the linear operator $G_u(\lambda, \gamma^*, u^*)$ is regular for all $\lambda \in (\lambda^* - \epsilon, \lambda^* + \epsilon)$, $\lambda \neq \lambda^*$ ($\epsilon > 0$) [13]. This additional condition (14) is related to the use of the extra variable μ in system (13).

Method II.

An alternative approach is to consider the augmented system

$$(15) \quad K^{(2)}(y) = 0 \quad K^{(2)} : \mathbb{R} \times \mathbb{R} \times \mathbb{X} \times \mathbb{X} \times \mathbb{X} \rightarrow \mathbb{Y} \times \mathbb{Y} \times \mathbb{Y} \times \mathbb{R} \times \mathbb{R}.$$

with $y = (\lambda, \gamma, u, p, w)$

and

$$(16) \quad K^{(2)}(\lambda, \gamma, u, p, w) = \begin{cases} G(\lambda, \gamma, u) \\ G_u(\lambda, \gamma, u)p \\ G_{uu}(\lambda, \gamma, u)w + G_{uu}(\lambda, \gamma, u)\phi\phi \\ l(p)-1 \\ k(p, w) \end{cases}$$

Here $l : X \rightarrow \mathbb{R}$ and $k : X \times X \rightarrow \mathbb{R}$ are (not necessarily linear) functions such that

$$(17) \quad \frac{l_p(\phi)}{p} \neq 0$$

$$(18) \quad k_w(\phi, v)\phi \neq 0$$

Since $y^0 = (\lambda^0, \gamma^0, u^0, \phi^0, v)$ with v defined by (6) is a solution of $K^2(y) = 0$, a double turning point can be computed by solving system (15). It can be proved [2] that $K_y^{(2)}(y^0)$ is regular if the generic condition (11) is satisfied. No additional information like (14) is required.

4. Convergence of discretizations and numerical procedure

If $G(\lambda, \gamma, u) = 0$ is a system of differential equations or integral equations, the following strategies can be adopted :

- The augmented system (12) or (15) corresponding to the original equation is constructed analytically. The resulting system of differential or integral equations (with special "side conditions") can be solved numerically to compute an approximation of the cusp point of the original equation.
- The equation $G(\lambda, \gamma, u) = 0$ is discretized to obtain a nonlinear system of algebraic equations $G_h(\lambda, \gamma, u_h) = 0$. This system can be enlarged to compute the cusp point of the approximate problem $G_h(\lambda, \gamma, u_h) = 0$.

If strategy a) is adopted, the standard convergence results are valid, since a cusp point is an isolated solution of systems (12) and (15) (see e.g. [8]). The presence of some real parameters in our problem doesn't prohibit from

using the standard theory. The existence of an asymptotic expansion for the discretization error of the form

$$(19) \quad \lambda_h^{\circ} = \lambda^{\circ} + c_1 h^{p_1} + c_2 h^{p_2} + \dots + c_n h^{p_n} + O(h^{p_{n+1}})$$

(and similar expressions for γ_h° and u_h°) can be proved for a large class of discretization algorithms using a theorem of Stetter [19].

Strategy b) can also be used since a cusp catastrophe is stable in a two parameter problem [21]. We can consider the discrete problem as a perturbed bifurcation problem and $G_h(\lambda, \gamma, u_h) = 0$ will have a cusp point $(\lambda_h^{\circ}, \gamma_h^{\circ}, u_h^{\circ})$ near $(\lambda^{\circ}, \gamma^{\circ}, u^{\circ})$. In this case, based on the results of Spence and Werner [18; Theorem 3.1], we can use the convergence theory for simple turning points, developed by Moore and Spence [11]. If an $O(h^p)$ discretization is used, it follows that

$$(20) \quad |\lambda^{\circ} - \lambda_h^{\circ}|, |\gamma^{\circ} - \gamma_h^{\circ}|, \|u^{\circ} - u_h^{\circ}\| = O(h^p).$$

However, for a large class of discretization methods, strategy b) results in the same augmented system as could be obtained with approach a). Then the convergence results of case a) hold.

Good starting values for the resulting system can be obtained e.g. by a continuation process along a "branch of simple turning points" (see [18], Theorem 3.1). At a simple turning point $(\lambda^t, \gamma^t, u^t)$ the value of $\psi_{uu}^t(G_{uu}(\lambda^t, \gamma^t, u^t) \phi^t \phi^t)$ can be computed. If this value approaches zero, the nearby double turning point can be determined by solving system (12) or (15), using the information of the last continuation step to generate starting values. One can also start with a large discretization step h_0 . The resulting system (12) or (15) will be of low dimension and can be solved by a robust method to obtain a first approximation $(\lambda_{h_0}^{\circ}, \gamma_{h_0}^{\circ}, u_{h_0}^{\circ})$ of the cusp point.

A steplength sequence $\{h_i : h_i < h_{i-1}, i=1,2,3,\dots\}$ can be used to obtain more accurate approximations. Starting values for $h = h_i$ can be generated from the solution with $h = h_{i-1}$. In the case of integral equations, this can be done by the Nyström-method [1].

If approximations with several stepsizes h_i ($i=0,1,2,\dots$) are available, repeated Richardson extrapolation [7] can be used to improve the results, based on an expansion of the form (19).

In the case of a nonlinear system of algebraic equations, Eq. (12) or (15) can be solved by Newton's method

$$(21a) \quad K_y(y^k) \delta y^k = -K(y^k) \quad (k = 0, 1, 2, \dots)$$

$$(21b) \quad y^{k+1} = y^k + \delta y^k$$

One can take advantage of the structure of the systems (12) and (15) to reduce the computational work for the solution of the linear system (21a) of dimension $(3n+3)$ or $(3n+2)$. For the solution of system (12), one can use the technique described in [13], which is closely related to the method developed in [10]. Then the solution of (21a) consists of the LU-decomposition of two $n \times n$ matrices and 9 backsubstitutions. A similar technique for the solution of system (15) requires only one LU-decomposition of a $n \times n$ matrix.

5. Computation of cusp points for two point boundary value problems

If the original problem $G(\lambda, \gamma, u) = 0$ consists of a boundary value problem in ordinary differential equations, we can transform system (12) or system (15) into a system of differential equations in "standard" form, in order to use general-purpose software for the computation of the cusp point. Such a transformation is already used for the computation of simple turning points [14] and for bifurcation points and the bifurcating branches [9, 16, 20].

We consider boundary value problems of the form

$$(22a) \quad y' = f(x, y, \lambda, \gamma)$$

with boundary conditions

$$(22b) \quad r(y(a), y(b)) = 0$$

where $x \in [a, b]$, $f : [a, b] \times \Omega \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^n$, $0 \in \Omega \subset \mathbb{R}^n$. Then system (12) can be written as

$$(23) \quad \left\{ \begin{array}{l} y' = f(x, y, \lambda, \gamma) \\ p' = f_y(x, y, \lambda, \gamma)p \\ q' = -f_y^T(x, y, \mu, \gamma)q \\ w_1' = p^T p \\ w_2' = q^T q \\ z' = q [f_{yy}(x, y, \lambda, \gamma) pp] \\ \lambda' = 0 \\ \mu' = 0 \\ \gamma' = 0 \end{array} \right.$$

with boundary conditions

$$(24) \quad \left\{ \begin{array}{l} r(y(a), y(b)) = 0 \\ B_a p(a) + B_b p(b) = 0 \\ B_a^* q(a) + B_b^* q(b) = 0 \\ w_1(a) = 0; w_1(b) = 1 \\ w_2(a) = 0; w_2(b) = 1 \\ z(a) = 0; z(b) = 0 \end{array} \right.$$

Here B_a , B_b , B_a^* , B_b^* are $n \times n$ matrices with $\text{rank } [B_a \ B_b] = \text{rank } [B_a^* \ B_b^*] = n$, defined by (see e.g. [9, 20])

$$(25) \quad B_a = \frac{\partial r(y(a), y(b))}{\partial y(a)}; \quad B_b = \frac{\partial r(y(a), y(b))}{\partial y(b)}; \quad B_a B_a^{*T} - B_b B_b^{*T} = 0$$

Here we used the scalar product in L_2

$$(26) \quad \langle u, v \rangle = \int_a^b u(x) v(x) dx$$

in the equation which express the condition for a double turning point. The functions p and q are scaled by requiring $\langle p, p \rangle = \langle q, q \rangle = 1$. This scaling can also be incorporated in the boundary conditions, without requiring two extra equations in the system (23) (see [14]). The resulting system

(23) can now be solved by any standard numerical method for two point boundary value problems. Similarly we can transform system (15) into a regular boundary value problem.

If the original problem is a scalar second order equation of the form

$$(27) \quad u'' = g(x, u, \lambda, \gamma)$$

$$R(u(a), u(b), u'(a), u'(b)) = 0$$

method I can be substantially simplified. Indeed, if Eq. (27) is transformed into a system of first order equations of the form (22), then

$$(28) \quad f_y = \begin{vmatrix} 0 & 1 \\ g_u & 0 \end{vmatrix}$$

If $p = (p_1(x), p_2(x))^T$ satisfies the linearized equation

$$(29) \quad p' = f_y(x, y, \lambda, \gamma)p \quad B_a^* p(a) + B_b^* p(b) = 0$$

then $q = (p_2(x), -p_1(x))^T$ satisfies the adjoint equation

$$(30) \quad q' = -f_y^T(x, y, \lambda, \gamma)q \quad B_a^* q(a) + B_b^* q(b) = 0$$

for the common types of boundary conditions.

Thus, the adjoint equation can be removed from system (23), together with the equation $u' = 0$.

6. Results

a. ordinary differential equations :

Heat and mass transfer in a tubular reactor [4] is described by

$$(31) \quad \frac{1}{Pe} \frac{d^2y}{dx^2} - \frac{dy}{dx} + Da(1-y) \exp\left(\frac{By}{1+\epsilon By}\right) = 0$$

$$y'(0) = Pe \cdot y(0) \quad ; \quad y'(1) = 0$$

If the parameters B and ϵ are fixed, the solution $y(x)$ shows a cusp catastrophe behaviour on the $Da - Pe$ plane. In the neighbourhood of the cusp

point $(Da^\circ, Pe^\circ, y^\circ)$, the solution $y^\circ(x)$ and the eigenfunction $\phi(x)$ exhibits a boundary layer at $x = 1$ (see Table 1). The cusp point can easily be computed by solving system (15) (method II) with (backward) shooting. We used the multiple shooting code BOUNDSOL of Bulirsch et al. (see e.g. [14] for references). Results are reported in Table 1 and can be compared with those obtained in [4]. Note that in this case the application of method I gives rise to severe numerical difficulties since the left eigenfunction $\psi(x)$ exhibits a boundary layer at $x = 0$.

	$\varepsilon = 0.05$		$\varepsilon = 0.0$	
Da°	0.11205		0.11170	
Pe°	31.480		630.67	
x	$y^\circ(x)$	$\phi(x)$	$y^\circ(x)$	$\phi(x)$
0	0.004	0.1E-8	0.0002	< 1.E-10
0.25	0.04	0.4E-8	0.03	"
0.50	0.09	0.4E-5	0.08	"
0.75	0.19	0.006	0.15	"
0.85	0.28	0.08	0.21	"
0.90	0.40	0.26	0.26	"
0.95	0.56	0.68	0.34	< 1.E-10
0.98	0.67	0.93	0.44	0.6E-4
1.00	0.70	1.00	0.85	1.00

Table 1 : Cusp point for Eq. (32) ($B = 10$)

b. partial differential equations.

Mass transfer and chemical reaction of the Langmuir-Hinshelwood mechanism in a porous catalyst is described by [5] :

$$(32) \quad \frac{\partial^2 Y}{\partial x^2} + \frac{\partial^2 Y}{\partial y^2} = \lambda \left(\frac{1+B}{1+BY} \right)^2 Y \left(\frac{Y+C}{1+C} \right) \quad (x, y) \in D = (-1, 1) \times (-1, 1)$$

$$Y(x, y) = 1 \quad (x, y) \in \partial D$$

Eq. (32) is discretized using the Stormer-Numerov finite difference scheme (see e.g. [13]), which is $O(h^4)$ accurate. For $C = 1$, approximations of the cusp point $(\lambda^\circ, B^\circ, Y^\circ)$ are computed using method I. Results are given in Table 2. The underlined figures are exact.

h	λ_h°	B_h°	Y_h°
1/4	<u>1.7960</u>	<u>16.93</u>	0.0 <u>895</u>
1/8	<u>1.800043</u>	<u>16.782447</u>	0.0 <u>868566</u>

Table 2 : Approximations of the cusp point of Eq. (32) ($C = 1$)

c. integral equations.

Heat and mass transfer in a porous catalyst [5] can be described by

$$(33) \quad f(x) - \lambda \int_0^1 f(y) \exp\left(\frac{\gamma\beta(1-f(y))}{1+\beta(1-f(y))}\right) k(x,y) dy = 1 \quad x \in [0,1]$$

where $k(x,y) = \max(x,y) - 1$.

The integral in (33) is discretized by the trapezoidal rule. For $\beta = 0.4$, method II is used to compute approximations of the cusp point $(\lambda^\circ, Y^\circ, f^\circ)$. The triangular table given in Table 3 is found by Richardson extrapolation on approximations λ_h° , based on an asymptotic expansion of the form (19) with $p_i = 2i$. Following the procedure defined by Dobrovol'skii [3] we find the most accurate result below in column 4 (marked with $*$).

$h=1/3$	0.22033		
		0.222858035	
$h=1/6$	0.22222	0.22286264304	
		0.222862355	0.222862606553
$h=1/12$	0.22270	0.22286260712	0.222862606624
		0.222862591	0.222862606624*
$h=1/24$	0.22282	0.22286260663	
		0.222862605	
$h=1/48$	0.22285		

Table 3 : Richardson extrapolation table for approximations λ_h° of Eq. (33)
 $(\beta = 0.4) \quad (\gamma^\circ = 14.4032220, f^\circ(0) = 0.526489302)$

7. Conclusion

We discussed two closely related methods for the computation of a cusp catastrophe point of an operator equation. In both approaches, a three times larger system has to be solved, for which the cusp point is an isolated solution.

Comparing these two methods, we conclude that in general method II is superior. Method I has the drawback that the additional condition (10) must be satisfied, but this will not be a problem in practice. The adjoint operator $G_u^*(\lambda, \gamma, u)$ enters in the system and must be computed, which is a major disadvantage. In the case of a large system of nonlinear algebraic equations, the efficient implementation of the Newton iteration requires for method I about twice as much computations as for method II.

However, using method II, it can be difficult to provide a sufficiently good starting value for the unknown function v , which has not such a "physical meaning" as the unknown eigenfunction ψ in method I. If the cusp point is roughly located by monitoring a test function during a continuation process, an approximation of ψ will already be available. Method I is superior for the computation of a cusp point of a second order two point boundary value problem without first derivative using o.d.e.-software.

Note that the problem discussed here is related to the subject of the papers of Jepson and Spence in these proceedings [6,17].

Acknowledgement

The authors wish to thank Prof. R. Piessens and Prof. V. Hlavacek for the many valuable discussions. The work of one of the authors (D.R.) was supported by N.F.W.O. while visiting the State University of New York at Buffalo. This support is gratefully acknowledged.

References

- [1] K. Atkinson : A survey of numerical methods for the solution of Fredholm integral equations of the second kind. Philadelphia : SIAM (1976).
- [2] R. Caluwaerts : A direct method for the determination of non-simple turning points. In preparation.

- [3] I.P. Dobrovol'skii : Richardson extrapolation in the approximate solution of Fredholm integral equations of the second kind. U.S.S.R. Comput. Maths. Math. Phys. 21, 139-149 (1981).
- [4] V. Hlavacek and H. Hofmann : Modelling of chemical reactors - XVI & XVII. Steady state axial heat and mass transfer in tubular reactors. Chem. Engng. Sci. 25, 173-199 (1970).
- [5] V. Hlavacek, M. Kubicek, J. Caha : Qualitative analysis of the behaviour of nonlinear parabolic equations - II. Chem. Engng. Sci. 26, 1743-1752 (1971).
- [6] A.D. Jepson and A. Spence : Paths of singular points and their computation. These proceedings.
- [7] D.C. Joyce : Survey of extrapolation processes in Numerical Analysis. SIAM Review 13, 435-488 (1971).
- [8] H.B. Keller : Approximation methods for nonlinear problems with application to two-point boundary value problems. Math. Comp. 29, 464-474 (1975).
- [9] W.F. Langford : Numerical solution of bifurcation problems for ordinary differential equations. Numer. Math. 28, 171-190 (1977).
- [10] G. Moore and A. Spence : The calculation of turning points of nonlinear equations. SIAM J. Numer. Anal. 17, 567-576 (1980).
- [11] G. Moore and A. Spence : The convergence of operator equations at turning points. IMA J. Numer. Anal. 1, 23-38 (1981).
- [12] T. Poston and I.N. Stewart : Catastrophe theory and its applications. London : Pitmann Press 1978.
- [13] D. Roose and R. Piessens : Numerical computation of nonsimple turning points and cusps. Report TW60, Dept. of Computer Science, Universiteit Leuven (1983). (submitted to Numer. Math.).
- [14] R. Seydel : Numerical computation of branch points in ordinary differential equations. Numer. Math. 32, 51-68 (1979).
- [15] R. Seydel : Numerical computation of branch points in nonlinear equations. Numer. Math. 33, 339-352 (1979).
- [16] R. Seydel : Branch switching in bifurcation problems for ordinary differential equations. Numer. Math. 41, 93-116 (1983).
- [17] A. Spence and A.D. Jepson : The numerical calculation of cusps, bifurcation points and isolas formation points in two parameter problems. These proceedings.

- [18] A. Spence and B. Werner : Nonsimple turning points and cusps. IMA J. Numer. Anal. 2, 413-427 (1982).
- [19] H. Stetter : Asymptotic expansions for the error of discretization algorithms for non-linear functional equations. Numer. Math. 7, 18-31 (1965).
- [20] H. Weber : Shooting methods for bifurcation problems in ordinary differential equations. In : H.D. Mittelmann, H. Weber (eds.), Bifurcation problems and their numerical solution, ISNM Vol. 54, pp. 185-210. Basel : Birkhäuser-Verlag 1980.
- [21] E.C. Zeeman : Bifurcation, Catastrophe and Turbulence. In : P.J. Hilton, G.S. Young (eds.), New directions in Applied Mathematics, pp. 109-153. New York : Springer-Verlag 1982.

Dirk Roose
Renaat Caluwaerts
Department of Computer Science
Katholieke Universiteit Leuven
Celestijnenlaan 200A
B - 3030 Leuven (Belgium)

On the Axisymmetric Buckling of Thin Spherical Shells.

R.Scheidl, Vienna.

I. Introduction:

In this work a post-buckling analysis of a thin spherical shell under the action of uniform external pressure is given. A similar investigation was done by Knightly & Sather ([1]) using a more functional analytic approach. Further in this paper a versal unfolding (see [2]) of the bifurcation equations is carried out. Nevertheless these results have a limited practical importance, as the range of validity is restricted to a small neighborhood of the bifurcation point. This is caused by the fact that the eigenvalues lie closely spaced. Lange & Kriegsmann ([3,4]) applied asymptotic integration techniques to overcome these troubles. They obtained results, which are valid at larger distances of the bifurcation point and which give together with the classical bifurcation solution a nearly total understanding of the complete axisymmetric buckling behavior of the perfect shell. I will shortly comment on this at the end of the paper.

II. Shell Equations:

The basic equations governing the finite axisymmetric deformations of thin shells of revolution were formulated by E.Reissner ([5]). We make some reasonable simplifications and assumptions as we restrict ourselves to small but finite deflections and consider the so called dead-loading. This type of loading which directs always to the center of the shell can be hardly realized experimentally. The alternative type would be a real pressure loading pointing normal to the deformed shell surface and usually occurring in praxis. But it leads to mathematical complications and therefore we impose the dead loading type like many other authors have done also ([3,6,7]).

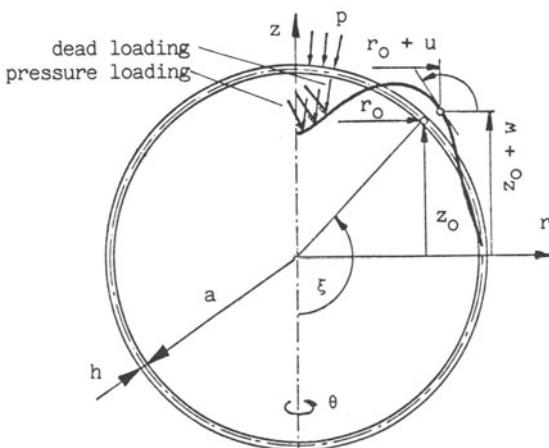


Fig.1: Geometry and loading of the shell.

The equations are, if nonlinearities up to the third order are retained:

$$\begin{aligned} \beta'' + \beta' \operatorname{ctg} \xi + \beta (2\lambda/\delta - \underline{\nu} - \operatorname{ctg}^2 \xi) + \underline{\psi}/\delta = \\ \underline{\beta^2 (\frac{3-\nu}{2}) \operatorname{ctg} \xi} + \underline{\beta \psi \operatorname{ctg} \xi / \delta} - \underline{\beta^3 [\frac{2}{3} (\operatorname{ctg}^2 \xi - 1)]} + \\ \underline{\sin \xi (1+\nu)/6} - \underline{\lambda/3\delta} + \underline{\beta^2 \psi / 2\delta} \end{aligned} \quad (1)$$

$$\begin{aligned} \psi'' + \psi' \operatorname{ctg} \xi + \psi [\underline{\nu} - \operatorname{ctg}^2 \xi] - \underline{\beta/\delta} + \underline{2\lambda\beta [1-\nu]} = \\ \underline{\psi [\beta \operatorname{ctg} \xi (2+\nu) + \nu \beta']} - \underline{\beta^2 \operatorname{ctg} \xi / 2\delta} + \underline{2\lambda [\nu \beta \beta' + \beta^2 \operatorname{ctg} \xi]} \\ + \underline{\psi [\nu \beta^2 / 2 - \beta \beta' \nu \operatorname{ctg} \xi + \beta^2 - \beta^2 \operatorname{ctg} \xi]} - \underline{\beta^3 / 6\delta} + \underline{\lambda \beta^3 (4-\nu) / 3} \end{aligned} \quad (2)$$

with boundary conditions: $\beta(0)=\beta(\pi)=\psi(0)=\psi(\pi)$ (3)

and with: $()' = \frac{d}{d\xi}$, $\psi = (\hat{\psi} - \hat{\psi}_0)/a^2 p_k$, $p_k = \frac{E(h/a)^2}{\sqrt{12(1-\nu^2)}}$ (4, 5)

$$\lambda = \frac{p}{4p_k}, \quad \delta = \frac{h}{a} \sqrt{\frac{1}{12(1-\nu^2)}}, \quad \hat{\psi}_0 = -\frac{a^2 p_k}{4} \sin 2\xi \quad (6, 7, 8)$$

$\delta = \xi - \phi$ is a deformation variable and gives the tangent rotation. ξ is the azimuthal angle (see also Fig. 1). ψ is a dimensionless form of the stress function $\hat{\psi} = r_0 H$, with H being the horizontal component of the stress resultant in meridional direction. $\hat{\psi}_0$ is the value of $\hat{\psi}$ for the uniformly contracted shell. E is Young's modulus and ν Poisson's ratio, two material constants. The ratio of wall thickness and shell radius is contained in δ , which takes small values for thin walled shells. The load intensity in comparison with a certain critical value is expressed by λ .

For small values of δ we safely can omit the underlined terms, which yields:

$$\beta'' + \beta' \operatorname{ctg} \xi + \beta (2\lambda/\delta - \operatorname{ctg}^2 \xi) + \psi/\delta = \beta \psi \operatorname{ctg} \xi / \delta + \beta^2 \psi / 2\delta + \beta^3 \lambda / 3\delta \quad (9)$$

$$\psi'' + \psi' \operatorname{ctg} \xi - \psi \operatorname{ctg}^2 \xi - \beta/\delta = -\beta^2 \operatorname{ctg} \xi / 2\delta - \beta^3 / 6\delta \quad (10)$$

with boundary conditions (3).

III. Bifurcation Point:

Equations (9, 10) possesses the trivial solutions $\beta(\xi)=\psi(\xi)\equiv 0$ for all values of λ and δ . From Implicit Function Theorem follows, that bifurcation only can occur if the Fréchet derivative becomes singular.

Fréchet derivative in our example is given by the left hand side of equations (9,10). Some computations show, that the Legendre-Polynomials

$$\beta = A_n P_n^1(\cos \xi) , \quad \psi = B_n P_n^1(\cos \xi) \quad (11,12)$$

are eigenfunctions of the corresponding eigenvalue problem. An eigenvalue zero occurs if the following equation is satisfied:

$$\lambda = \lambda_n = (\delta \mu_n + 1 / \delta \mu_n) / 2 \quad \text{with} \quad \mu_n = n(n+1)-1 \quad (13,14)$$

$$\text{and} \quad B_n = -A_n / \mu_n \delta \quad (16)$$

In order to detect the smallest value of λ we treat n as a continuous variable, which leads to the rough condition:

$$\mu_n = 1 / \delta \quad (15)$$

The situation for the eigenvalues is shown in Fig.2, where different eigenvalue curves λ_n are plotted in a λ - δ diagram. We can see that in general a simple eigenvalue occurs. But there exist special values of δ for which double eigenvalues appear.

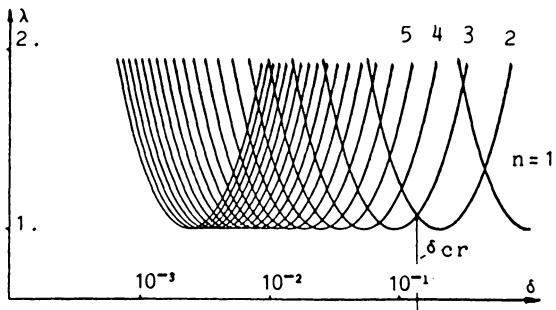


Fig.2: Eigenvalue curves in a δ - λ diagram.
 δ : thickness/radius; λ : loadfactor

IV. Bifurcating Solution:

IV.1 Liapunov-Schmidt Method:

Liapunov-Schmidt method provides a mean for largely reducing the dimension of a bifurcation problem. (For details see [8,9,10]). The result is a system of bifurcation equations giving the complete solution of the full system in a neighborhood of the bifurcation point. This reduction procedure requires a decomposition of the variables into a part belonging to the kernel of the linear operator (critical variables) and a part which is an element of its orthogonal complement (non-critical variables). Using the regular part of the operator the non-critical variables can be expressed uniquely in terms of the critical variables.

Introducing these expressions into the singular part of the operator yields

the so called bifurcation equations.

IV.2 Determinacy and Universal Unfolding of Bifurcation Equations.

Usually Liapunov-Schmidt process is carried out expanding equations in power series, which leads of course to power series representation of the bifurcation equations. Truncating these power series can be justified using the concept of determinacy as developed in Catastrophe Theory (|2|). Bifurcation equations are called n-determinate if terms of order $n+1$ or higher do not influence the local behavior.

Karman and Tsien firstly discovered the decisive influence of imperfections on the stability behavior of shells (|11|). Such imperfections result in constant terms in the bifurcation equations. They are fixed parts of every unfolding of any degenerate (bifurcation) equation. Such universal unfoldings for a lot of practical important cases can be found in the famous list of the Seven Elementary Catastrophes (|2|).

IV.3 Spectral Representation of the Operator.

In order to carry out this Liapunov-Schmidt reduction process we give a spectral representation of our equations. We make an ansatz for the both unknowns β and ψ in terms of series of eigenfunctions:

$$\beta = \sum_k a_k P_k^1(\cos \xi), \quad \psi = \sum_k b_k P_k^1(\cos \xi) \quad (17, 18)$$

Introducing this ansatz to our shell equations and projecting on each of the eigenfunctions

$$\int_0^\pi L_j(\beta, \psi, \lambda) P_i^1(\cos \xi) \sin \xi d\xi = \int_0^\pi N_j(\beta, \psi, \lambda) P_i^1(\cos \xi) \sin \xi d\xi \quad (19)$$

$$j = 1, 2; i = 1, 2, 3, \dots$$

- L_j and N_j are the linear and nonlinear parts of the equations - yields an infinite dimensional system of algebraic equations for the coefficients a_i, b_i :

$$(2\lambda/\delta - \mu_i) a_i + b_i/\delta = \sum_{k,l} r_{ikl} a_k b_l + \sum_{k,l,m} r_{iklm} (\frac{1}{2} a_k a_l b_m + \frac{\lambda}{3} a_k a_l a_m) \quad (20)$$

$$-a_i/\delta - \mu_i b_i = -\frac{1}{2} \sum_{k,l} r_{ikl} a_k a_l - \frac{1}{6} \sum_{k,l,m} r_{iklm} a_k a_l a_m \quad (21)$$

with the coefficients r_{ikl}, r_{iklm} given by:

$$r_{ikl} = \frac{1}{\delta} \frac{2i+1}{2i(i+1)} \int_0^{\pi} P_k^1(\cos\xi) P_l^1(\cos\xi) P_i^1(\cos\xi) \cos\xi d\xi \quad (22)$$

$$r_{iklm} = \frac{1}{\delta} \frac{2i+1}{2i(i+1)} \int_0^{\pi} P_k^1(\cos\xi) P_l^1(\cos\xi) P_m^1(\cos\xi) P_i^1(\cos\xi) \sin\xi d\xi \quad (23)$$

From equation (21) b_i can be computed in terms of a_j . Introducing this to the equation (20) we get a system of equations for the a_i only:

$$\begin{aligned} \omega_i a_i &= -\frac{1}{\delta} \sum_{k,l} r_{ikl} (1/2\mu_i + 1/\mu_1) a_k a_l + \sum_{k,l,m} a_k a_l a_m [-r_{iklm}/6\delta\mu_i + \\ &\quad \frac{1}{2} \sum_j r_{ikj} r_{jlm}/\mu_j - r_{iklm}/2\delta\mu_m + r_{iklm}\lambda/3] \end{aligned} \quad (24)$$

with $\omega_i = 2\lambda/\delta - \mu_i - 1/\delta^2\mu_i$, which is the i -th eigenvalue. (25)

At a bifurcation point one (or two in the case of a double eigenvalue) of these ω_i become zero. ($\lambda = \lambda_\alpha (= \lambda_{\alpha+1}) + \omega_\alpha (\omega_{\alpha+1}) = 0$). Therefore the equation with index $\alpha(\alpha+1)$ in the system (25) has vanishing linear part and constitutes the singular part of the operator. From the regular part of the operator (equations in (25) with index $i \neq \alpha(\alpha+1)$) we can compute the a_i in terms of the $a_\alpha (a_{\alpha+1})$.

$$a_i = f_i(a_\alpha (a_{\alpha+1})) = \frac{1}{\delta} \sum_{k,l=\alpha}^{(\alpha+1)} r_{ikl} a_k a_l (1/2\mu_i + 1/\mu_1) + O(|a_k + a_l|^3) \quad (26)$$

V. Unfolded Bifurcation Equations:

V.1 Simple Eigenvalue:

The bifurcation equation in the simple eigenvalue case is:

$$\omega_\alpha a_\alpha = k_2 a_\alpha^2 + k_3 a_\alpha^3 \quad (27)$$

$$\text{with } k_2 = -r_{\alpha\alpha\alpha} 3/2\delta\mu_\alpha \quad (28)$$

$$\begin{aligned} \text{and } k_3 &= \sum_i r_{\alpha\alpha i} r_{i\alpha\alpha} (1/2\mu_i + 1/\mu_\alpha) (2/\mu_\alpha + 1/\mu_i) / \delta^2 \omega_i \\ &\quad i \neq \alpha \\ &+ \frac{1}{2} \sum_i r_{\alpha\alpha i} r_{i\alpha\alpha} / \mu_i - r_{\alpha\alpha\alpha\alpha} (2/3\delta\mu_\alpha - \lambda/3) \end{aligned} \quad (29)$$

There is a qualitative difference whether α is even or odd.

α even: As $k_2 \neq 0$ the bifurcation equation is 2-determinate and of type fold catastrophe ([2]). Unfolding requires a so called imperfection term. (For physical meaning see appendix).

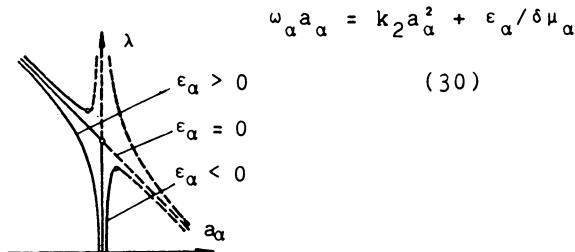


Fig.3: solutions for even order

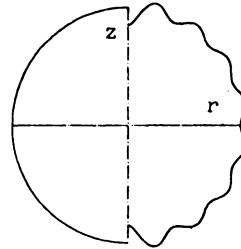


Fig.4: buckling pattern for even order

α odd: Here $k_2 = 0$. Therefore we have to retain the third order term. This bifurcation equation is of type cusp catastrophe. The unfolded equation is:

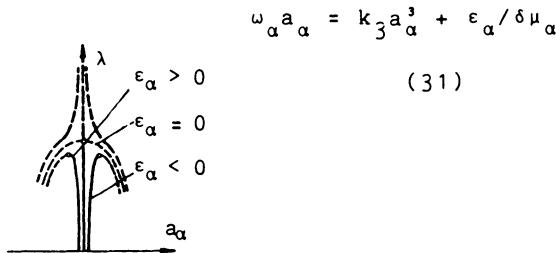


Fig.5: solutions for odd order

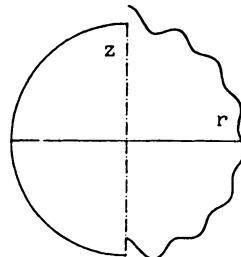


Fig.6: buckling pattern for odd order

As $k_3 < 0$ the bifurcating path exists for $\lambda < \lambda_{cr}$ and is unstable. Remarkably, if we would neglect the contribution of the f_i in the evaluation of k_3 it would have positive sign. This would be of course a qualitatively wrong result.

V.2 Double Eigenvalue Case:

The one-parameter unfolded bifurcation equations are, provided α is even:

$$a_\alpha \omega_\alpha = - \frac{3}{2\delta\mu_\alpha} [r_{\alpha\alpha\alpha} a_\alpha^2 + r_{\alpha\beta\beta} a_\beta^2 (\frac{1}{3} + \frac{2\mu_\alpha}{3\mu_\beta})] \quad (32)$$

$$a_\beta \omega_\beta = - \frac{3}{2\delta\mu_\alpha} r_{\beta\alpha\beta} (\frac{1}{3} + \frac{2\mu_\alpha}{3\mu_\beta}) a_\alpha a_\beta \quad (33)$$

They can be simplified, if we apply the approximate relations:

$$\mu_\alpha / \mu_\beta \approx 1, \quad r_{\alpha\beta\beta} / r_{\alpha\alpha\alpha} \approx r_{\beta\alpha\beta} / r_{\alpha\alpha\alpha} \approx 1, \quad \delta\mu_\alpha \approx 1 \quad (34)$$

to: $a_\alpha \Delta \tilde{\lambda} = -a_\alpha^2 - a_\beta^2$ (35)

$$a_\beta \Delta \tilde{\lambda} = -2a_\alpha a_\beta$$
 (36)

with $\Delta \tilde{\lambda} = 4(\lambda - \lambda_{cr})/3\delta r_{\alpha\alpha}$ (37)

These are bifurcation equations of type hyperbolic umbilic catastrophe (|2|). The versal unfolding requires imperfection terms in both equations and additionally a deviation $\Delta\delta$ of δ from its critical value δ_{cr} , for which the double eigenvalue occurs. (See Fig.2).

The unfolded equations are:

$$a_\alpha \Delta \tilde{\lambda} = -a_\alpha^2 - a_\beta^2 + \epsilon_\alpha / (\frac{3}{2} r_{\alpha\alpha})$$
 (38)

$$a_\beta (\Delta \tilde{\lambda} - \Delta \tilde{\delta}) = -2a_\alpha a_\beta + \epsilon_\beta / (\frac{3}{2} r_{\alpha\alpha})$$
 (39)

with : $\Delta \tilde{\delta} = 2\Delta\delta\alpha / (\frac{3}{2}\delta r_{\alpha\alpha})$, $\Delta \tilde{\lambda} = 2\Delta\lambda / (\frac{3}{2}\delta r_{\alpha\alpha})$ (40, 41)

$$\Delta\lambda = \lambda - \lambda_{cr}(\delta_{cr}) + \Delta\delta\alpha$$
 (42)

They can be transformed to the normal form of the unfolded hyperbolic umbilic catastrophe:

$$2ty = -3y^2 - x^2 - q$$
 (43)

$$2tx = -2xy - r$$
 (44)

by means of the following transformations:

$$y = \sqrt{3}a_\alpha - \Delta \tilde{\lambda}/2 - \Delta \tilde{\delta}/4 , \quad x = a_\beta$$
 (45, 46)

$$2t = -\Delta \tilde{\delta}\sqrt{3}/2 , \quad r = \epsilon_\beta / (\sqrt{27}r_{\alpha\alpha}/2)$$
 (47, 48)

$$q = \epsilon_\alpha (3r_{\alpha\alpha}/2) - \Delta \tilde{\lambda}^2/4 + \Delta \tilde{\delta}^2/16$$
 (49)

The advantage of this transformation is, that we immediately can make use of the bifurcation set given in Fig.7, where we have four distinct regions with qualitatively different solutions. With its help we also can classify a

set of bifurcation diagrams which present the solutions in a $\lambda-a_\alpha-a_\beta$ diagram. There are five classes of such structural stable bifurcation diagrams. A representative of each is shown in Fig.7. This classification is obtained evaluating the transformation equations (45-49). As $\Delta\lambda$ only influences the value of the parameter q , we get, varying $\Delta\lambda$, a straight line in the $q-r-t$ space oriented in the negative q -direction. The five different classes of bifurcation diagrams correspond to the five different possibilities for the intersection of these semi-rays with the four regions of different solutions.

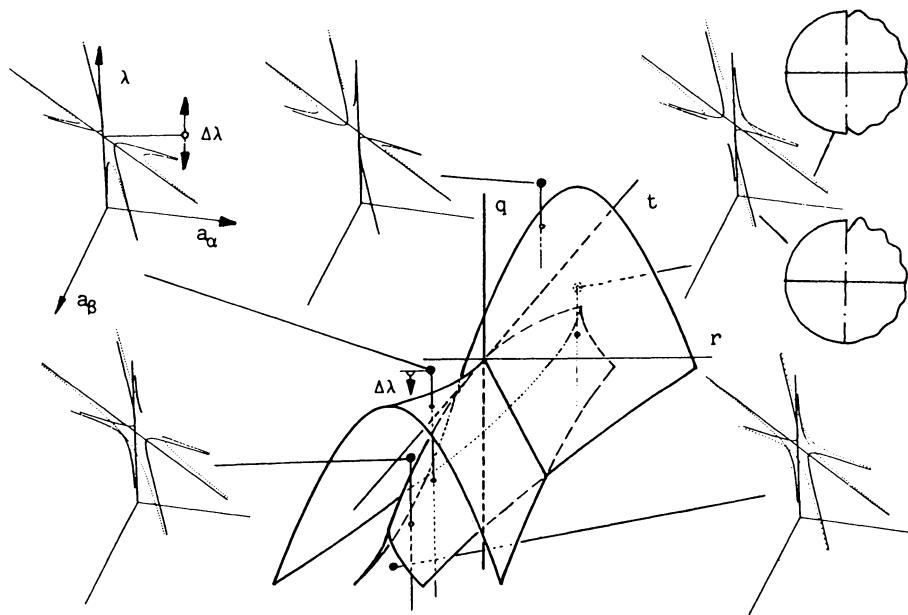


Fig.7: Bifurcation set in parameter space $q-r-t$. (hyperbolic umbilic catastrophe) and representation of bifurcation solutions in a $\lambda-a_\alpha-a_\beta$ diagram.

..... solutions for the imperfect shell ($\epsilon_\alpha, \epsilon_\beta \neq 0$)
 ——— solutions for the perfect shell ($\epsilon_\alpha, \epsilon_\beta = 0$)

VI. Remarks on the Global Bifurcation behavior.

In Fig.2 is shown, that for thin shells (small δ) the eigenvalues lie closely spaced. The Liapunov-Schmidt reduction process as carried out in the above approach is therefore valid only in a close vicinity of the bifurcation point. ($\Delta\lambda=0(\delta)$).

Lange and Kriegsmann ([3]) have applied asymptotic integration techniques to get a continuation of this classical solution, which is valid for $\Delta\lambda=0(1)$.

This extended result is tangent to the classical one at the bifurcation point, (see Fig.8), and there is a remarkable change in the bifurcation pattern (Fig.9): The wavy pattern of the classical solution tends to a dimple at one or both poles, whereas the residual region flattens out.

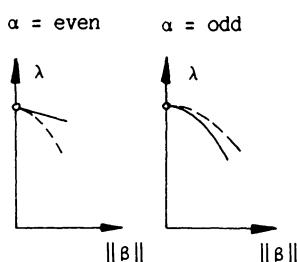


Fig.8: The extended solution
is tangent to the classical
solution.
classical
extended in |3|

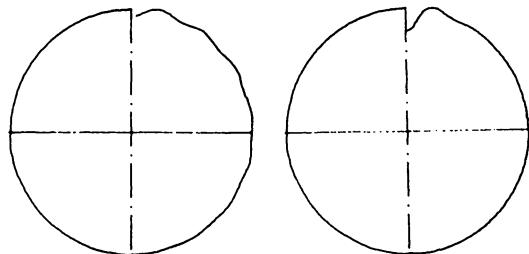


Fig.9: The classical buckling pattern
tends to a single dimple

In a second paper ([4]) both authors investigated large deflection states of spherical shells. For complete shells there exist deformation states consisting of a nearly undeformed region and a so called inverted cap at one or both poles. These two or three parts, respectively, are connected with boundary layers. The investigation is based on Reissner's equations with full nonlinearity, regarding both dead-loading and pressure-loading. Such deformation states exist only for λ is of order $O(\sqrt{\delta})$. It seems, that the extended solution provides the transition from the overall-deflection of the classical solution to this dimple solutions, which constitute the buckling patterns up to the deep post-buckling range. (see Fig.10,11 and for details [12]).

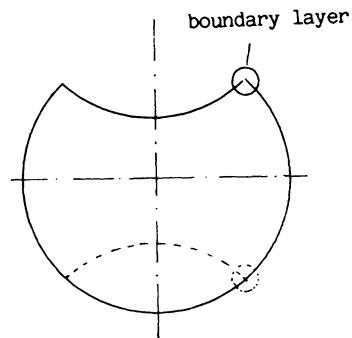


Fig.10: Single (double) dimple
solution as a combination of
an inverted cap and an unde-
formed part.

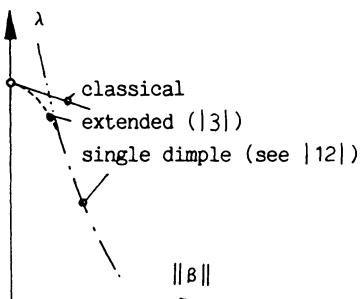


Fig.11: Post-buckling path up
to large deformations.

Appendix: Physical Meaning of Imperfection Terms.

There is a great variety of different physical reasons for imperfection, three of which are listed here:

- a slightly non uniform pressure,
- a not exact spherical shell middle surface,
- a non uniform wall thickness.

We have realized the last possibility. As h depends on ξ additional terms appear in the shell equations, which are given here up to quadratic terms:

$$\beta'' + \beta' \operatorname{ctg} \xi + \beta(2\lambda/\delta - \operatorname{ctg}^2 \xi) + \psi/\delta = \beta \psi \operatorname{ctg} \xi / \delta - 3(\delta'/\delta)(\beta' - v \beta \operatorname{ctg} \xi)$$

$$\psi'' + \psi' \operatorname{ctg} \xi - \psi \operatorname{ctg}^2 \xi - \beta/\delta = -\beta^2 \operatorname{ctg} \xi / 2\delta + 2\lambda(\delta'/\delta)(1 - v \cos \xi)$$

In the spectral representation this leads - if we only consider the lowest order additional term - to a constant in each equation

$$a_i(2\lambda/\delta - \mu_i - 1/\delta^2 \mu_i) = -\frac{1}{\delta} \sum_{k,l} r_{ikl} (2/\mu_i + 1/\mu_l) a_k a_l + \frac{1}{\delta \mu_i} \epsilon_i$$

which is approximately the projection of the thickness-variation on the corresponding eigenfunction.

$$\epsilon_i = \frac{1}{\delta} \frac{2i+1}{2i(i+1)} \int_0^\pi \delta'(\xi) 2\lambda(1 - v \cos \xi) P_i^1(\cos \xi) \sin \xi d\xi$$

References:

- 1 Knightly,G.H., Sather,D., Existence and Stability of Axisymmetric Buckled States of Spherical Shells,
- 2 Poston,T., Stewart,I., Catastrophe Theory and Its Applications, Pitman, London, 1978.
- 3 Lange,C.G., Kriegsmann,G.A., The Axisymmetric Branching Behavior of Complete Spherical Shells, Quarterly of Applied Mathematics, 39, 1981, 145-178.
- 4 Kriegsmann,G.A., Lange,C.G., On Large Axisymmetrical Deflection States of Complete Spherical Shells, J. of Elast., 10, 1980, 179-192.

- 5 Reissner,E., On Axisymmetrical Deformations of Thin Shells of Revolution, Proc. of Symposia in Appl. Math., 3, Amer. Math. Soc. (1950), 27-52.
- 6 Koiter,W.T., The Nonlinear Buckling Problem of a Complete Spherical Shell under Uniform External Pressure, Proc. K. ned. Akad. Wet., Series B, 72, 40, (1969).
- 7 Hutchinson,J.W., Imperfection Sensitivity of Externally Pressurized Shells, J.Appl.Mech., 34, (1967), 49-55.
- 8 Wainberg,M.M., Trenogin,W.A., Theorie der Lösungsverzweigung bei nicht-linearen Gleichungen, Akademie Verlag Berlin, 1973.
- 9 Sattinger,D.H., Bifurcation and Symmetry Breaking in Applied Mathematics, Bull. Amer. Math. Soc., 3, (1980), 779-819.
- 10 Chow,S.N., Hale, K.J., Methods of Bifurcation Theory, Springer, New York-Berlin-Heidelberg, 1982.
- 11 Karman, Th.von, Tsien, H.S., The Buckling of Spherical Shells by External Pressure, J. Aeron.Sci. 7, 43-45, 1939.
- 12 Scheidl, R., Troger, H., On the Buckling of Thin Spherical Shells, Proc. of Euromech. Coll. Nr.165, Munich, Springer 1983 in print, (Emmerling ed.).

Rudolf Scheidl, Institut für Mechanik, TU Wien, Karlsplatz 13, A-1040 Wien

ON THE RATE OF CONVERGENCE FOR THE
APPROXIMATION OF NONLINEAR PROBLEMS. *)

Reinhard Scholz

1.

In this paper we present an abstract theory to obtain error estimates in various norms for the approximation of solution branches of nonlinear equations by the aid of known estimates for corresponding linear problems.

In order to illustrate the results, we consider the following model problem. Let Ω be a bounded convex domain in \mathbb{R}^2 with sufficiently smooth boundary $\partial\Omega$. We are interested to approximate the solution of the boundary value problem

$$\begin{aligned} -\Delta u &= \lambda e^u \quad \text{in } \Omega , \\ u &= 0 \quad \text{on } \partial\Omega , \end{aligned} \tag{1.1}$$

where λ is a real parameter. It is well known that there exists a maximum value λ^* of the parameter λ such that Problem (1.1) has at least one solution $u \in H_0^1(\Omega) \cap L^\infty(\Omega)$; moreover there exists a unique solution $u^* \in H_0^1(\Omega) \cap L^\infty(\Omega)$ of (1.1) for $\lambda = \lambda^*$ and (u^*, λ^*) is a turning point.

Now let $\Gamma = \{(u(t), \lambda(t)) \mid |t| \leq t_0\}$ be a part of the solution branch such that $u(0) = u^*$, $\lambda(0) = \lambda^*$ holds. In order to

*) Summary of a joint paper with J. Descloux and J. Rappaz,
Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne,
Switzerland.

compute an approximation of Γ we consider a finite element method for discretizing Problem (1.1). If V_h is the finite element subspace of $H_0^1(\Omega)$ of piecewise linear polynomials with respect to a triangulation of Ω with meshsize $h > 0$, an approximation $u_h \in V_h$ of the solution of (1.1) is defined by

$$\int_{\Omega} \nabla u_h \cdot \nabla \varphi dx = \lambda \int_{\Omega} e^{u_h} \varphi dx \quad \text{for all } \varphi \in V_h . \quad (1.2)$$

Using the general results of Brezzi-Rappaz-Raviart [1] it is possible to prove that for $h \leq h_0$ there exists a branch of solutions $\Gamma_h = \{(u_h(t), \lambda_h(t)) \mid |t| \leq t_0\}$ of Problem (1.2) such that

$$\lim_{h \rightarrow 0} \sup_{|t| \leq t_0} \left\{ \|u(t) - u_h(t)\|_{H^1(\Omega)} + |\lambda(t) - \lambda_h(t)| \right\} = 0 .$$

Moreover error estimates for $|\lambda(t) - \lambda_h(t)|$ and $\|u(t) - u_h(t)\|_{H^1(\Omega)}$ are obtained, but using this theory it is not possible to get optimal error estimates for $u(t) - u_h(t)$ in the L^2 -norm or the L^∞ -norm, e.g.

In the following section we state an abstract result which permits to obtain error estimates in various norms for the approximation of solutions of nonlinear equations. (The proof can be found in [3].) This result, for instance, can be applied to the example above to obtain optimal L^2 - and L^∞ -estimates and to the Navier-Stokes problem using the "stream-function formulation" to derive optimal error estimates in the H^1 -norm.

2.

In this section, V and W will represent real Banach spaces; $L(W,V)$ is the space of bounded linear operators from W to V . Let T and T_h belong to $L(W,V)$, $G : V \times \mathbb{R} \rightarrow W$ be a nonlinear C^p -mapping with $p \geq 2$; h denotes a positive parameter the values of which have an accumulation point at 0. In a neighborhood of a point $(u_0, \lambda_0) \in V \times \mathbb{R}$, we consider the nonlinear equations

$$F(u, \lambda) = 0, \quad F_h(u, \lambda) = 0, \quad (2.1)$$

where F and $F_h : V \times \mathbb{R} \rightarrow V$ are the nonlinear mappings defined by

$$F(u, \lambda) = u + TG(u, \lambda), \quad F_h(u, \lambda) = u + T_h G(u, \lambda). \quad (2.2)$$

We first suppose:

$$a) \quad F(u_0, \lambda_0) = 0, \quad (2.3)$$

$$b) \quad T \text{ is a compact operator}, \quad (2.4)$$

$$c) \quad \lim_{h \rightarrow 0} \|T - T_h\| = 0. \quad (2.5)$$

Denoting by $F'(u, \lambda) \in L(V \times \mathbb{R}, V)$, by $D_u F(u, \lambda) \in L(V, V)$ and by $D_\lambda F(u, \lambda) \in L(\mathbb{R}, V)$ respectively the total derivative of F at (u, λ) and the partial derivatives of F with respect to u and λ , we remark that Hypothesis (2.4) implies that $D_u F(u, \lambda)$ is a Fredholm operator of index 0 and consequently that $F'(u, \lambda)$ is a Fredholm operator of index 1.

We next suppose either that $D_u F(u_0, \lambda_0)$ is an isomorphism from V into itself or that $D_u F(u_0, \lambda_0)$ has a kernel of dimension 1 and $D_\lambda F(u_0, \lambda_0)$ does not belong to the range of

$D_u F(u_o, \lambda_o)$. In the first case (u_o, λ_o) is a "regular point"; in the second case (u_o, λ_o) is a "simple limit point". This assumption can be written simply as

$$d) \quad \text{Range } F'(u_o, \lambda_o) = V . \quad (2.6)$$

The following result can be found in Descloux-Rappaz [2].

THEOREM 1:

Under Hypotheses (2.1) - (2.6), there exist $h_o > 0$ and a neighborhood of $(u_o, \lambda_o) \in V \times \mathbb{R}$ such that for $h \leq h_o$ and in this neighborhood, each of the equations $F(u, \lambda) = 0$ and $F_h(u, \lambda) = 0$ possess an unique branch of solutions. These branches can be parametrized as

$(u(t), \lambda(t)), (u_h(t), \lambda_h(t))$, $|t| \leq t_o$, $t_o > 0$, with the following properties:

- a) $(u(t), \lambda(t))$ and $(u_h(t), \lambda_h(t))$ are of class C^p ;
 $(u(0), \lambda(0)) = (u_o, \lambda_o)$; $u'(0) \neq 0$;
- b) $\lim_{h \rightarrow 0} \sup_{|t| \leq t_o} \left\{ \|u^{(k)}(t) - u_h^{(k)}(t)\|_V + |\lambda^{(k)}(t) - \lambda_h^{(k)}(t)| \right\} = 0$,
 $k = 0, 1, \dots, p-1$,
where $u^{(k)}, \lambda^{(k)}, \dots$ are the k -th derivative of u, λ, \dots ;
- c) there exists a constant C such that, for
 $|t| \leq t_o$, $h \leq h_o$, $k=0, \dots, p-1$,
we have

$$\begin{aligned} & \|u^{(k)}(t) - u_h^{(k)}(t)\|_V + |\lambda^{(k)}(t) - \lambda_h^{(k)}(t)| \\ & \leq C \sum_{l=0}^k \left\| \frac{d^l}{dt^l} F_h(u(t), \lambda(t)) \right\|_V . \end{aligned} \quad \#$$

The purpose is to derive error estimates for $u^{(k)}(t) - u_h^{(k)}(t)$ in a norm different from $\|\cdot\|_V$. Let H be a Banach space for which we suppose

- e) $V \subset H$ with continuous injection; (2.7)
- f) there exists a constant C such that along the solution branch of the exact problem $(u(t), \lambda(t))$ defined by Theorem 1, we have

$$\|D_\lambda^{k-1} D_u^1 G(u(t), \lambda(t)) [v_1, \dots, v_k]\|_W \leq C \|v_1\|_H \prod_{i=1}^{k-1} \|v_i\|_V \quad (2.8)$$

for $|t| \leq t_o, k=1, \dots, p-1$, $1 \leq k \leq k$ and for all $v_1, \dots, v_k \in V$.

The main result is contained in the following

THEOREM 2:

We assume that Hypotheses (2.1) - (2.8) are satisfied. Then, for the exact and the approximate branches of solutions defined by Theorem 1, there exist constants t_o, h_o, C and parametrizations $(u(t), \lambda(t)), (u_h(t), \lambda_h(t))$, such that for $k=0, \dots, p-2$, $|t| \leq t_o$, $h \leq h_o$ we have

$$\|u^{(k)}(t) - u_h^{(k)}(t)\|_H + |\lambda^{(k)}(t) - \lambda_h^{(k)}(t)|$$

$$\leq C \left\{ \sum_{l=0}^k \left\| \frac{d^l}{dt^l} F_h(u(t), \lambda(t)) \right\|_H + \|F_h(u(t), \lambda(t))\|_V^2 \right\} \#$$

REMARK 1:

In Theorem 1 and 2, the parametrizations of the exact respectively of the approximate solution branch are not necessarily identical; however, in the proofs we can show that it is possible to choose the same one.

REMARK 2:

In general, Theorem 2 gives better bounds for $|\lambda^{(k)}(t) - \lambda_h^{(k)}(t)|$ than Theorem 1.

REMARK 3:

In many examples, Hypothesis (2.8) will be verified by using regularity properties of the solutions of the exact problem.

REMARK 4:

Hypotheses (2.4) and (2.5) can be weakened by using results of Descloux-Rappaz [2].

REFERENCES

- [1] BREZZI, F., RAPPAZ, J., RAVIART, P.A.:
Finite dimensional approximation of nonlinear
problems.
Part I: Branches of nonsingular solutions.
Numer. Math. 36 (1980), 1-25.
Part II: Limit points.
Numer. Math. 37 (1981), 1-28.
Part III: Simple bifurcation points.
Numer. Math. 38 (1981), 1-30.
- [2] DESCLOUX, J., RAPPAZ, J.:
Approximation of solution branches of nonlinear
equations.
RAIRO, Anal. Numér. 16 (1982), 319-349.
- [3] DESCLOUX, J., RAPPAZ, J., SCHOLZ, R.:
On the rate of convergence for the approximation
of nonlinear problems.

Reinhard Scholz

Institut für Angewandte Mathematik
Albert-Ludwigs-Universität
Hermann-Herder-Str. 10
7800 Freiburg
Federal Republic of Germany

ALGORITHMS FOR FINITE-DIMENSIONAL TURNING POINT PROBLEMS FROM
VIEWPOINT TO RELATIONSHIPS WITH CONSTRAINED OPTIMIZATION METHODS

Hubert Schwetlick

The aim of this paper is to survey most of recent turning point algorithms and to show their relations to optimization methods which arise from the characterization of a turning point as solution of an equality constrained optimization problem.

1. Introduction

The modelling of certain real processes leads immediately or after appropriate discretization to nonlinear systems in finite dimensions of the type

$$G(x, t) = 0, \quad G : R^n \times R \rightarrow R^n. \quad (1)$$

In general, x denotes the vector of state variables whereas t plays the role of an additional parameter that has a special meaning for the underlying problem. For simplification of notation we write equation (1) in the equivalent form

$$G(u) = 0, \quad u := \begin{pmatrix} x \\ t \end{pmatrix} \quad (2)$$

and consider G as a function $G : R^{n+1} \rightarrow R^n$.

In the following we assume that

(A₁) there is a vector $u^* = (x^*, t^*)^T$ with

$$G(u^*) = 0$$

(A₂) the mapping G is sufficiently smooth and there holds

$$\text{rank } G_u(u^*) = n$$

where

$$G_u(u) = (\partial G_j(u)/\partial u_j) = (G_x(x, t) \mid G_t(x, t)) \quad (3)$$

denotes the $(n, n+1)$ -Jacobian of G with respect to u .

From (A₁) it follows that

$$\ker(G_u(u^*)) = \text{span}\{v^*\} \quad \text{with} \quad G_u(u^*)v^* = 0, \|v^*\| = 1; \quad (4)$$

the norm is always the Euclidean vector norm. Moreover, for sufficiently small $\delta > 0$ the solution set

$$\mathcal{L} := \{u \in \mathbb{R}^{n+1} : G(u) = 0, \|u - u^*\| < \delta\} \quad (5)$$

is a one-dimensional smooth arc that can be parametrized in the form

$$\mathcal{L} = \{u \in \mathbb{R}^{n+1} : u = z(\tau), \tau \in T\}$$

where T is an open interval, $z : T \rightarrow \mathbb{R}^{n+1}$ is a smooth function with

$$z(\tau^*) = u^*, \dot{z}(\tau^*) = \alpha^* v^* \text{ for some } \tau^* \in T \text{ with } \alpha^* \neq 0, \quad (6)$$

and the parametrization is regular in the sense of $\dot{z}(\tau) \neq 0 \forall \tau \in T$. As an example the parametrization defined by (2) and the additional equation

$$(v^*)^T(u - u^*) = \tau \quad (7)$$

can be considered. In this case we have $\tau^* = 0$ and $\alpha^* = 1$, see fig.1.

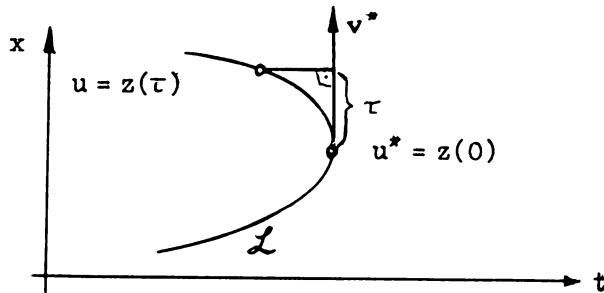


Fig.1. Parametrization of \mathcal{L} in the neighbourhood of u^*

The simplest critical points on the solution path \mathcal{L} are the turning points in which the path looks like in fig.1. Therefore we define a turning point u^* as solution of the equality constrained optimization problem

$$(TP) \quad \begin{aligned} t &= (e^{n+1})^T u \rightarrow \text{Extremum} \\ &\text{subject to } G(u) = 0. \end{aligned}$$

Solution paths with turning points arise, e.g., in nonlinear mechanics, see POSTON/STEWART[78], TROGER[83], in the description of chemical reacting systems, see RAY[77], BOHL[83], KUBICEK[83], and in the analysis of resistive electronic circuits, see CHUA/LIN[76].

In the last ten years various algorithms for computing turning points have been developed, analyzed, and implemented. Today most of real problems containing turning points can be solved by efficient and reliable algorithms the description of which is one of the aims of this paper. The other consists in showing the relations between the algorithms known from literature and the extremal formulation (TP) of a turning point. It turns out that almost all algorithms can be derived from this formulation in a very natural way.

Numerical tests are not contained in this survey. They can be found, e.g., in the paper of MELHEM/RHEINBOLDT[82]. For general remarks about solving equations with parameters and computing turning points we refer to the review articles SCHWETLICK[81,82].

2. First Order Turning Point Conditions

Let

$$L(u, \psi) := (e^{n+1})^T u - G(u)^T \psi, \quad L: \mathbb{R}^{n+1} \times \mathbb{R}^n \rightarrow \mathbb{R}^n \quad (8)$$

denote the Lagrangian of (TP) where $\psi \in \mathbb{R}^n$ is the vector of Lagrange multipliers.

Theorem 1. Let u^* be a turning point of G . Then there exists a unique $\psi^* \in \mathbb{R}^n$ such that the pair (u^*, ψ^*) satisfies the conditions

$$(L_0) \quad \nabla_\psi L(u^*, \psi^*) = G(u^*) = 0 ,$$

$$(L_1) \quad \nabla_u L(u^*, \psi^*) = e^{n+1} - G_u(u^*)^T \psi^* = 0.$$

Moreover, the Lagrange condition (L_1) is equivalent to each of the following conditions

the dual_zero_vector_condition

$$(DZ) \quad G_x(u^*)^T \psi^* = 0, \quad G_t(u^*)^T \psi^* = 1,$$

the tangent condition

$$(T) \quad (e^{n+1})^T v^* = 0, \quad G_u(u^*)v^* = 0, \quad \|v^*\| = 1,$$

the primal_zero_vector_condition

$$(PZ) \quad G_x(u^*)\psi^* = 0, \quad \|\psi^*\| = 1,$$

the eigenvalue condition

$$(E) \quad \lambda(G_x(u^*)) = 0$$

where $\lambda(A)$ denotes the absolutely smallest eigenvalue of the matrix A ,

the determinantal condition

$$(D) \quad \det(G_x(u^*)) = 0.$$

Proof. From Lagrange theory it is clear that (L_0) and (L_1) are necessary for u^* to be a solution of (TP). Since $G_u(u^*)$ has full rank the multipliers ψ^* are uniquely determined. Condition (DZ) is identical with (L_1) but written in a partitioned form, compare (3). On the other hand, (L_1) means that e^{n+1} is in the range of $G_u(u^*)^T$ or equivalently is orthogonal to the kernel of $G_u(u^*)$. Since the latter is spanned by v^* , see (4), we obtain (T). Moreover, v^* is the tangent direction of \mathcal{L} at u^* , see (6) and fig.1. If v^* is partitioned according to $v^* = (\psi^*, \alpha^*)^T$ the condition (T) is seen to be equivalent to $\alpha^* = 0$ and (PZ). Both (DZ) and (PZ) say that $G_x(u^*)$ is singular which is equivalent to (E) and (D).

3. Second Order Turning Point Conditions

Let

$$H(u, \psi) := \nabla_{uu} L(u, \psi)$$

denote the Hessian of L with respect to u . Then again Lagrange theory gives the following sufficient condition for u^* to be a solution of (TP), see e.g. GILL/MURRAY/WRIGHT[80].

Theorem 2. Let (u^*, ψ) satisfy the necessary first order conditions of Theorem 1. Then u^* is a strong local extremum of (TF) if the second order Lagrange condition

$$(L_2) \quad H(u^*, \psi^*) \text{ is definite on } \ker G_u(u^*)$$

is fulfilled. The condition (L_2) is equivalent to each of the following ones

$$(L_{2,1}) \quad (\psi^*)^T G_{uu}(u^*) v^* v^* \neq 0 ,$$

$$(L_{2,2}) \quad (\psi^*)^T G_{xx}(u^*) \psi^* \psi^* \neq 0 ,$$

$$(L_{2,3}) \quad (e^{n+1})^T \begin{pmatrix} G_u(u^*) \\ (v^*)^T \end{pmatrix}^{-1} \begin{pmatrix} G_{uu}(u^*) v^* v^* \\ 0 \end{pmatrix} \neq 0 .$$

We show only the equivalence of all conditions. Because of (4) condition (L_2) is equivalent to

$$(v^*)^T H(u^*, \psi^*) v^* \neq 0 . \quad (9)$$

Now there holds

$$g^T H(u, \psi) h = \psi^T G_{uu}(u) g h \quad \forall g, h \in \mathbb{R}^{n+1} \quad (10)$$

where G_{uu} is the second derivative of G with respect to u . By using (10) we immediately obtain $(L_{2,1})$. Due to $v^* = (\psi^*, 0)^T$ this is the same as $(L_{2,2})$. Further from (PZ) we conclude

$$e^{n+1} = G_u(u^*)^T \psi^* = (G_u(u^*)^T | v^*) \begin{pmatrix} \psi^* \\ 0 \end{pmatrix} = B(u^*, v^*)^T \begin{pmatrix} \psi^* \\ 0 \end{pmatrix} \quad (11)$$

where the matrix

$$B(u^*, v^*) := \begin{pmatrix} G_u(u^*) \\ \hline \dots \\ (v^*)^T \end{pmatrix} \quad (12)$$

is regular. Therefore we get

$$\begin{pmatrix} \psi^* \\ 0 \end{pmatrix} = B(u^*, v^*)^{-T} e^{n+1} .$$

Substituting this expression into $(L_{2,1})$ yields condition $(L_{2,3})$.

If at u^* the necessary first order conditions (L_0) ,

(L_1) as well as the sufficient second order condition (L_2) are satisfied then u^* is called a simple (or strong or regular) turning point. In the following we suppose u^* to be a simple turning point in the sense just defined.

4. The Principle of Direct Turning Point Methods

The basic idea of the direct turning point method consists in augmenting the underdetermined system (2) of size $(n, n+1)$ by an additional system

$$M(u, y) = 0 , \quad M : R^{n+1} \times R^{1-1} \rightarrow R^1 \quad (13)$$

of size $(1, n+1)$ eventually involving auxiliary variables y in order to get the system

$$F(u, y) := \begin{pmatrix} G(u) \\ M(u, y) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} , \quad F : R^{n+1} \times R^{1-1} \rightarrow R^{n+1} \quad (14)$$

of size $(n+1, n+1)$. The function M is to be chosen such that

(P₁) there exist a vector y^* with $F(u^*, y^*) = 0$,

(P₂) the derivative

$$F_{(u,y)}(u^*, y^*) = \begin{bmatrix} \frac{\partial G_u(u^*)}{\partial u} & 0 \\ \frac{\partial M_u(u^*, y^*)}{\partial u} & \frac{\partial M_y(u^*, y^*)}{\partial y} \end{bmatrix} \quad (15)$$

with respect to both u and y is regular.

Then (u^*, y^*) is an isolated solution of the so-called defining equations (14) and can be found by applying a method for solving regular nonlinear equations.

Depending on the choice of

(i) the augmenting equation $M(u, y) = 0$ and

(ii) the method for solving the inflated system $F(u, y) = 0$

various types of turning point methods can be derived. All methods known are of Newton-type and therefore quickly but only locally convergent. The latter fact, however, is not a disadvantage since arbitrarily good initial points can be computed by a path following algorithm, see ALLGOWER [81] or SCHWETLICK [82] for a review of such methods.

5. Direct Methods Using Second Derivatives

The most natural choice of the method for solving (14) is certainly Newton's method. Its k -th step has the form

$$u^{k+1} := u^k + \delta u^k, \quad y^{k+1} := y^k + \delta y^k \quad (16)$$

where the corrections $\delta u^k, \delta y^k$ are defined by the linear system

$$G(u^k) + G_u(u^k)\delta u^k = 0 \quad (17)$$

$$M(u^k, y^k) + M_u(u^k, y^k)\delta u^k + M_y(u^k, y^k)\delta y^k = 0. \quad (18)$$

Since the matrix of (17), (18) is the Jacobian $F_{(u,y)}(u^k, y^k)$ having the special structure (15) the system can be solved by the following 2-stage strategy

Step 1. Determine the general solution of the first equation (17) in the form

$$u^k = s^k + \lambda v^k, \quad \lambda \in \mathbb{R} \quad (19)$$

where s^k, v^k solve the linear systems

$$G(u^k) + G_u(u^k)s^k = 0, \quad G_u(u^k)v^k = 0 \text{ with } \|v^k\| = 1. \quad (20)$$

Note that s^k, v^k may be computed from the regular systems

$$B(u^k, r)s^k = - \begin{pmatrix} G(u^k) \\ 0 \end{pmatrix}, \quad B(u^k, r)\tilde{v}^k = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad (21)$$

by setting $v^k := \tilde{v}^k / \|\tilde{v}^k\|$. The vector $r \in \mathbb{R}^{n+1}$, $\|r\| = 1$, is such that $r^T v^* \neq 0$ with v^* from (T). Usually $r = r^k$ is chosen as

$$r^k := \pm e^{jk} \text{ with } |(e^{jk})^T v^{k-1}| \geq \omega \cdot \max_j |(e^j)^T v^{k-1}|$$

and a threshold $0 < \omega \leq 1$.

Step 2. Substitute (19) into the second equation (18). Solve the arising linear system

$$M(u^k, y^k) + M_u(u^k, y^k)s^k + (M_u(u^k, y^k)v^k)\lambda_k + M_y(u^k, y^k)\delta y^k = 0$$

for $(\lambda_k, \delta y^k)$ and set $\delta u^k := s^k + \lambda_k v^k$. (22)

In some special cases this strategy has been used by ABBOTT

[77,78] and PÖNISCH/SCHWETLICK [81].

If $l = 1$, i.e. if no auxiliary variables occur the function $M = M(u)$ becomes a scalar one and step 2 reduces to

$$\lambda = \lambda_k = - \frac{M(u^k) + \nabla_u M(u^k)^T s^k}{\nabla_u M(u^k)^T v^k} \quad (23)$$

where $\nabla_u M(u) = (\partial M(u)/\partial u_i)$ denotes the gradient of M .

5.1 Methods Based on the Determinantal Condition (D)

Set $l := 1$ and

$$M(u) := \det(G_x(u)) . \quad (24)$$

Use Newton's method for solving (14), see KUBÍČEK [75] and ABBOTT [77,78].

In this case the evaluation of the gradient $\nabla_u M$ requires to compute n determinants of order n so that this method can not be recommended for large n .

5.2 Methods Based on the Eigenvalue Condition (E)

Suppose that $G_u(u)$ is symmetric, set $l := 1$ and

$$M(u) := \lambda(G_x(u)) . \quad (25)$$

Since $\lambda=0$ is a simple eigenvalue of $G_x(u^*)$ the same is true for $\lambda(u) := M(u)$ if u is sufficiently near u^* . From eigenvalue perturbation theory one gets

$\nabla_u \lambda(u)^T h = \varphi(u)^T G_{xu}(u) \varphi(u) h = \varphi(u)^T G_{xx}(u) \varphi(u) \bar{h} + \varphi(u)^T G_{xt}(u) \varphi(u) h_{n+1}$
for all $h = (\bar{h}, h_{n+1}) \in \mathbb{R}^{n+1}$ where $\varphi(u)$ is the eigenvector of norm 1 belonging to $\lambda(u)$.

PAUMIER [81] replaces $\nabla_u \lambda$ by its component in direction $w(u) := (\varphi(u), 0)^T$, i.e. by

$$\widetilde{\nabla}_u \lambda(u) := k(u) w(u) \text{ where } k(u) := \varphi(u)^T G_{xx}(u) \varphi(u) \varphi(u). \quad (27)$$

Using this approximation equation (18) goes over into

$$\lambda(u^k) + k(u^k) w(u^k)^T \delta u^k = 0 \quad (28)$$

Then the iteration (17), (28) can be written as $u^{k+1} := \tilde{\Phi}(u^k)$, and $R := \tilde{\Phi}_u(u^*)$ characterizes the convergence behaviour. In PAUMIER's paper the relation

$$Ry = f_o(y)w^* \text{ with } w^* := (\varphi^*, 0)^T \quad (29)$$

and $f_o(y) := (\tilde{\nabla}_u \lambda(u^*) - \nabla_u \lambda(u^*))^T y / k(u^*)$ is derived where $f_o(w^*) = 0$ which implies that the spectral radius of R is zero and, hence, the method converges R-superlinearly. A sharper result can be obtained from the fact that $R^2 = 0$ which follows from (29). Therefore the convergence is 2-step quadratic in the sense of

$$\|u^{k+2} - u^*\| \leq Q \|u^k - u^*\|^2 \quad \forall k \text{ with } Q > 0$$

and, consequently, of R-order $\sqrt{2}$.

The method is realized by using an inverse iteration approach for computing the x-part of δu^k , see the original paper for details.

5.3 Methods Based on the Primal Zero Vector Condition (PZ)

Set $l := n+1$, $y := \varphi$, and

$$M(u,) := \begin{pmatrix} G_x(u) \\ q^T \varphi - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad M : \mathbb{R}^{n+1} \times \mathbb{R}^n \rightarrow \mathbb{R}^{n+1} \quad (30)$$

with some normalizing vector $q \in \mathbb{R}^n$, $\|q\| = 1$. Use Newton's method for solving (14), see SEYDEL [79] and MOORE/SPENCE [80].

In (30) the normalizing condition $\|\varphi\| = 1$ of (PZ) for the zero vector φ is replaced by the condition $q^T \varphi = 1$ that is easier to handle since it is linear in φ . The vector q is to be chosen such that $q^T \varphi^* \neq 0$, compare the remarks at the beginning of this chapter. In the original papers only the case $q = e_j$ is considered. MOORE/SPENCE showed that δu^k , δy^k can be obtained by solving 4 linear systems of order n with the same matrix. The 2-stage strategy described above also requires to solve 4 linear systems namely the two of (21) and two further system which are needed for solving equation (22), i.e. the system

$$\begin{pmatrix} G_x(u^k)\varphi^k \\ q^T \varphi^k - 1 \end{pmatrix} + \begin{pmatrix} G_{xu}(u^k)\varphi^k s^k \\ 0 \end{pmatrix} + \begin{pmatrix} G_x(u^k) & G_{xu}(u^k)\varphi^k s^k \\ -q^T & 0 \end{pmatrix} \begin{pmatrix} \delta\varphi^k \\ \lambda_k \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \quad (31)$$

by using the LR-factorization of $B(u^k, r)$ with $r := (q, 0)^T$.

5.4 Methods Based on the Tangent Condition (T)

By introducing $v = (\varphi, x)^T$ and r as above equation (30) may be written in the form

$$M_0(u, v) := \begin{pmatrix} G_u(u)v \\ r^T v - 1 \\ (e^{n+1})^T v - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \quad (32)$$

with $l := n+1$ and the auxiliary variables v . This is the analogon of the method 5.3 for (T) instead of (PZ). It is immediately seen that Newton's method for (2), (32) with the starting point (u^0, φ^0) generates the same sequence as for (2), (30) with the starting point (u^0, v^0) provided that $v^0 = (\varphi^0, 0)^T$ holds.

The linearity of (32) with respect to v can be used to eliminate the variables v . The first two equations are solved for v which results in

$$v = v(u) = v(u, r) = B(u, r)^{-1} e^{n+1} = \begin{pmatrix} G_u(u) \\ r^T \end{pmatrix}^{-1} e^{n+1}. \quad (33)$$

By substituting this expression into the third equation we obtain the condition

$$M(u) = (e^{n+1})^T v(u) = 0, \quad M: \mathbb{R}^{n+1} \rightarrow \mathbb{R} \quad (34)$$

that does not contain v explicitly, i.e. we have $l = 1$. Now Newton's method can be applied for solving (2), (34) which is just the method proposed by ABBOTT[78] and PÖNISCH/SCHWETLICK [81] with $r = e^j$ and with general r , respectively. It can be considered as a pre-eliminated version of the method based on (32). The gradient of (34) is defined by

$$\nabla_u M(u)^T h = - (e^{n+1})^T B(u, r)^{-1} \begin{pmatrix} G_{uu}(u)v(u)h \\ 0 \end{pmatrix}. \quad (35)$$

If v^k in step 1 is calculated from (21) then there holds

$$v(u^k) = \tilde{v}^k = v^k / r^T v^k ,$$

and step 2 gives

$$\lambda_k = \frac{(e^{n+1})^T v^k - (e^{n+1})^T a^k}{(e^{n+1})^T b^k} , \quad (36)$$

see (23), where a^k , b^k are defined as solutions of the systems

$$B(u^k, r)a^k = \begin{pmatrix} G_{uu}(u^k) v^k \\ s^k \\ 0 \end{pmatrix}, \quad B(u^k, r)b^k = \begin{pmatrix} G_{uu}(u^k) v^k \\ v^k \\ 0 \end{pmatrix} \quad (37)$$

which have the same matrix as those of (21). Therefore, as in 5.3, one Newton step for (2), (34) requires to solve 4 linear systems with the matrix $B(u^k, r)$.

It should be remarked that the idea to eliminate linearly occurring variables before applying Newton's method is, of course, not new and commonly used e.g. in the so-called variable projection methods for solving nonlinear regression problems with separated variables, see GOLUB/PEREYRA [73].

6. The Relation between the ABBOTT/PÖNISCH/SCHWETLICK method

and WILSON's Method for Constrained Optimization

In recent nonlinear optimization so-called Wilson-type methods have proved very successful, see e.g. GILL/MURRAY/WRIGHT [81]. In the special case of the problem (TP) Wilson's method consists in solving the nonlinear system $(L_0), (L_1)$ for (u^*, ψ^*) by applying Newton's method. In the context of chapter 5 this method can be considered as Newton's method for (14) with the augmenting equations

$$M(u, \psi) = \nabla_u L(u, \psi) = e^{n+1} - G_u(u)^T \psi = 0 , \quad (38)$$

$M: \mathbb{R}^{n+1} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$. This is, of course, the dual counterpart of the method of chapter 5.3 if there (PZ) is replaced by (DZ). If Newton's method for (2), (38) is realized via the 2-stage strategy the system (22) to be solved in step 2 becomes

$$e^{n+1} - G_u(u^k)^T \psi^k - H(u^k, \psi^k) s^k - \lambda_k H(u^k, \psi^k) v^k - G_u(u^k)^T \delta \psi^k = 0$$

since $M_u(u, \psi) = -H(u, \psi)$ and $M_\psi(u, \psi) = -G_u(u)^T$. Multiplying from the left by $(v^k)^T$ and considering (10) and (20) gives

$$(e^{n+1})^T v^k - (\psi^k)^T G_{uu}(u^k) v^k s^k - \lambda_k (\psi^k)^T G_{uu}(u^k) v^k v^k = 0$$

and, hence,

$$\lambda_k = \frac{(e^{n+1})^T v^k - (\psi^k)^T G_{uu}(u^k) v^k s^k}{(\psi^k)^T G_{uu}(u^k) v^k v^k} \quad (39)$$

In order to simplify the method that is a coupled iteration in the (u, ψ) -space we modify it in a way often used in optimization, namely by defining ψ^k as least squares solution of the linear system

$$e^{n+1} - G_u(u^k)^T \psi = 0 ,$$

i.e. the system (L_1) for fixed $u = u^k$. This leads to the estimate

$$\psi^k = [G_u(u^k)^T]^+ e^{n+1} \quad (40)$$

where A^+ denotes the Moore-Penrose inverse of A . Thus ψ^k is a function of u^k alone, and the iteration is decoupled.

If we define

$$z^k := \begin{pmatrix} \psi^k \\ \varsigma_k \end{pmatrix} \quad \text{with} \quad \varsigma_k := (e^{n+1})^T v^k$$

then it follows

$$\begin{aligned} B(u^k, v^k)^T z^k &= \left(G_u(u^k)^T | v^k \right) \begin{pmatrix} \psi^k \\ \varsigma_k \end{pmatrix} = G_u(u^k)^T \psi^k + \varsigma_k v^k = \\ &= G_u(u^k)^T (G_u(u^k)^T)^+ e^{n+1} + v^k (v^k)^T e^{n+1} . \end{aligned} \quad (41)$$

Because of the properties of the Moore-Penrose inverse there holds

$$G_u^T (G_u^T)^+ = (G_u^+ G_u)^T = G_u^+ G_u .$$

Now $G_u^+ G_u$ is the orthogonal projector on $R(G_u^T)$ which is the orthogonal complement of $\ker(G_u(u^k))$. Since the latter is spanned by v^k we get

$$G_u^T G_u = P_{R(G_u^T)} = I - P_{\ker(G_u)} = I - v^k (v^k)^T$$

Therefore, (41) leads to

$$B(u^k, v^k)^T z^k = e^{n+1}, \text{ i.e. } z^k = B(u^k, v^k)^{-T} e^{n+1}$$

and, finally, to

$$(\gamma^k)^T G_{uu}(u^k) v^k p = (e^{n+1})^T B(u^k, v^k)^{-1} \begin{pmatrix} G_{uu}(u^k) v^k p \\ 0 \end{pmatrix} \quad (42)$$

for all $p \in R^{n+1}$. With the definitions

$$\tilde{a}^k := B(u^k, v^k)^{-1} \begin{pmatrix} G_{uu}(u^k) v^k s^k \\ 0 \end{pmatrix}, \quad \tilde{b}^k := B(u^k, v^k)^{-1} \begin{pmatrix} G_{uu}(u^k) v^k v^k \\ 0 \end{pmatrix} \quad (43)$$

expression (39) now can be written as

$$\lambda_k = \frac{(e^{n+1})^T v^k - (e^{n+1})^T \tilde{a}^k}{(e^{n+1})^T \tilde{b}^k}. \quad (44)$$

By comparing (43), (44) with (37), (36) the following result is obtained, see SCHWETLICK [83].

Theorem 3. If in the method of ABBOTT/PÖNISCH/SCHWETLICK at the k -th step the normalizing direction is chosen as $r = v^k$ then this method is identical with Wilson's method for solving (TP) if there the least squares estimate (40) of the Lagrange multipliers γ^k is used.

7. Direct Methods with Approximated Second Derivatives

All the methods described in chapter 5 use the second derivative G_{uu} explicitly. The user has to provide subroutines for computing the partial derivatives $\partial^2 G_i / \partial u_i \partial u_j$, and these derivatives have to be evaluated once per step. Also if the subroutines are generated by an automatic differentiation routine alone the evaluation of all these subroutines is expensive, in general. Therefore it is sometimes desirable to approximate the terms containing second derivatives by first derivatives or function values only.

7.1 Approximations Using First Derivatives

In case $l = 1$ ABBOTT [77,78] proposed an approximation principle that is a block version of the well-known Brown-Brent technique. This principle can easily be generalized to augmenting equations of type

$$M(u, y) := K(u)y + g(u) = 0, \quad M : R^{n+1} \times R^{l-1} \rightarrow R^l \quad (45)$$

where $K(u)$ is a $(l, l-1)$ -matrix and $g(u)$ is an l -vector both smoothly depending on u . Note that all the functions M of chapter 5 are of this type.

The first step is the same as in the 2-stage strategy of chapter 5 and gives

$$u = u^k + s^k + \lambda v^k, \quad \lambda \in R \quad (46)$$

as solution of (45). In the second step, (46) is substituted for u in (45) resulting in

$$N^k(\lambda, y) := M(u^k + s^k + \lambda v^k, y) = 0, \quad N^k : R \times R^{l-1} \rightarrow R^l. \quad (47)$$

Now (47) is linearized at $(\lambda, y) = (0, y^k)$ with respect to λ in the form

$$N^k(0, y^k) + [(N^k(M_k, y^k) - N^k(0, y^k)) / M_k] \lambda_k + K(u^k + s^k) \delta y^k = 0. \quad (48)$$

Then (48) is solved for $(\lambda_k, \delta y^k)$, and the next iterate is defined as in chapter 5. Note that the second term of (48) is a difference approximation to $N_\lambda^k(0, y^k)(\lambda - 0)$ using the stepsize $M_k \neq 0$ whereas the third one is exactly $N_y^k(0, y^k)\delta y^k$. In case $l = 1$ equation (48) reduces to

$$\lambda_k = - \frac{N^k(0, y^k)}{(N^k(M_k, y^k) - N^k(0, y^k)) / M_k} \quad (49)$$

a formula that can be found in ABBOTT's papers. Let us note that the evaluation of G_{uu} is avoided on costs of 2 extra evaluations of G_u at $u^k + s^k$ and $u^k + s^k + M_k v^k$ since M depends on G_u . Let us further remark that the 2-stage strategy of chapter 5 is just the method described here but applied to the linearized equation $M(u, y) = 0$.

Another possibility for approximating $F(u, y)$ proposed by GRIEWANK/REDDIEN [83] is to take the exact Jacobian of G but a Broyden update for the Jacobian of M .

7.2 Approximations Using Function Values

PÖNISCH/SCHWETLICK [81] introduced the following approximation principle for the method of chapter 5.4. From (37) it is seen that G_{uu} there occurs only within the directional derivative

$$G_{uu}(u^k)v^k p^k \quad \text{with } p^k = v^k \text{ or } p^k = s^k / \|s^k\| . \quad (50)$$

Because of $G_{uu}vs = \|s\|G_{uu}v(s/\|s\|)$ the scaling of s^k is no restriction but it makes the arguments v^k , p^k of G_{uu} of equal norm. Now the terms (50) are approximated by the differences

$$\delta^2 G(u^k, v^k, p^k, \mu_k) = \frac{1}{\mu_k} [G(u^k + \mu_k v^k) - G(u^k + \mu_k v^k - \mu_k p^k) + G(u^k - \mu_k p^k) - G(u^k)] \quad (51)$$

using the stepsize $\mu_k \neq 0$. This approach requires only 4 extra evaluations of G .

8. The Principle of Indirect Turning Point Methods

In contrast to the direct turning point methods the indirect methods use a local parametrization

$$u = z^k(\tau), \quad \tau \in T_k \quad (52)$$

of the solution path \mathcal{L} in the neighbourhood of the current iterate u^k . Without loss of generality we assume that

$$z^k(0) = u^k \quad \text{and} \quad z^k(\tau_k^*) = u^*$$

holds for some $\tau_k^* \in T_k$. In the following we restrict ourselves to the parametrization defined by (2) and

$$(r^k)^T(u - u^k) = \tau \quad (53)$$

where $r = r^k$ with $\|r^k\| = 1$ is chosen as in chapter 5; see MENZEL/SCHWETLICK [78] for the history of this parametrization. Substituting (52) into (2), (53) and differentiating with respect to τ yields

$$B(z^k(\tau), r^k) \dot{z}^k(\tau) = e^{n+1}, \text{ i.e. } \dot{z}^k(\tau) = v(z^k(\tau), r^k) \quad (54)$$

with $v(u, r)$ from (33). For other parametrizations we refer to KELLER [77, 78] and MENZEL/SCHWETLICK [83].

If the turning point is characterized by (TP) and the parametrization (52) is substituted into this formulation then it reduces to the scalar problem

$$f_k(\tau) := (e^{n+1})^T z^k(\tau) \rightarrow \text{Extremum} \quad (55)$$

since $G(z^k(\tau)) = 0$ is identically fulfilled. By differentiating (55) the necessary condition

$$\dot{f}_k(\tau) = (e^{n+1})^T \dot{z}^k(\tau) = (e^{n+1})^T v(z^k(\tau), r^k) = 0 \quad (56)$$

is obtained. Obviously (56) is equivalent to the condition that arises if (52) is substituted into the equations (2), (34) by which the turning point is characterized in chapter 5.4.

Now the k -th step of an indirect method has the following structure

Step 1. Perform one step of an appropriate method for solving the scalar problem (55) or (56) in order to get an approximation τ_{k+1} for τ_k^* .

Step 2. Compute $u^{k+1} := z^k(\tau_{k+1})$ as solution of (2), (53) for $\tau = \tau_{k+1}$.

9. Interpolatory Methods for Determining τ_{k+1}

The basic idea is to interpolate f_k or \dot{f}_k at the nodes $\tau_{k,j} := (r^k)^T (u^{k-j} - u^k)$ that correspond to the preceding points u^{k-j} ($j=0, 1, \dots$) by a polynomial p_k and to solve (55) or (56) for the approximating polynomial in order to get τ_{k+1} .

9.1 Methods Based on (55)

(i) Define the cubic polynomial p_k by the conditions

$$f_k(\tau_{k,j}) = p_k(\tau_{k,j}), \quad \dot{f}_k(\tau_{k,j}) = \dot{p}_k(\tau_{k,j}) \quad (j=0, 1).$$

Take τ_{k+1} as the absolutely smallest root of the quadratic equation $\dot{p}_k(\tau) = 0$, see PÖNISCH/SCHWETLICK [82].

(ii) Define the quadratic polynomial p_k by

$$f_k(\tau_{k,j}) = p_k(\tau_{k,j}) \quad (j=0,1,2)$$

Take τ_{k+1} as zero of the linear function $p_k(\tau)$, see PÖNISCH[79].

9.2 Methods Based on (56)

(iii) Interpolate \dot{f}_k by a linear function p_k according to

$$\dot{f}_k(\tau_{k,j}) = p_k(\tau_{k,j}) \quad (j=0,1)$$

Take τ_{k+1} as zero of $p_k(\tau)$, see KELLER [78] and PÖNISCH [79].

(iv) As (iii) but with $v(u,r)$ in (56) replaced by

$$\hat{v}(u) := \pm v(u,r)/\|v(u,r)\|$$

where the sign is chosen such that $\det B(u, v(u)) > 0$, see RHEINBOLDT [82] and MELHEM/RHEINBOLDT [82].

Other methods can be derived by replacing (56) by the condition

$$M(z^k(\tau)) = 0 \tag{57}$$

where M is one of the functions (24) or (25), see again PÖNISCH[79].

It should be noted that because of (54) all the information necessary for computing τ_{k+1} in the methods (i) - (iii) can easily be obtained from u^{k-j} and the corresponding tangent directions v^{k-j} , see PÖNISCH/SCHWETLICK [82] for details.

10. Predictor-Corrector Schemes for Computing $u^{k+1} = z^k(\tau_{k+1})$

The common way for computing u^{k+1} as solution of (2), (53) for $\tau = \tau_{k+1}$ is to use the following PC-scheme.

Predictor Step. Approximate $z^k(\tau)$ by a simple function $d^k(\tau)$, e.g. by the tangent at u^k defined by

$$d^k(\tau) := u^k + v^k / [(r^k)^T v^k], \quad v^k \text{ as in (20),(21),}$$

see KELLER [78], PÖNISCH/SCHWETLICK [82] and others, or by the secant between u^k and u^{k-1} that is defined by

$$d^k(\tau) := u^k + \tau(u^k - u^{k-1}) / [(r^k)^T(u^k - u^{k-1})] ,$$

see RHEINBOLDT [82] and MELHEM/RHEINBOLDT [82].

Determine the predictor point $u^{k,0} = z^k(\cdot|_{k,0})$ so that $(r^k)^T(u^{k,0} - u^k) = \tau_{k+1}$. In the 2 examples one gets $\tau_{k,0} = \tau_{k+1}$.

Corrector Step. Take $u^{k,0}$ as starting point for Newton's method for solving (2), (53). Of course, Newton's method may be replaced by the modified Newton method or a quasi-Newton method.

11. Further Problems

There are some problems which could not be treated in this survey, namely

(i) How to solve linear systems with the matrix $B(u, r)$ in case that G_x is symmetric, sparse, and of special structure, see KELLER [77,78], BATOZ/DHATT [79], RHEINBOLDT [81], PAUMIER [81], CHAN [83].

(ii) How to compute curves of turning points and their cusps when G has two parameters, see RHEINBOLDT [82] and SPENCE/WERNER [82].

(iii) What happens if equation (1) is obtained by discretization of an operator equation defined in a function space, see KIKUCHI [79], SCHOLZ [80], SPENCE/MOORE [80], BREZZI/RAPPAZ/RAVIART [81], MOORE/SPENCE [81], PAUMIER [81], FINK/RHEINBOLDT [82].

12. References

ABBOTT, J.P.: Numerical continuation methods for nonlinear equations and bifurcation problems. Ph.D. Thesis. Austral. Nat. Univ., Canberra, 1977

ABBOTT, J.P.: An efficient algorithm for the determination of certain bifurcation points. J.Comput.Appl.Math.4 (1978), 19-27

- BATOZ, J.-L., DHATT, G.: Incremental displacement algorithms for nonlinear problems. *Internat.J.Numer.Methods Engrg.* 14 (1979), 1262-1267
- BOHL, E.: Bifurcation driven by diffusion, in these Proceedings
- BREZZI, F., RAPPAZ, J., RAVIART, P.A.: Finite dimensional approximation of nonlinear problems. Part II: Limit points. *Numer.Math.* 37 (1981), 1-28
- CHAN, T.: Techniques for large sparse systems arising from continuation methods, in these Proceedings
- CHUA, L.O., LIN, F.M.: Computer-aided analysis of electronic circuits: Algorithms and computational techniques. Prentice-Hall, Englewood Cliffs, 1976
- FINK, J.P., RHEINBOLDT, W.C.: On the descretization error of parametrized nonlinear equations. Technical Report ICMA-82-40. University of Pittsburgh, Pittsburgh, 1982
- GILL, P.E., MURRAY, W., WRIGHT, M.H.: Practical optimization. Academic Press, London, 1981
- GOLUB, G.H., PEREYRA, V.: The differentiation of pseudoinverses and nonlinear least squares problems whose variables separate. *SIAM J.Numer.Anal.* 10 (1973), 413-432
- GRIEWANK, A., REDDIEN, G.W.: Characterization and computation of generalized turning points. Manuscript 1983. Submitted to *SIAM J.Numer.Anal.*
- KELLER, H.B.: Numerical solution of bifurcation and nonlinear eigenvalue problems, in: RABINOWITZ, P.E. (Ed.): Application of bifurcation theory, p. 359-384. Academic Press, New York, 1977
- KELLER, H.B.: Global homotopies and Newton methods, in: DE BOOR, C., GOLUB, G.H. (Eds.): Recent advances in numerical analysis, p. 73-94. Academic Press, New York, 1978
- KIKUCHI, F.: Finite element approximations to bifurcation problems of turning point type. *Theoretical and Applied Mechanics* 27 (1979), 99-114
- KUBIČEK, M.: Evaluation of branching points for nonlinear boundary-value problems based on the GPM-technique. *Appl. Math.Comput.* 1 (1975), 341-352
- KUBIČEK, M.: Numerical determination of bifurcation points in steady state and periodic solutions - numerical algorithms and examples, in these Proceedings
- KUBIČEK, M., MAREK, I.: Evaluation of limit and bifurcation points for algebraic equations and nonlinear boundary value problems. *Appl.Math.Comput.* 5 (1979), 253-264
- MELHEM, R.G., RHEINBOLDT, W.C.: A comparison of methods for determining turning points of nonlinear equations. *Computing* 29 (1982), 201-226.

- MENZEL, R., SCHWETLICK, H.: Zur Lösung parameterabhängiger nichtlinearer Gleichungen mit singulären Jacobi-Matrizen. *Numer.Math.* 30 (1978), 65-79
- MENZEL, R., SCHWETLICK, H.: Parametrization via secant length and application to path following. Manuscript 1983. Submitted to *Numer.Math.*
- MOORE, G., SPENCE, A.: The calculation of turning points of nonlinear equations. *SIAM J.Numer.Anal.* 17 (1980), 567-576
- MOORE, G., SPENCE, A.: The convergence of operator approximations at turning points. *IMA J.Numer.Anal.* 1 (1981), 23-38
- PAUMIER, J.-C.: Une méthode numérique pour le calcul des points de retournement. Application à un problème aux limites non-linéaire. I. Etude théorique et expérimentation de la méthode. II. Analyse numérique d'un problème aux limites non-linéaire. *Numer.Math.* 37 (1981), 433-444; 445-452
- PÖNISCH, G.: Verfahren zur numerischen Bestimmung von Rückkehrpunkten implizit definierter Raumkurven. Dissertation A. Technische Universität Dresden, Dresden, 1979
- PÖNISCH, G., SCHWETLICK, H.: Computing turning points of curves implicitly defined by nonlinear equations depending on a parameter. *Computing* 26 (1981), 107-121
- PÖNISCH, G., SCHWETLICK, H.: Ein lokal überlinear konvergentes Verfahren zur Bestimmung von Rückkehrpunkten implizit definierter Raumkurven. *Numer.Math.* 38 (1982), 455-466; also Preprint 07-30-77 der Sektion Mathematik. Technische Universität Dresden, Dresden, 1977
- POSTON, T., STEWART, I.: Catastrophe theory and its applications. Pitman, London, 1978
- RAY, H.W.: Bifurcation phenomena in chemically reacting systems, in: RABINOWITZ, P.H. (Ed.): Applications of bifurcation theory, p. 285-315. Academic Press, New York, 1977
- RHEINBOLDT, W.C.: Numerical analysis of continuation methods for nonlinear structural problems. *Comput. & Structures* 13 (1981), 103-113
- RHEINBOLDT, W.C.: Computation of critical boundaries on equilibrium manifolds. *SIAM J.Numer.Anal.* 19 (1982), 653-669
- SCHOLZ, R.: Computation of turning points of the stationary Navier-Stokes equations using mixed finite elements, in: MITTELMANN, H.D., WEBER, H. (Eds.): Bifurcation problems and their numerical solution. ISNM 54, p. 147-162. Birkhäuser, Basel, 1980
- SCHWETLICK, H.: Numerische Lösung nichtlinearer Gleichungen. Dtsch. Verlag d. Wissenschaften, Berlin, 1979; Oldenbourg, München, 1979
- SCHWETLICK, H.: Effective methods for computing turning points of curves implicitly defined by nonlinear equations. Preprint

der Sektion Mathematik Nr.46. Martin-Luther-Universität
Halle-Wittenberg, Halle, 1981

SCHWETLICK, H.: Zur numerischen Behandlung nichtlinearer parameterabhängiger Gleichungen. Preprint der Sektion Mathematik Nr. 76-77. Martin-Luther-Universität Halle-Wittenberg, Halle, 1982

SCHWETLICK, H.: On the relation between some turning point algorithms and Wilson's method for constrained optimization. Manuscript 1983. Submitted to Numer.Math.

SEYDEL, R.: Numerical computation of branch points in nonlinear equations. Numer. math. 33 (1979), 339-352

SIMPSON, R.B.: A method for the numerical determination of bifurcation states of nonlinear systems of equations. SIAM J.Numer.Anal. 12 (1975), 439-451

SPENCE, A., MOORE, G.: A convergence analysis for turning points of nonlinear compact operator equations, in: ALBRECHT, J., COLLATZ, L. (Eds.): Numerical treatment of integral equations. ISNM 53, p. 203-212. Birkhäuser, Basel, 1980

SPENCE, A., WERNER, B.: Non-simple turning points and cusps. IMA J.Numer.Anal. 2 (1982), 413-427

TROGER, H.: Application of bifurcation theory to problems in mechanical engineering, in these Proceedings

Hubert Schwetlick
Martin-Luther-Universität Halle-Wittenberg
Sektion Mathematik
Universitätsplatz 6
4010 Halle
German Democratic Republic

A CONTINUATION ALGORITHM WITH STEP CONTROL

R. Seydel

The problem of steplength algorithms in continuation methods is considered. The present paper proposes a step control that leads to a routine which is extremely short and easy to handle. A complicated numerical example illustrates the effectiveness of the approach. A FORTRAN-routine is included.

1. Introduction

Consider a boundary-value problem

$$(1) \quad y'(t) = f(t, y, \lambda), \quad r(y(a), y(b)) = 0,$$

where ' denotes differentiation with respect to t , $a \leq t \leq b$, y , f and r are n -vectors, f and r sufficiently smooth. We are interested in the dependence of solutions $y(t; \lambda)$ of (1) on the real parameter λ . This problem of tracing branches of solutions is handled by continuation algorithms. Several continuation methods have been published; see, for example, [1, 4-12, 17, 18].

In 1977, a new continuation method was outlined [13]. Based on these ideas, an algorithm has been developed and tested on a wide range of problems (for examples see [15]). This continuation algorithm with step control is presented in this contribution. The algorithm is easily implemented, the corresponding routine in its basic version consisting of about 40 statements only. The algorithm works globally, turning points pose no difficulties. The method is not limited to ordinary differential equations, it handles systems of nonlinear

equations as well. The routine is to be used along with a solver for boundary-value problems or nonlinear equations respectively; such solvers are available on most computers.

2. Parameter Transformation

Often, for example at turning points, it is impossible to parametrize a branch of solutions by the "natural" parameter λ . Therefore, a parameter transformation $\lambda \leftrightarrow n$ is reasonable. By the existence and uniqueness theorem for ordinary differential equations with Lipschitz-continuous right-hand side, different solutions of (1) have different initial values. This situation suggests a parameter transformation by

$$y_k(a; \lambda) = n .$$

For fixed initial value n , the corresponding solution and value $\lambda = \lambda(n)$ can be calculated by solving the $(n+1)$ -boundary-value problem

$$(2) \quad \begin{pmatrix} y \\ \lambda \end{pmatrix}' = \begin{pmatrix} f(t, y, \lambda) \\ 0 \end{pmatrix}, \quad \begin{pmatrix} r(y(a), y(b)) \\ y_k(a) - n \end{pmatrix} = 0 .$$

This boundary-value problem (2) contains system (1). Thus, it is sufficient to solve (2) for suitable values of k and n in order to trace a branch. Thereby, the continuation is controlled by means of one boundary condition only. Fixing λ fits into this framework if we set

$$y_{n+1} := \lambda$$

Such sort of parameter transformations have shown to be very useful; see, for example, [11, 12, 14, 15]. Other extended systems may be used instead of (2); system (7) of [15] enables a continuation by asymmetry.

In the following two sections, criteria will be given for the choice of k and n .

3. Choice of Index k

For notational convenience, we use the abbreviation

$$z_i := y_i(a), \quad i=1, \dots, n+1$$

The initial values of the finally calculated solution of (2) will be labelled by z^n ("new") whereas those of the previous solution will be denoted by z^0 ("old").

The index k has to be chosen such that the branch can be parametrized by z_k in the next continuation step. This objective is very likely to be achieved if the relative changes in z_k are maximal, that is to say,

$$\left| \frac{z_k^n - z_k^0}{z_k^n} \right| := \max_{i=1, \dots, n+1} \left| \frac{z_i^n - z_i^0}{z_i^n} \right|$$

Sometimes, a preferential treatment of λ may be advisable for two reasons: First, the parameter λ may have a physical meaning. The second argument concerns the occurrence of bifurcations. Often, emanating branches are leaving the "main" branch perpendicular to the λ -axis. Thus, selecting $n=\lambda$, $k=n+1$, may be useful in order to remain on the main branch. Note that the same effect can be produced, if an L_2 -norm is prescribed by means of an additional variable; see, for example, [2].

These considerations lead to the following strategy for the choice of k : Let $w \in \mathbb{R}$ be a weight on selecting λ ($w=1$: no preference), then k is implicitly defined by

$$(3) \quad \begin{aligned} d_i &:= |z_i^n - z_i^0| / |z_i^n|, \quad i=1, \dots, n \\ d_{n+1} &:= w \cdot |\lambda^n - \lambda^0| / |\lambda^n| \\ d_k &:= \max \{ d_i \mid i=1, \dots, n+1 \}. \end{aligned}$$

4. Choice of Initial Value n , Estimation of Stepsizes

The value n of the fixed initial condition is the sum of z_k^n and a suitably chosen increment $s \in \mathbb{R}$,

$$n = z_k^n + s_k$$

s fixes the stepsize of the next continuation step. If $|s|$ is chosen too small, the continuation procedure involves too many steps and becomes inefficient. If $|s|$ is chosen too large, the iteration procedure for solving (2) ("SOLVER") may at best converge with difficulties. In the first case, $|s|$ should be enlarged; in the second case, $|s|$ has to be reduced.

A simple criterion for a reasonable improvement of the stepsize is based on the number (It) of iterations performed by SOLVER to obtain a solution of (2). This number It should be in accordance with a desired value I_0 . A desired medium number of iterations I_0 depends on the type of iteration method utilized and on the prescribed tolerance specification. For example, solving the boundary-value problem (2) by multiple shooting [16] with a relative error tolerance of 10^{-4} , the choice of $I_0=6$ is advisable. This number $I_0=6$ takes into account the adaption of the Newton-relaxation factor and the use of rank-one approximations for Jacobian matrices [3, 16]. As a consequence, the next step of the continuation usually remains in the domain of attraction of Newton's method. The way of updating the last step given by

$$\bar{s} := z_k^n - z_k^0$$

is to modify it by the factor I_0/It ,

$$s = \bar{s} I_0 / It .$$

Thus, after having determined the index k by (3), the parameter n is given by

$$(4) \quad n = z_k^n + (z_k^n - z_k^0) \cdot I_0 / It$$

The sensitivity of the present approach is influenced by the relation between the order of magnitude of I_0 and the stepsizes. Satisfactory results have been obtained using solvers that are of quasi-Newton type.

5. The Algorithm

Often, the costs of a continuation procedure are significantly reduced, if the initial guess to the next solution is calculated by "extrapolation", based on y^n and y^0 . This option can be realized in the following algorithm. For sake of clearness, no end condition is specified.

input parameters: k, s : initial step
 I_0 : desired number of iterations for SOLVER
 w : weight factor influencing the choice $\eta = \lambda$

solution arrays: Y : current iterate / initial guess
 YN : last solution of (2)
 $Y0$: older solution of (2)
 $Z, ZN, Z0$: initial values

algorithm:

```

go to 2
1   k=0 , d0=0
loop (i=1,...,n+1):
    if |ZN(i)|<10-3 then d=0
        else d=|(ZN(i)-Z0(i))/ZN(i)|
    if i=n+1 then d=d*w
    if d>d0 then k=i , d0=d
    s=ZN(k)-Z0(k)
    α=I0/It
    s=s*α

2   ifail=0
21  n=ZN(k)+s
    if extrapolation then Y=(1+α)*YN-α*Y0
        else Y=YN
    call SOLVER
    It= number of iterations
    if failure go to 3
    Y0=YN

```

```

YN=Y
go to 1

3   if ifail>2  stop
      ifail=ifail+1
      s=s/5
      a=a/5
      go to 21

```

6. Systems of Nonlinear Equations

Consider the system of nonlinear equations

$$(5) \quad g(x, \lambda) = 0 ,$$

$\lambda \in \mathbb{R}$, $x \in \mathbb{R}^n$, $g(x, \lambda) \in \mathbb{R}^n$. Here, initial values and solution arrays coincide, $x=z=y$. Similar to (2), instead of (5) the $(n+1)$ -system

$$(6) \quad \begin{cases} g(x, \lambda) \\ x_k - n \end{cases} = 0$$

is solved. The same algorithm as in the case of ordinary differential equations is valid.

7. Example

The continuation algorithm established above has been tested on several examples. We demonstrate the success of the algorithm by means of a Duffing equation that exhibits an extremely complicated dependence of the solutions on the parameter λ (exciting frequency). This Duffing equation

$$\ddot{x} + \frac{1}{25} \dot{x} - \frac{1}{5} x + \frac{8}{15} x^3 = \frac{2}{5} \cos \lambda t$$

and the computation of its harmonic solutions was suggested as a standard example for testing continuation methods; details are given in [2], compare also [15]. All the corresponding boundary-value problems were solved by the multiple shooting code BOUNDS from [3].

In the following three figures, the values $y_1(0)=x(0)$

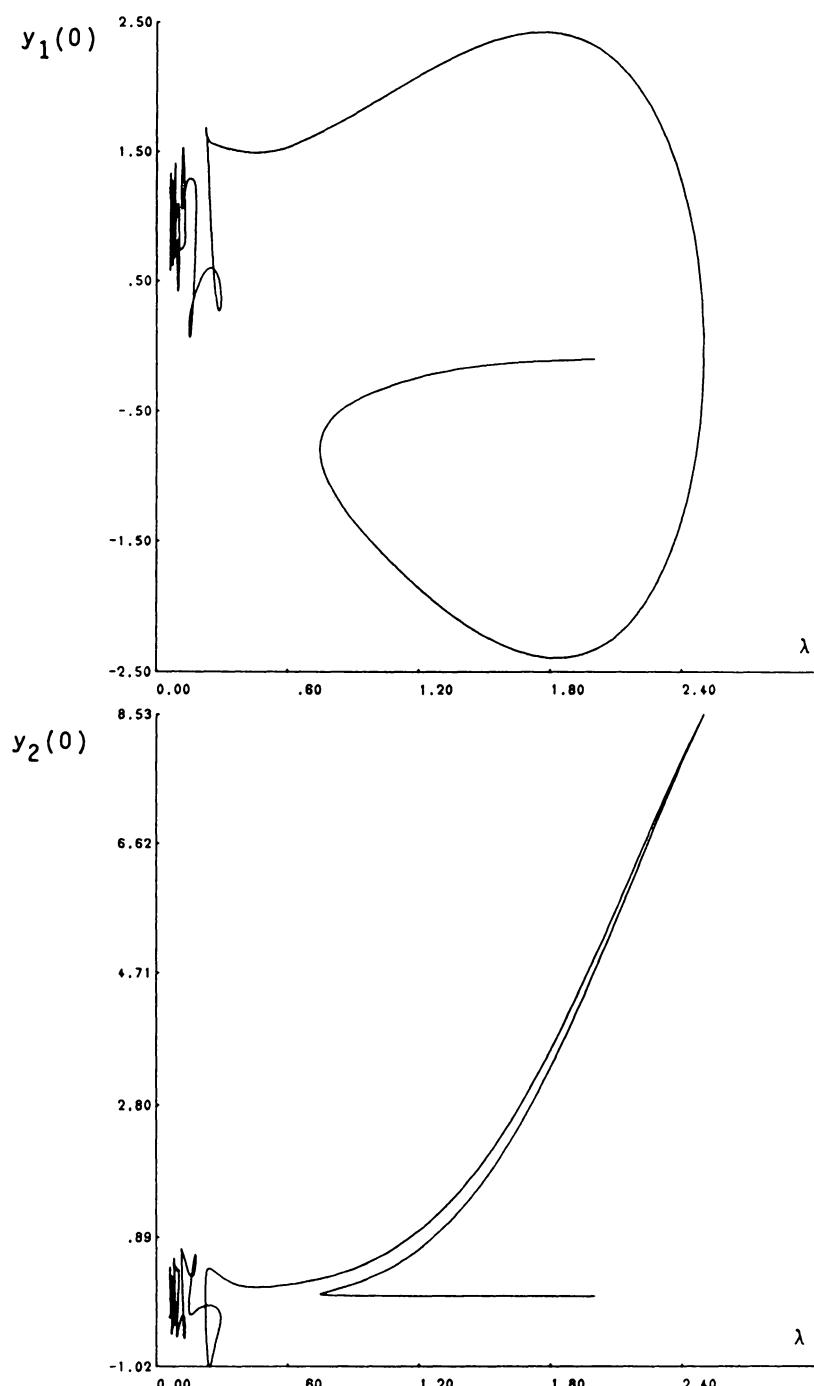


Fig.1 branching diagram (symmetric solutions)

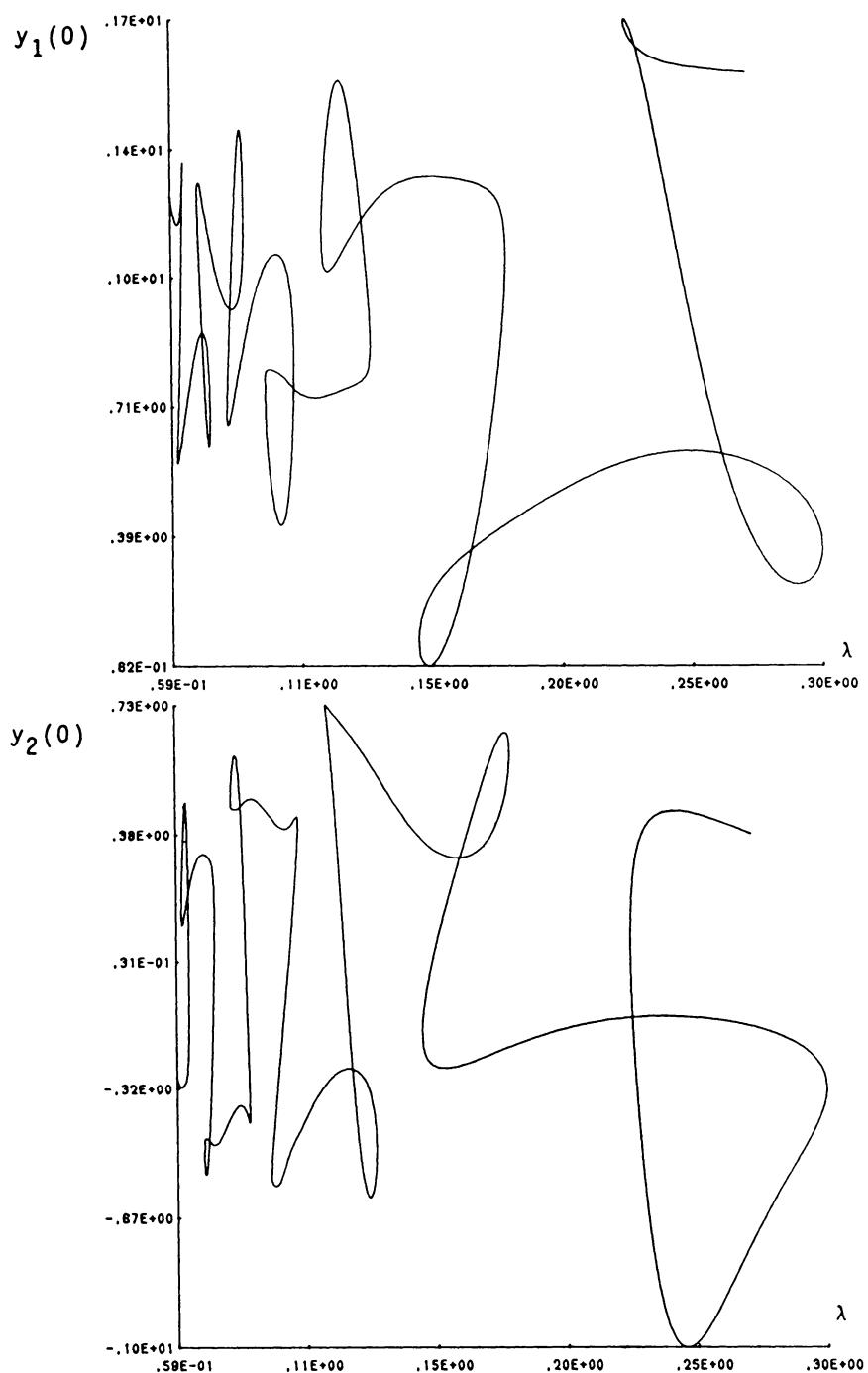


Fig.2 detail of Fig.1

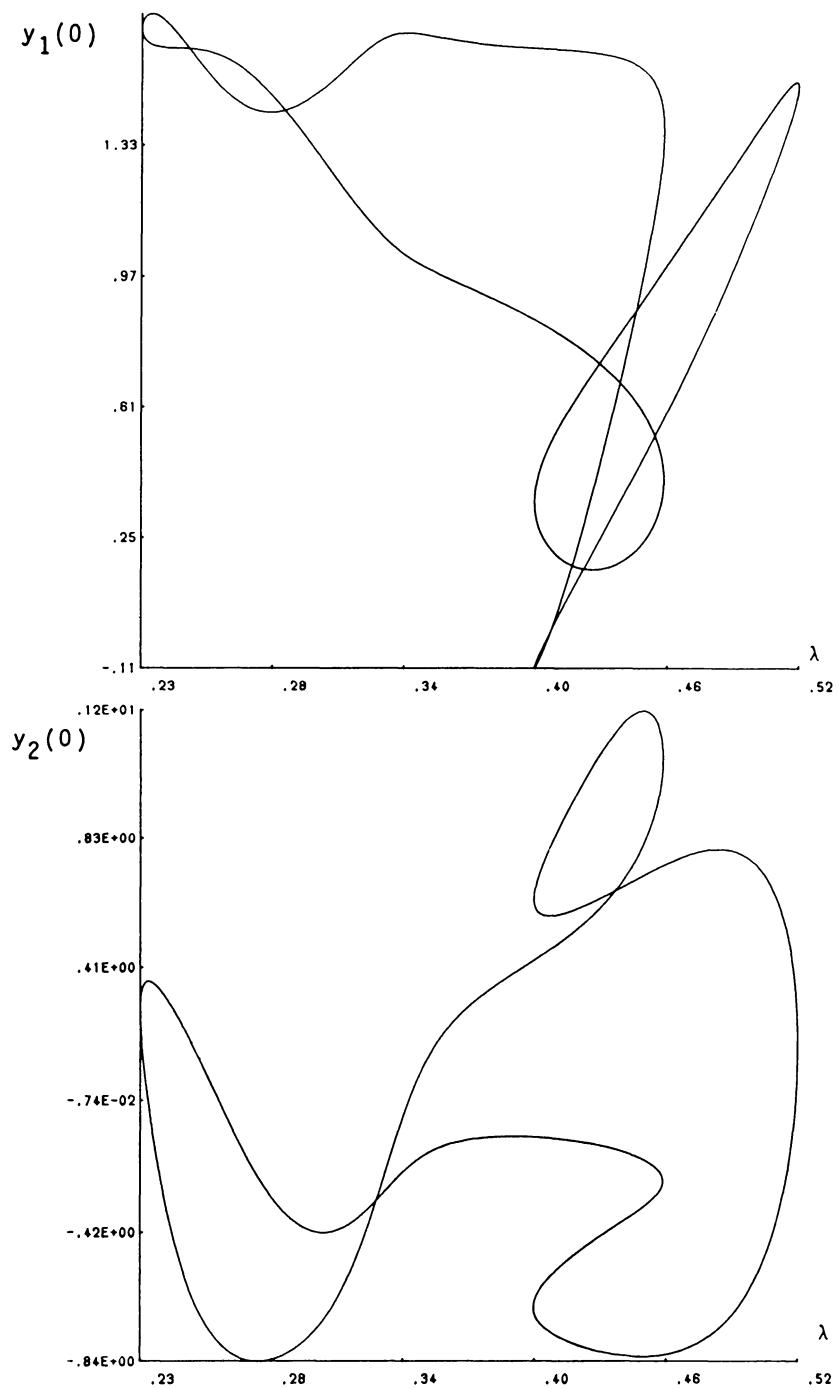


Fig.3 emanating branch

and $y_2(0)=x(0)$ of the harmonic solutions versus the exciting frequency λ are plotted. Fig.1 contains the results of one run of the algorithm, starting at $\lambda=3$ ($\|x\|_2=0.044$) with prescribed end at $\lambda=0.06$ ($\|x\|_2=8.94$). For input parameters the values $I_0=6$, $w=10$, $k=3$, $s=-1$ were chosen. As can be seen by the results, the algorithm easily traces the branch in spite of its complicated structure (313 solutions were calculated). The loops for $\lambda < 0.3$ are enlarged in Fig.2.

Figures 1 and 2 show the "main" branch, its solutions in the phase plane being symmetric with respect to the origin. Many bifurcation points were passed during the course of the continuation. One of the emanating branches (asymmetric solutions) is shown in Fig.3. The corresponding results of this closed branch were again obtained by means of only one run of the algorithm of Section 5.

Appendix

In this appendix, the FORTRAN 77 subroutine HOM is listed. This routine is a version of the algorithm of Section 5 that performs a restart after failure. By equation (3), two indices k , $k2$ are determined. In case of failure of continuation by z_k , continuation switches over to z_{k2} .

The arrays in HOM are adapted to solving boundary-value problems by multiple shooting (BOUNDS from [3]). The arrays Y , YA , YN have dimensions of at least $Y(N+1,M)$. In case of nonlinear equations, the arrays should be modified to vectors ($M=1$), T may be skipped then.

calling sequence:

N	number of variables n (without parameter λ)
M	number of nodes (for multiple shooting)
T	vector of nodes
YN	array of current solution for starting continuation
	$YN(i,j)$, $i=1,\dots,N+1$: solution vector at the node $T(j)$ ($j=1,\dots,M$)

EPS relative error tolerance for SOLVER
 INDEX in the starting step the corresponding component
 k=INDEX is fixed
 S initial stepsize to start continuation
 MAX continuation terminates after MAX steps
 ITAIM desired number of iteration steps I_o for SOLVER
 GEW weighting factor for preferred use of parameter λ
 IEX =1 : extrapolation, using secant
 =0 : no extrapolation, HOM uses YN for initial guess

definition of the boundary-value problem:

- a) The subroutine of the right-hand side has to define $y_{n+1}'=0$.
- b) The subroutine of the boundary conditions has to define $r_{n+1}=y_k(a)-n$, the index k and the value n are contained in

$$\text{COMMON ETA,K}$$
- c) Solution of (2) (or (6)) by SOLVER:
 IT is the number of iteration steps performed to approximate a solution Y with relative error tolerance EPS. SOLVER has to set IT=-1 if convergence fails.

listing of HOM:

```

SUBROUTINE HOM (N,M,T,YN,EPS,INDEX,S,MAX,ITAIM,GEW,IEX)
C.. VERSION WITH RESTART AFTER FAILURE

COMMON ETA,K
DIMENSION Y(12,7),YA(12,7),YN(12,7),T(7)
K=IMDEX
IHOH=1
N1=N+1
GO TO 3
*****
```

```

2      IUDM=IMOM+1
C..  DETERMINATION OF INDEX K:
      K=0
      D1=0.
      K2=0
      D2=0.
      DO 25 I=1,N1
      IF (ABS(YN(I,1)).LT.1.E-3) D=0.
      IF (ABS(YN(I,1)).GE.1.E-3)
      *    D=ABS((YN(I,1)-YA(I,1))/YN(I,1))
      IF (I.EQ.N+1) D=D*GEW
      IF (D.GT.D1) THEN
          D2=D1
          K2=K
          D1=D
          K=I
      ELSE
          IF (D.GE.D2) THEN
              D2=D
              K2=I
          END IF
      END IF
25    CONTINUE
      IF (D1.LT.EPS) GO TO 9
C..  NEW STEPSIZE:
28    S=YH(K,1)-YA(K,1)
      ALPHA=FLOAT(ITAIM)/FLOAT(IT)
      S=S*ALPHA

3      IFAIL=0
C..  IINITIAL GUESS:
31    DO 35 I=1,N1
      DO 35 J=1,M
      IF (IHOM.GT.1 .AND. IEX.EQ.1)
      1      Y(I,J)=YN(I,J)*(1.+ALPHA)-ALPHA*YA(I,J)
      IF (IHOM.EQ.1 .OR. IEX.EQ.0) Y(I,J)=YN(I,J)
      35  COMTINUE
      ETA=YH(K,1)+S
      Y(K,1)=ETA
      CALL SOLVER (N1,M,T,Y,EPS,IT)
      IF (IT.GT.0) THEN
          DO 5 I=1,N1
          DO 5 J=1,M
              YA(I,J)=YN(I,J)
              YM(I,J)=Y(I,J)
5      WRITE (6,600) IHOM,IT,K,S,(Y(I,J),I=1,N1)
      WRITE (7,601) IHOM,N1,(Y(K,J),J=1,M)
      IF (IHOM.LT.MAX) GO TO 2

C***** ****

```

```

C.. FAIL-- AND END-CONDITIONS:
      INDEX=K
      GO TO 99
    ELSE
      WRITE (6,600) IHOM,IT,K,S
      IF (IFAIL.LT.2) THEN
        C..
          REDUCTION AFTER FAILURE:
          IFAIL=IFAIL+1
          S=S*0.2
          IF (IHOM.GT.1) ALPHA=ALPHA*0.2
          GO TO 31
        ELSE
          IF (K.EQ.K2.OR.K2.EQ.0) THEN
            WRITE (6,*) 'NO CONVERGENCE'
            GO TO 9
            END IF
        C..
          TRY NEXT INDEX:
          K=K2
          GO TO 28
        END IF
      END IF

9      INDEX==1
99     RETURN
600     FORMAT ("0",I3,"/IT",I2,"/K=",I2,"/S=",E10.4,"   ",
+ 3E17.9,2(/30X,3E17.9))
601     FORMAT (I2,I3,5E15.8)
END

```

References

- 1 E. Allgower, K. Georg: Simplicial and continuation methods for approximating fixed points and solutions to systems of equations. SIAM Review 22 (1980), pp. 28-85
- 2 K.-H. Becker, R. Seydel: A Duffing equation with more than 20 branch points. in: Numerical solution of nonlinear equations, E. Allgower et al., ed., Lecture Notes in Math. 878, Springer, Berlin-Heidelberg- New York, 1981, pp. 98-107
- 3 R. Bulirsch, J. Stoer, P. Deuflhard: Numerical solution of nonlinear two-point boundary value problems I. to appear in Numer. Math., Handbook Series Approximation
- 4 C. Den Heijer, W.C. Rheinboldt: On steplength algorithms for a class of continuation methods. SIAM J. Numer. Anal. 18 (1981), pp. 925-948

- 5 P. Deuflhard: A stepsize control for continuation methods and its special application to multiple shooting techniques. *Numer. Math.* 33 (1979), pp. 115-146
- 6 H.B. Keller: Numerical solution of bifurcation and nonlinear eigenvalue problems. in: *Applications of bifurcation theory*, P.H. Rabinowitz, ed., Academic Press, New York, 1977, pp. 359-384
- 7 E. Kosin: Ein Homotopieverfahren zur numerischen Behandlung von Lösungszweigen nichtlinearer Gleichungssysteme mit spezieller Anwendung auf die Mehrzielmethode. Dissertation, Technische Universität München, 1983
- 8 M. Kubíček: Algorithm 502. Dependence of solutions of nonlinear systems on a parameter. *ACM Trans. Math. Software* 2 (1976), pp. 98-107
- 9 P. Lory: Enlarging the domain of convergence for multiple shooting by the homotopy method. *Numer. Math.* 35 (1980), pp. 231-240
- 10 R. Menzel, H. Schwetlick: Zur Lösung parameterabhängiger nichtlinearer Gleichungen mit singulären Jacobi-Matrizen. *Numer. Math.* 30 (1978), pp. 65-79
- 11 W.C. Rheinboldt: Solution fields of nonlinear equations and continuation methods. *SIAM J. Numer. Anal.* 17 (1980), pp. 221-237
- 12 W.C. Rheinboldt, J.v.Burkardt: A locally parametrized continuation process. *ACM Tans. Math. Software* 9 (1983), pp. 215-235
- 13 R. Seydel: Numerische Berechnung von Verzweigungen bei gewöhnlichen Differentialgleichungen. Dissertation, Technische Universität München, 1977
- 14 R. Seydel: Numerical computation of primary bifurcation points in ordinary differential equations. *ISNM* 48 (1979), pp. 161-169
- 15 R. Seydel: Branch switching in bifurcation problems for ordinary differential equations. *Numer. Math.* 41 (1983), pp. 93-116
- 16 J. Stoer, R. Bulirsch: *Introduction to numerical analysis*. Springer, Berlin-Heidelberg- New York, 1980
- 17 H. Wacker (ed.): *Continuation methods*. Academic Press, New York, 1978

- 18 L.T. Watson: An algorithm that is globally convergent with probability one for a class of nonlinear two-point boundary value problems. SIAM J. Numer. Anal. 16 (1979), pp. 394-401

R. Seydel
Department of Chemical Engineering
State University of New York at Buffalo
507 Clifford C. Furnas Hall
Amherst
New York 14260

Bifurcation near multiple eigenvalues for the flow
between concentric counterrotating cylinders *

R.C. DiPrima, Rensselaer Polytechnic Institute, Troy
P.M. Eagles, City University, London
J. Sijbrand, University of Utrecht **

We study the flow between counterrotating cylinders of infinite length, particularly the bifurcation of a variety of flow regimes from Couette flow. Let a , b and Ω_1 , Ω_2 denote the radii and angular velocities of the inner and outer cylinders, respectively and let $n = a/b$. We also introduce the Reynolds numbers:

$$(1) \quad R_i = \Omega_1 \frac{a(b-a)}{v} \text{ and } R_o = \Omega_2 \frac{b(b-a)}{v}.$$

The Navier Stokes equations linearized at the Couette flow permit a solution depending on the axial (z) and radial (r) coordinates but not on the azimuthal (θ) coordinate:

$$(2) \quad e^{\sigma t} + i\lambda z \vec{f}(r).$$

In (2) the eigenvalue σ and the eigenvector \vec{f} depend on the Reynolds numbers and on the wavelength λ . We shall not address the problem of wavelength selection here so let us assume that λ is fixed. In the R_o , R_i parameter plane (see figure 1) there is a curve $\sigma(R_o, R_i) = 0$ which separates stable from unstable perturbations of the form (2). The real eigenvalue σ is degenerate with eigenvectors $e^{i\lambda z} f(r)$ and $e^{-i\lambda z} f^*(r)$.

* This research was partially supported by the Army Research Office and the Fluid Mechanics Branch of the Office of Naval Research, and by the Netherlands Organization for the Advancement of Pure Research (Z.W.O.)

** Present address: Shell Laboratories, Amsterdam

Likewise, the linearized equations have solutions of the form

$$(3) \quad e^{\nu t} + i\lambda z + i\theta g(r).$$

However, in this case the eigenvalue ν is complex. When (R_o, R_i) is below the curve $\nu_r(R_o, R_i) = 0$, Couette flow is stable with respect to perturbations of the type 3; here ν_r denotes the real part of ν . To every eigenvalue ν there are two eigenvectors $e^{i\lambda z + i\theta} g_1(r)$ and $e^{i\lambda z - i\theta} g_2(r)$, and the complex conjugate eigenvalue $\tilde{\nu}$ introduces the complex conjugates of these vectors.

As remarked by Krueger, Gross and DiPrima (1966), for η near 1 there is a combination of parameter values (\hat{R}_o, \hat{R}_i) for which the perturbations (2) and (3) are simultaneously critical, leading to a total of 6 critical eigenvectors at (\hat{R}_o, \hat{R}_i) . The remainder of this contribution deals with the flow patterns occurring for R_o, R_i close to \hat{R}_o, \hat{R}_i .

The analysis starts by representing the components of a disturbance by

$$(4) \quad \begin{aligned} & F_c f_c(r) \cos \lambda z + F_s f_s(r) \sin \lambda z + \\ & H_c g_c(r) \cos \lambda z e^{i\theta} + H_s g_s(r) \sin \lambda z e^{i\theta} + \\ & \tilde{H}_c \tilde{g}_c(r) \cos \lambda z e^{-i\theta} + \tilde{H}_s \tilde{g}_s(r) \sin \lambda z e^{-i\theta} + \psi \end{aligned}$$

where $F_c, F_s, H_c, H_s, \tilde{H}_c, \tilde{H}_s$ are scalar functions of time and ψ is complementary to these 6 modes. By center-manifold techniques an invariant manifold is then constructed of the form

$$(5) \quad \psi = \psi(F_c, F_s, H_c, H_s, \tilde{H}_c, \tilde{H}_s).$$

The function ψ may be expanded in a (non-convergent) asymptotic expansion without constant or linear terms (because of the tangency of the invariant manifold to the critical eigenspace). Thus the asymptotic expansion of ψ starts with quadratic terms and the coefficients in the expansion are functions of r, z and θ defined by inhomogeneous boundary value problems, which can be solved by numerical methods.

On the center manifold the flow is described by a system of 6 ordinary differential equations:

$$\begin{aligned}
 \frac{dF_C}{dt} &= \sigma F_C + a_1 F_C^3 + a_1 F_C F_S^2 + a_3 F_C |H_C|^2 + a_4 F_C |H_S|^2 + \\
 &\quad + \delta_5 F_S H_C \tilde{H}_S + \tilde{\delta}_5 F_S \tilde{H}_C H_S + 5^{\text{th}} \text{ order terms} \\
 \frac{dF_S}{dt} &= \sigma F_S + \text{similar cubic terms} + 5^{\text{th}} \text{ order terms} \\
 \frac{dH_C}{dt} &= v H_C + \text{similar cubic terms} + 5^{\text{th}} \text{ order terms} \\
 (6) \quad \frac{d\tilde{H}_S}{dt} &= v H_S + \text{similar cubic terms} + 5^{\text{th}} \text{ order terms} \\
 \frac{d\tilde{H}_C}{dt} &= \tilde{v} \tilde{H}_C + \text{similar cubic terms} + 5^{\text{th}} \text{ order terms} \\
 \frac{dH_S}{dt} &= \tilde{v} \tilde{H}_S + \text{similar cubic terms} + 5^{\text{th}} \text{ order terms}
 \end{aligned}$$

These O.D.E.'s are the same as those derived by Davey, DiPrima and Stuart (1968) for the case $R_o=0$. The equations for \tilde{H}_C , \tilde{H}_S are the complex conjugates of the equations for H_C , H_S . For the coefficients a_i , δ_i etc. appearing in these equations closed-form formulas have been derived, in the form of integrals over r of the adjoint eigenfunctions of the Navier Stokes equations multiplying complicated expressions involving the coefficient functions of ψ discussed above. These integrals have been evaluated numerically for selected values of η and λ at the corresponding values of (\hat{R}_1, \hat{R}_o) leading to concrete values for the a_i , δ_i .

Knowing the coefficients we now turn to the analysis of (6). The structure of the system (6) permits us to put $F_S = H_C = H_S = 0$ corresponding to an initial state $F_C \neq 0$, $F_S = H_C = H_S = 0$. Then the solution will have a vanishing contribution by F_S , H_C , H_S modes for all t . In this case the system (6) reduces to

$$(7) \quad \frac{dF_C}{dt} = \sigma F_C + a_1 F_C^3 + 5^{\text{th}} \text{ order terms.}$$

The solution $F_C = (-\sigma/a_1)^{\frac{1}{2}}$ represents the 'Taylor vortex', defined for $\sigma < 0$, (when $a_1 > 0$).

A different reduction to a simple solution is obtained by setting $F_C = F_S = H_S = 0$. The resulting solution is the so-called 'non-axisymmetric simple mode'.

A more complicated solution is obtained by putting $F_S = H_C = 0$. The resulting set of 3 O.D.E.'s is:

$$(8) \quad \begin{aligned} \frac{dF_C}{dt} &= \sigma F_C + a_1 F_C^3 + a_4 F_C |H_S|^2 \\ \frac{dH_S}{dt} &= v H_S + b_1 H_S |H_S|^2 + b_4 F_C^2 H_S \\ \frac{d\bar{H}_S}{dt} &= \text{complex conjugate.} \end{aligned}$$

System (8) can be further reduced to two degrees of freedom by a transformation $F_C = x$, $H_S = ye^{i(vt+\phi)}$ which leads to an uncoupling of the phase ϕ from the amplitudes x and y :

$$(9) \quad \begin{aligned} \frac{dx}{dt} &= \sigma x + a_1 x^3 + a_4 xy^2 \\ \frac{dy}{dt} &= v_r y + b_{1r} y^3 + b_{4r} x^2 y \end{aligned}$$

This equation has a solution with $y=0$ representing the Taylor vortex, and a solution with $x=0$ representing the non-axisymmetric simple mode, and more important, a solution with nonvanishing contributions from both x and y which represents the wavy vortex explicitly given by:

$$(10) \quad \begin{pmatrix} a_1 & a_4 \\ b_{4r} & b_{1r} \end{pmatrix} \begin{pmatrix} x^2 \\ y^2 \end{pmatrix} = \begin{pmatrix} -\sigma \\ -v_r \end{pmatrix}$$

From (10) one sees that the wavy vortex exists if and only if

$$(11) \quad \frac{a_4 v_r - b_{1r} \sigma}{a_1 b_{1r} - a_4 b_{4r}} > 0 \text{ and } \frac{b_{4r} \sigma - a_1 v_r}{a_1 b_{1r} - a_4 b_{4r}} > 0.$$

The conditions for existence of the various solutions of (6) all take the form of inequalities in terms of σ and v_r . It is convenient to consider σ and v_r as the problem parameters rather than R_o, R_i , taking into account the 1-1 correspondence of a neighborhood of $(0,0)$ in the (σ, v_r) plane and a neighborhood of (\hat{R}_o, \hat{R}_i) in the (R_o, R_i) plane (see figure 2).

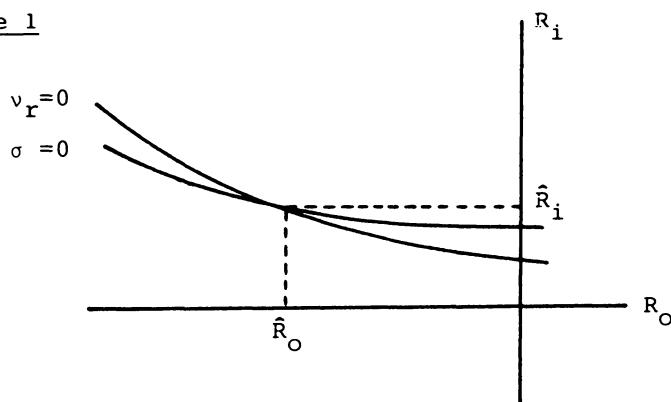
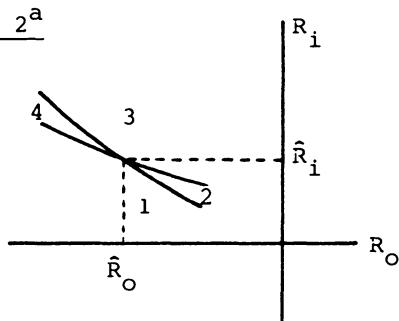
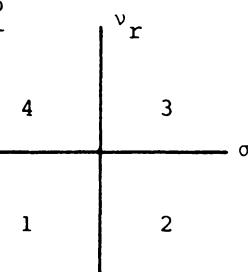
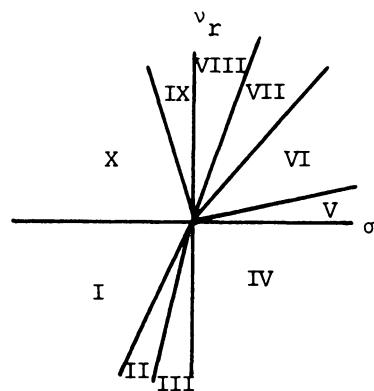
In the (σ, v_r) plane the set of points for which the wavy vortex exists now takes the shape of a sector bounded by the rays $v_r/\sigma = b_{1r}/a_4$ and $v_r/\sigma = b_{4r}/a_1$. At $v_r/\sigma = b_{1r}/a_4$ the wavy vortex branches out of the Taylor vortex, at $v_r/\sigma = b_{4r}/a_1$ it branches into the non-axisymmetric simple mode. A classification of the possible relative positions of these bifurcation lines (due to different parameters a_1, a_4, b_{1r}, b_{4r}) and the consequent differences in the phase planes of (9) for (σ, v_r) in the various sectors has been carried out by Keener (1976), Holmes (1980) and Langford and Iooss (1980). There are two other ways to reduce (6) to a consistent subsystem with 2 degrees of freedom, leading to a 'non-axisymmetric vortex' and a spiral vortex. All of these solutions exist in their respective sectors of the (σ, v_r) plane, leading to a total of some 10 sectors as depicted in figure 3. When crossing from one sector into the next there will be one type of solution branching out of or into another type of solution. The relative position of the sectors depends on the values of η and λ .

The analysis described up to now is not entirely satisfactory for the following reasons: 1) the sectors mentioned above in which some types of solution are predicted to exist, do not match experimental evidence in all cases. For instance, the calculated coefficient a_1 may be positive, indicating subcritical bifurcation of the Taylor vortex, while there is experimental evidence of supercritical bifurcation, certainly in the case of small enough R_o . This kind of problem is resolved by studying the full fifth order system rather than the third order system (6). 2) All bifurcations which occur upon crossing from one sector into the next are accompanied by a change of

stability of the solutions involved. However, we have observed two instances of a solution changing its stability without one of the types of solution mentioned above (i.e. Taylor vortex, wavy vortex, non-axisymmetric simple mode, non-axisymmetric vortex, spiral vortex) emerging from this point of critical stability. In both instances we have been able to prove that bifurcation does occur from these critical point but that the bifurcating solution has to be described by more than two degrees of freedom, thereby leading to a solution which is essentially more complicated than those mentioned up to now. Details on the analysis of the cubic as well as the fifth order systems, and a description of the bifurcation mentioned in the previous paragraph will appear in forthcoming papers.

References

- A. Davey, R.C. DiPrima and J.T. Stuart 1968, On the instability of Taylor vortices. *J. Fluid Mech.* 31, 17-52.
- P. Holmes 1980, Unfolding a degenerate nonlinear oscillator: a codimension two bifurcation. *Annals N.Y. Acad. Sci.* 357, 473-488.
- J.P. Keener 1976, Secondary bifurcations in nonlinear diffusion reaction equations. *Studies in Appl. Math.* 55, 187-211.
- R.R. Krueger, A. Gross and R.C. DiPrima 1966, On the relative importance of Taylor-vortex and non-axisymmetric modes in flow between rotating cylinders. *J. Fluid Mech.* 24, 521-538.
- W.F. Langford and G. Iooss 1980, Interaction of Hopf and pitch-forkbifurcations, in: H.D. Mittelmann, H. Weber (eds.), *Bifurcation problems and their numerical solution*, pp. 103 - 134, ISNM 54, Birkhäuser-Verlag, Bern 1980

figure 1figure 2^afigure 2^bfigure 3

THE NUMERICAL CALCULATION OF CUSPS, BIFURCATION POINTS AND ISOLA
FORMATION POINTS IN TWO PARAMETER PROBLEMS

A. Spence and A.D. Jepson

1. Introduction

In this paper we discuss the numerical computation of solutions of the *nonlinear, two parameter, problem*

$$(1.1) \quad f(x, \lambda, \alpha) = 0, \quad f : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^n,$$

where $x \in \mathbb{R}^n$ is a state variable, λ and α are parameters, and f is a smooth function. Many physical systems can be described by equations like (1.1), see for example, [1], [3], [12] and [7], [17], where there are more than two parameters.

Interest centres on the *singular points* of (1.1), i.e. the points where $f_x(x, \lambda, \alpha)$ is singular, since at such points there is often a change in the stability of the time dependent problem $x_t = f(x, \lambda, \alpha)$. The main aims of this paper are to describe what types of singular point occur generically in (1.1) when $\text{Rank}(f_x) = n - 1$, and to outline a general approach for finding these singular points. We also discuss briefly the case $\text{Rank}(f_x) = n - 2$.

The basic idea of our approach is, first, to find a simple quadratic turning point of (1.1) (see Section 2), and then to follow a path of such turning points, a fold curve (see Section 3). We show that cusp points, bifurcation points and points of isola formation are themselves turning points in the fold curve and this has important consequences for the numerical calculation of these singularities. A key feature of our theoretical approach is that it provides quantities which characterise the various types of singularity in (1.1) and which can be readily recovered in our numerical approach. Thus a particularly efficient monitoring process can be carried out as the fold curve is computed, and hence the geometry of the solution (equilibrium) surface of (1.1) near the fold curve can be easily ascertained.

Much of the material in this paper is a combination of the ideas in [8] and [15]. Also relevant for problems like (1.1), especially for the case $\alpha \in \mathbb{R}^P$, $P > 1$, is the work in [9], where connections with singularity theory are explored.

The plan of the paper is as follows. In Section 2 some theoretical results for one parameter problems are reproduced for convenient reference. The key theoretical results are given in Section 3. Also in Section 3 the case $\text{Rank}(f_x) = n - 2$ is discussed since this also leads to simple turning points in the fold curve. In Section 4 we give some numerical details.

2. Theoretical results for one parameter problems

In this section we discuss briefly the theory of one parameter problems and introduce our terminology. Consider the problem

$$(2.1) \quad g(x, \lambda) = 0, \quad g: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n,$$

where g is a smooth mapping.

Assume $g(x_0, \lambda_0) = 0$. If $g_x^0 := g_x(x_0, \lambda_0)$ is nonsingular then (x_0, λ_0) is a regular point and the Implicit Function Theorem gives results about the existence and uniqueness of a solution of (2.1) near (x_0, λ_0) . If g_x^0 is singular then various types of behaviour can occur. In this case let us assume

$$(2.2) \quad \begin{aligned} a) \quad \text{Null}(g_x^0) &= \text{span}\{\phi_1\}, \quad \phi_1 \in \mathbb{R}^n \setminus \{0\}, \\ b) \quad \text{Range}(g_x^0) &= \{y \in \mathbb{R}^n : \psi_1^T y = 0\}, \quad \psi_1 \in \mathbb{R}^n \setminus \{0\}. \end{aligned}$$

We call (x_0, λ_0) a simple singular point if (2.2) holds. Also the singular mapping g_x^0 induces a natural decomposition of \mathbb{R}^n described by

$$(2.3) \quad \mathbb{R}^n = \text{Null}(g_x^0) \oplus V.$$

Now, for convenience, we gather together some definitions. Define

$$(2.4) \quad a_g := \psi_1^T g_{xx}^0 \phi_1 \phi_1;$$

$$(2.5) \quad \begin{cases} v_0 = 0 & \text{if } a_g \neq 0, \\ g_x^0 v_0 = -g_{xx}^0 \phi_1 \phi_1, & v_0 \in V, \quad \text{if } a_g = 0; \end{cases}$$

$$(2.6) \quad d_g := \psi_1^T g_{xxx}^0 \phi_1 \phi_1 \phi_1 + 3\psi_1^T g_{xx}^0 \phi_1 v_0;$$

$$(2.7) \quad \begin{cases} z_o = 0, & \text{if } \psi_1^T g_\lambda^o \neq 0, \\ g_x^o z_o = -g_\lambda^o, & z_o \in V, \text{ if } \psi_1^T g_\lambda^o = 0; \end{cases}$$

$$(2.8) \quad b_{g\lambda} := \psi_1^T g_{x\lambda}^o \phi_1 + \psi_1^T g_{xx}^o \phi_1 z_o;$$

$$(2.9) \quad c_{g\lambda} := \psi_1^T g_{\lambda\lambda}^o + 2\psi_1^T g_{x\lambda}^o z_o + \psi_1^T g_{xx}^o z_o^2;$$

and, finally

$$(2.10) \quad D_1 := b_{g\lambda}^2 - a_g c_{g\lambda}.$$

(Here we have made use of the notation $g_{xx}^o := g_{xx}(x_o, \lambda_o)$, etc.). Now, if $\psi_1^T g_\lambda^o \neq 0$, $a_g \neq 0$, then (x_o, λ_o) is a *simple quadratic turning point*; if $\psi_1^T g_\lambda^o \neq 0$, $a_g = 0$, $d_g \neq 0$, then (x_o, λ_o) is a *simple cubic turning point (cusp point)* if $\psi_1^T g_\lambda^o = 0$, $D_1 > 0$, $a_g \neq 0$, then (x_o, λ_o) is a *simple transcritical bifurcation point*; and if $\psi_1^T g_\lambda^o = 0$, $D_1 < 0$, then (x_o, λ_o) is a *simple point of isola formation*, (see Figure 3.2).

We shall see that all four types of singular point occur generically in problems like (1.1) and, the last three types (as well as being important in physical situations) are of interest in the discussion of qualitatively similar bifurcation diagrams (see [9], [10]). Simple quadratic turning points occur generically in one parameter systems like (2.1) and so we discuss their calculation in this section.

Many methods exist for the calculation of simple quadratic turning points and we refer the reader to [13] and the references therein. We shall discuss only one method here. We introduce the extended system

$$(2.11) \quad G(y) := \begin{bmatrix} g(x, \lambda) \\ \psi^T g_x(x, \lambda) \phi \end{bmatrix}, \quad y = (x, \lambda),$$

where $\psi = \psi(x, \lambda)$, $\phi = \phi(x, \lambda)$ are the *singular vectors* of $g_x(x, \lambda)$, associated with a *singular value* σ i.e.

$$(2.12) \quad g_x(x, \lambda) \phi = \sigma \psi, \quad g_x^T(x, \lambda) \psi = \sigma \phi.$$

We note that the Jacobian of $G(y)$ has the form

$$(2.13) \quad G_y(y) := \begin{bmatrix} g_x & g_\lambda \\ \psi^T g_{xx} \phi + \psi^T g_x \phi + \psi^T g_x \phi_x & \psi^T g_{x\lambda} \phi + \psi^T g_\lambda \phi + \psi^T g_x \phi_\lambda \end{bmatrix}$$

and that the derivatives ϕ_x , ϕ_λ , etc., exist since singular vectors are eigenvectors of symmetric matrices.

At a singular point (x_o, λ_o) of (2.1) $\sigma = 0$, and the corresponding singular vectors are right and left eigenvectors. Thus, at $y_o = (x_o, \lambda_o)$, $G_y(y)$ has the form

$$(2.14) \quad G_y^o := G_y(y_o) = \begin{bmatrix} g_x^o & g_\lambda^o \\ \psi_1^T g_{xx}^o \phi_1 & \psi_1^T g_{x\lambda}^o \phi_1 \end{bmatrix}, \quad y_o = (x_o, \lambda_o).$$

It is easy to prove the following theorem.

Theorem 2.15 Assume (2.2) holds. Define the 2×2 matrix E by

$$(2.16) \quad E := \begin{bmatrix} 0 & \psi_1^T g_\lambda^o \\ a_g & b_{g\lambda} \end{bmatrix}$$

where a_g and $b_{g\lambda}$ are given by (2.4) and (2.8) respectively. Then

$$\dim \text{Null}(G_y^o) = \dim \text{Null}(E).$$

An immediate corollary is:

Corollary 2.17 Under the assumptions of Theorem 2.15, $y_o = (x_o, \lambda_o)$ is an isolated solution of $G(y) = 0$ if and only if (x_o, λ_o) is a simple quadratic turning point of (2.1).

This last result indicates that the only "stable" singular points of one parameter problems are simple quadratic turning points. Other singular points of (2.1), like those mentioned above, will be "lost" under arbitrary perturbations of (2.1). Of course Corollary 2.17 has important implications for the convergence of Newton's method or Newton-like methods (see Section 4).

Finally we remark that the numerical approach discussed in Section 4 provides an approximation to the matrix E and hence approximations to $\psi_1^T g_\lambda^0$ and a_g will be known.

3. Two parameter problems

We consider now the problem

$$(3.1) \quad f(x, \lambda, \alpha) = 0 \quad f : \mathbb{R}^n \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^n$$

introduced in Section 1. To analyse the singular points of (3.1) let us introduce the extended system

$$(3.2) \quad F(y, \alpha) : = \begin{bmatrix} f(x, \lambda, \alpha) \\ \psi^T f_x(x, \lambda, \alpha) \phi \end{bmatrix} = 0$$

where ϕ and ψ are singular vectors of $f_x(x, \lambda, \alpha)$ (cf. (2.12), (2.13)). Let us assume that, for fixed α , say $\alpha = \alpha_0$, $f(x, \lambda, \alpha_0) = 0$ has a simple quadratic turning point at $(x_0, \lambda_0, \alpha_0)$. Then $F_y(y_0, \alpha_0)$ is nonsingular (Corollary 2.17), and (y_0, α_0) is a regular point of (3.2). Now, if α is allowed to vary, the Implicit Function Theorem ensures the existence of a path of regular solutions of (3.2), in a neighbourhood of (y_0, α_0) , which can be parameterised by α . We call $y(\alpha)$ the *fold curve* and the solution surface of (3.1) near a fold curve is called a *fold* (see Figure 3.1).

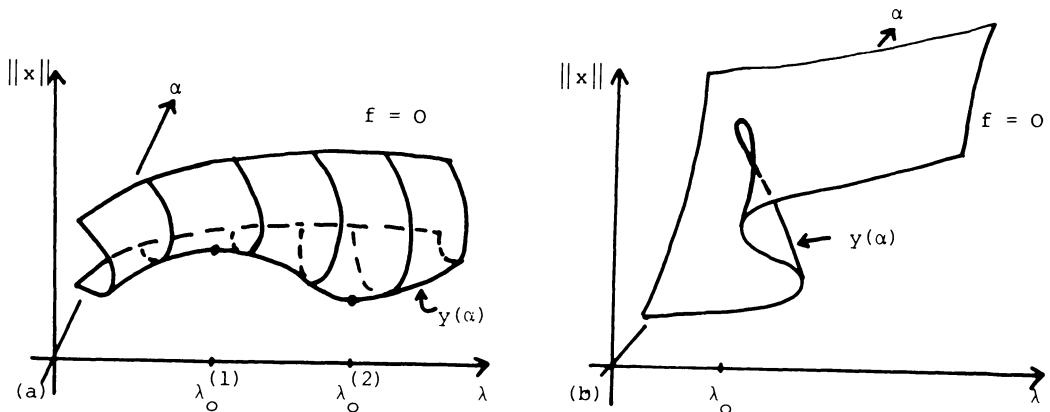


Figure 3.1 Different types of fold curve and solution surface for (3.1).

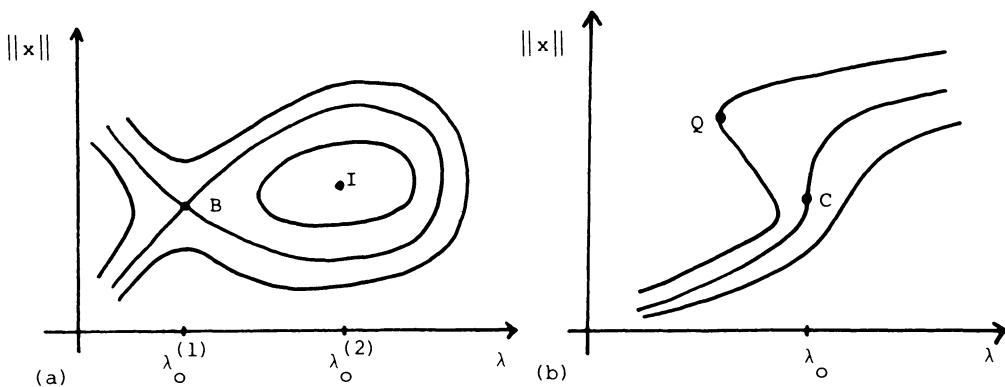


Figure 3.2 Solution diagrams for (3.1) for various values of α , showing a quadratic turning point (Q), a bifurcation point (B), a point of isola formation (I), and a cubic turning point (C).

Clearly the points $y(\alpha)$ on the fold curve remain regular points of (3.2) provided $\psi_1^T f_\lambda^O \neq 0$ and $a_f : = \psi_1^T f_{xx}^O \phi_1 \phi_1 \neq 0$ and hence they can be computed using standard continuation methods applied to $F(y, \alpha) = 0$.

What happens if one of these conditions is violated? The situation is described by the following theorem.

Theorem 3.3 Assume (2.2) holds and that $F(y_o, \alpha_o) = 0$. Then (y_o, α_o) is a simple turning point of $F(y, \alpha) = 0$ if and only if

$$\text{either } A) \quad a_f \neq 0, \quad \psi_1^T f_\lambda^O = 0,$$

(3.4)

$$\psi_1^T f_\alpha^O \neq 0, \quad (\text{see Fig. 3.1(a)});$$

$$\text{or } B) \quad a_f = 0, \quad \text{rank}(E) = 1,$$

(3.5)

$$\psi_1^T f_\lambda^O \neq 0, \quad b_{f\hat{\alpha}} : = \psi_1^T f_{xx}^O \phi_1 z_\alpha + \psi_1^T f_{x\hat{\alpha}}^O \phi_1 \neq 0, \quad (\text{see Fig. 3.1(b)}).$$

Here we have used the definitions

$$\begin{pmatrix} \hat{\lambda} \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} c_1 & c_2 \\ -c_2 & c_1 \end{pmatrix} \begin{pmatrix} \lambda \\ \alpha \end{pmatrix}, \quad c_1 = \psi_1^T f_\lambda^O, \quad c_2 = \psi_1^T f_\alpha^O,$$

and

$$\mathbf{f}_x^O \mathbf{z}_{\hat{\alpha}} = -\mathbf{f}_{\hat{\alpha}}^O, \quad \mathbf{z}_{\hat{\alpha}} \in V.$$

The proof is straightforward and is omitted. One can easily prove that the turning points in Theorem 3.3 are quadratic if,

in case A)

$$(3.6) \quad D_1 = b_{f\lambda}^2 - a_f c_{f\lambda} \neq 0 \quad (\text{see (2.10)}).$$

in case B)

$$(3.7) \quad \psi_1^T \mathbf{f}_{\lambda}^O \neq 0, \quad d_f \neq 0 \quad (\text{see (2.6) and [16]}).$$

Theorem 3.2 is important in that it indicates that bifurcation points, points of isolas formation, and cubic turning points (cusp points) which correspond to cases A), A), and B) respectively, can be computed using standard turning point algorithms for (3.2), see [13]. If (3.6) (respectively (3.7)) holds then the turning points are quadratic and the solution surfaces and curves for fixed values of α are given in Figures 3.1 and 3.2.

Theorem 3.2 also tells us precisely what numerical quantities to compute to distinguish between the two cases, namely, the quantities $\psi_1^T \mathbf{f}_{\lambda}^O$ and a_f . Our numerical approach, described briefly in Section 4, is specially designed to provide these quantities directly. Also we remark that these quantities tell us the shape of the solution surface near the fold curve, see [8].

Finally we discuss briefly the case where (2.2) fails. It is straightforward to verify that if, at (y_o, α_o) , $\text{rank}(\mathbf{f}_x^O) < n - 2$ then (y_o, α_o) cannot be a simple turning point of $F(y, \alpha) = 0$. Indeed we have the following theorem (cf. Theorem 4.15 of [8]).

Theorem 3.8 Assume (y_o, α_o) satisfies (3.2) and that (2.2) is not satisfied. Then (y_o, α_o) is a simple turning point of (3.2) if and only if

a) $\text{Null}(f_x^O) = \text{span}\{\phi_1, \phi_2\}$,

b) $\text{Range}(f_x^O) = \{z \in \mathbb{R}^n : \psi_1 z = 0, \psi_2 z = 0\}$,

(3.9) c) $A_f = \begin{pmatrix} \psi_1^T f_{xx}^O \phi_1 \phi_1 & \psi_1^T f_{xx}^O \phi_1 \phi_2 \\ \psi_2^T f_{xx}^O \phi_1 \phi_1 & \psi_2^T f_{xx}^O \phi_1 \phi_2 \end{pmatrix}$ is nonsingular,

d) $C_f = \begin{pmatrix} \psi_1^T f_\lambda^O & \psi_1^T f_\alpha^O \\ \psi_2^T f_\lambda^O & \psi_2^T f_\alpha^O \end{pmatrix}$ is nonsingular.

We omit the proof. Note that theorems 3.3 and 3.6 characterize all types of simple turning point of (3.2).

4. Numerical aspects

The basic idea of our approach is to follow paths of quadratic turning points, and so clearly we must be able to compute solutions of the extended system (2.11). Also, as was emphasised in Section 3, the quantities $\psi^T f_\lambda$ and $\psi^T f_{xx} \phi \phi$ play a key role in our attempt to gain information about the fold curve and the solution surface of (1.1) in the neighbourhood of the fold curve. Hence the numerical approach discussed here is motivated by the need to know these quantities at each computed point on the fold curve.

Any suitable continuation method can be used to compute the fold curve $y(\alpha)$. Here we describe only the computation of the solutions of (2.11) i.e.

(4.1) $G(y) \equiv \begin{pmatrix} g(x, \lambda) \\ \psi^T g_x(x, \lambda) \phi \end{pmatrix} = 0 \quad y = (x, \lambda) \in \mathbb{R}^{n+1},$

where ϕ and ψ are the singular vectors satisfying (2.12). We assume that a starting value close to (x_0, λ_0) , a quadratic turning point of (2.1), is known, and that a Newton-like method, is to be used to solve (4.1). We also assume that g_x is a full $n \times n$ matrix, i.e. it does not exhibit any special structure such as that which might arise from a discretization of an ordinary or partial differential equation. We note however that numerical techniques for such

problems based on following paths of singular points have been successfully used in the past, for example, [3],[4],[8],[14], and [16].

Clearly to set up $G(y)$ and $G_y(y)$ we need to know ϕ and ψ . In fact we neglect the terms $\psi_x^T g_x \phi$, $\psi g_{xx}^T \phi$, $\psi_\lambda^T g_x \phi$, and $\psi g_x^T \phi$ in the exact expression for $G_y(y)$, see (2.13), since $g_x \phi = O(\sigma)$, $\psi g_x^T = O(\sigma)$, and $\sigma \rightarrow 0$. The resulting Newton-like method has the form

$$(4.2) \quad B^{(k)} \begin{bmatrix} \Delta x^{(k)} \\ \Delta \lambda^{(k)} \end{bmatrix} = - \begin{bmatrix} g^{(k)} \\ \psi^{(k)} g_x^{(k)} \phi^{(k)} \end{bmatrix}$$

where

$$(4.3) \quad B^{(k)} : = \begin{bmatrix} g_x^{(k)} & g_\lambda^{(k)} \\ \psi^{(k)} g_{xx}^{(k)} \phi^{(k)} & \psi^{(k)} g_{x\lambda}^{(k)} \phi^{(k)} \end{bmatrix}$$

and

$$\lambda^{(k+1)} = \lambda^{(k)} + \Delta \lambda^{(k)}, \quad x^{(k+1)} = x^{(k)} + \Delta x^{(k)} \quad k = 0, 1, 2, \dots .$$

The iteration (4.2) will exhibit quadratic convergence provided the usual conditions are satisfied (see Corollary 2.17 and Theorem 3.4 in [5]).

Also we do not need to find the exact singular vectors $\phi^{(k)}, \psi^{(k)}$ at each step of the iteration. We can use approximate singular vectors, which are produced using a technique introduced in [11], and discussed in [8],[15]. Again the quadratic convergence is not affected. The details of the procedure (where, for convenience, the superscript k is dropped) are as follows. We assume that $\text{rank}(f_x) \geq n - 1$.

(i) Set up g , g_λ and g_x .

(ii) Perform a LU factorization of g_x i.e.

$$g_x = PLUQ^T$$

where P and Q are permutation matrices, L and U are lower and upper triangular matrices respectively. It is important to use a factorization technique which forces a small pivot to the (n,n) position of U , see for example [2] and [8]. Thus we may write

$$(4.4) \quad L = \begin{bmatrix} L_{11} & 0 \\ L_{12} & 1 \end{bmatrix}, \quad U = \begin{bmatrix} U_{11} & U_{12} \\ 0 & \epsilon \end{bmatrix}$$

and we assume, recall assumptions (2.2), that L_{11} and U_{11} are well-conditioned.

(iii) Approximations to the singular vectors ϕ and ψ are given by,
 L_{11} ,

$$(4.5) \quad \hat{\psi}^T : = (-L_{21}L_{11}^{-1}, 1)^T, \quad \hat{\phi} : = Q \begin{pmatrix} -U_{11}^{-1} & U_{12} \\ 1 & \end{pmatrix}$$

(Note that this assumes that the ϵ in (4.4) is "small").

(iv) Set up $\hat{\psi}^T g_{xx} \hat{\phi} \hat{\phi}$ and $\hat{\psi}^T g_{x\lambda} \hat{\phi}$.

(v) Solve (4.2). Since we already have a LU factorization of g_x , a LU factorization of $B^{(k)}$ in (4.2) can be easily obtained. The key point is to avoid any division by ϵ , and to only solve systems involving L_{11} and U_{11} . We have

$$B^{(k)} = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{12} & 1 & 0 \\ L_{13} & 0 & 1 \end{bmatrix} \begin{bmatrix} U_{11} & U_{12} & U_{13} \\ 0 & \epsilon & U_{23} \\ 0 & U_{32} & U_{33} \end{bmatrix}$$

where $L_{13}, U_{13}, U_{23}, U_{32}$, and U_{33} are easily obtained. The 2×2 matrix

$$(4.6) \quad \hat{E} : = \begin{bmatrix} \epsilon & U_{23} \\ U_{32} & U_{33} \end{bmatrix}$$

is easily shown to approximate E given by (2.16), with $U_{23} \approx \hat{\psi}^T f_\lambda$ and $U_{32} \approx \hat{\psi}^T f_{xx} \hat{\phi} \hat{\phi}$. Hence, from Corollary 2.17, \hat{E} is nonsingular at a quadratic turning point. Standard (block) forward and back substitutions provide the solution to (4.2).

A work count shows that, assuming g_x , g_λ , $g_{\lambda x}$, g_{xx} are known analytically, the work involved in steps (ii) - (v) is, essentially,

- (1) 1 LU factorization of g_x .
- (2) 6 forward or back substitutions.
- (3) the cost of computing $\hat{\psi}^T g_{xx} \hat{\phi}$ and $\hat{\psi}^T g_{\lambda x} \hat{\phi}$.

Thus points on the fold curve $y(\alpha) = (x(\alpha), \lambda(\alpha))$ of (3.1) can be computed. Also the values of $\psi_1^T f_\lambda(\alpha)$, and $\psi_1^T f_{xx} \phi_1 \psi_1(\alpha)$ can be readily monitored, using (4.6), to check for sign changes. Care must be taken, however, in the normalization of ψ_1 and ϕ_1 and this is discussed in [8].

Finally we remark that in past work, [15], we have used approximate eigenvectors instead of singular vectors. However it is well-known that singular vectors are better conditioned than eigenvectors, [6], and that the vectors in (4.5) are good approximations to singular vectors. It is also known that if a matrix has LU factors given by (4.4) then it has a singular value $\leq \epsilon$, and this justifies the Newton-like method which uses (4.3) rather than the full Jacobian.

Acknowledgements

The research of A.S. was supported by U.S. Army Contract 2FCZ519. The research of A.D.J. was supported by the National Science Foundation, the Office of Naval Research, the Army Research, the Air Force Office of Scientific Research, the University of Toronto, NSERC Canada and the British Council.

References

- [1] Bazley, N.W. and Wake, G.C. The disappearance of criticality in the theory of thermal ignition. ZAMP, 29, (1979), p.971-976.
- [2] Chan, T.F., On the existence and computation of LU-factorizations with small pivots (to appear in Maths. Comp.).
- [3] Cliffe, K.A., Numerical calculations of two-cell and single cell Taylor flows. (To appear in J. Fluid Mech.).

- [4] Cliffe, K.A. and Spence, A., The Calculation of High Order Singularities in the Taylor Problem (this proceedings).
- [5] Dennis, J.E. and Moré, J.J., (1977) Quasi-Newton methods, motivation and theory. SIAM Rev. 19, 46-89.
- [6] Golub, G.H. and Wilkinson, J.H., (1976) Ill-conditioned eigensystems and the computation of the Jordan canonical form. SIAM Rev. 18, p.578-619.
- [7] Heinemann, R.F. and Poore, A.B., (1981) Multiplicity, Stability and Oscillatory Dynamics of the Tubular Reactor, Chem. Eng. Sci. 36, pp.1411-1419.
- [8] Jepson, A. and Spence, A., (1982) Folds in solutions of two parameter systems and their calculation: Part I. Stanford University Technical Report, (submitted to SIAM JNA).
- [9] Jepson, A. and Spence, A., Paths of Singular Points and their Computation (this proceedings).
- [10] Jepson A. and Spence, A., (1983) The numerical solution of nonlinear equations having several parameters, Part I: Scalar Equations. (submitted).
- [11] Keller, H.B., Singular Systems, Inverse Iterations and least squares (private communication).
- [12] Keller, H.B. and Szeto, R.K-H., (1980) Calculation of flows between rotating disks, in "Computing Methods in Applied Sciences and Engineering" ed. R. Glowinski and J.L. Lions, North Holland, p.51-61.
- [13] Melhem, R.G. and Rheinboldt, W.C. (1982) A comparison of methods for determining turning points of nonlinear equations, Computing, 29, p.201-226.
- [14] Rheinboldt, W.G. and Burkardt, J.V., (1983) "A locally parameterized continuation process", ACM TOMS, 9, p.215-235.
- [15] Spence, A. and Jepson, A., (1982) The numerical computation of turning points of nonlinear equations in "Treatment of Integral Equations by Numerical Methods" ed. C.T.H. Baker, G.F. Miller, Academic Press London pp.169-189.
- [16] Spence, A. and Werner, B. (1982) Nonsimple turning points and cusps, IMA J. of Numer. Anal. 2, p.413-427.

- [17] Uppal, A., Ray, W.H. and Poore, A.B. (1974) On the dynamic behaviour of continuously stirred tank reactors, Chem. Eng. Sci., 29, p.967-985.

Alastair Spence,
School of Mathematics,
University of Bath,
BATH,
BA2 7AY,
U.K.

Allan D. Jepson,
Department of Computer Science,
University of Toronto,
TORONTO, M5S 1A7,
Canada.

Mode jumping of imperfect, buckled, rectangular plates.

H. Steinrück, H. Troger, R. Weiss (Wien)

1. Introduction.

One of the most fascinating phenomena in the buckling and post-buckling behavior of rectangular plates is mode jumping. By mode jumping we understand a sudden change in the wave number of the buckled deflection surface during a loading process.

Different types of loadings can be responsible for mode jumping, depending on the boundary conditions employed. We shall consider here two typical situations.

Case A: This case, where there is only one loading p , acting in the plane of the plate, is related to the classical mode jumping experiment by Stein [1]. Steins fundamental findings were: A plate with width 1 and length l will buckle with five half waves at a critical value p_k , provided l is chosen slightly smaller a critical value l_c , for which the linearized problem has a double eigenvalue. However, when p is increased beyond a second critical value p_s , then the deflection surface changes suddenly and the number of half waves increases by one.

A theoretical explanation of this phenomenon has been given by D. Schaeffer and M. Golubitsky in [1], who found that the boundary conditions at the four edges of the plate play a crucial role in its occurrence. They show that clamped boundaries at the edges where p is applied and simply supported boundaries at the other edges are sufficient for the described phenomenon to occur. On the other hand, mode jumping as described here cannot occur when all four edges of the plate are simply supported.

Case B: In this case mode jumping can occur even if we have simply supported boundaries at all edges of the plate. But additionally to the loading p transversal loads must be applied (the loads μ and ν in Fig. 1). Again it is necessary to chose the length l of the plate in such a way that it is close to a value l_c for which a double eigenvalue occurs ([2]).

After shortly indicating how to derive the bifurcation equations, which in the terminology of Catastrophe

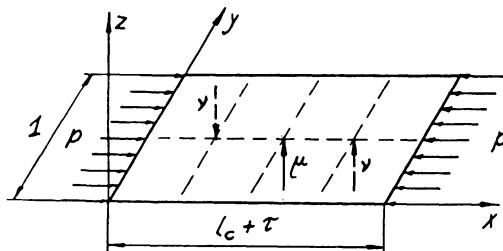


Fig.1: Plate at length $l=l_c+\tau$ close to a double eigenvalue (p ...thrust, μ, ν ...transversal loads).

theory ([3]) correspond to a Double Cusp catastrophe, we shall explain how we can do a physically meaningful analysis of the imperfect plate. Here we shall use the concept of Restricted Generic Bifurcation ([4]).

The main part of our paper is made up by the numerical results which, as we believe, give a good qualitative understanding of the behavior of the buckled plate when varying different parameters.

2. Plate equations and boundary conditions.

Mode jumping happens in the post-buckling range and is therefore basically a nonlinear phenomenon. An important question now is: Which mechanical and mathematical model of the plate should be used? Note that we are only interested in the buckled states, i.e. in the equilibrium positions of the plate. The transient process when the system changes from one equilibrium position to another is not of interest in our study. Since all forces acting on the plate are conservative and can be derived from a potential, the problem can be approached in the spirit of Catastrophe theory ([3]). This means also that only divergence bifurcations can occur. The appropriate mathematical plate model then seems to be the von Karman plate equations. Contrary to a wide spread belief these equations do not only follow from ad hoc assumptions in plate theory but also can be derived by an asymptotic procedure from the equations of three-dimensional nonlinear elasticity ([5]). We use them in non-dimensional variables ([6]) for the displacement w in z direction and the stress function f in the form

$$\begin{aligned} \Delta^2 f + \frac{1}{2}[w, w] &= 0 && \text{in } \Omega \\ \Delta^2 w - [w, f] + pw_{xx} &= 0, \end{aligned} \tag{1}$$

where $\Omega = \{(x,y), 0 < x < 1, 0 < y < 1\}$ is the domain of the plate. The operators are

$$\begin{aligned} \Delta^2 u &= u_{xxxx} + 2u_{xxyy} + u_{yyyy} \\ [u, v] &= u_{xx}v_{yy} - 2u_{xy}v_{xy} + u_{yy}v_{xx}. \end{aligned}$$

On the edges of the plate $\partial\Omega = \{(x,y), x = 0, 1 \quad 0 \leq y \leq 1, y = 0, 1 \quad 0 \leq x \leq 1\}$, we have the following boundary conditions:

Case A: $w = 0, \Delta w = 0$ at $y = 0, 1 \quad 0 \leq x \leq 1$ (simply supported) (2)

$w = 0, \frac{\partial w}{\partial n} = 0$ at $x = 0, 1 \quad 0 \leq y \leq 1$ (clamped)

$\frac{\partial f}{\partial n} = 0, \frac{\partial(\Delta f)}{\partial n} = 0$ at all boundaries. (3)

(Here n designates the direction normal to the boundary.)

$$\begin{aligned} \text{Case B: } w &= 0, \Delta w = 0 \\ &\quad \text{at all edges} \\ f &= 0, \Delta f = 0. \end{aligned} \tag{4}$$

A short comment on the boundary conditions seems to be appropriate. (4) are the boundary conditions usually employed for simply supported edges. However, the conditions (3) for the stress function f represent a more suitable approximation to the physical situation in the experiment by Stein [1]. The boundary conditions for the stress function enter into the calculation of the operator Δ^{-2} , which is required to eliminate f in (1). From the first equation of (1) we obtain

$$f = -\frac{1}{2} \Delta^{-2} [w, w]. \tag{5}$$

The calculation of Δ^{-2} is explicitly given in [1] for case A, where it is shown that the use of (3) allows a calculation of Δ^{-2} without infinite series expansions. The calculation of Δ^{-2} in case B is performed in [6] or [8].

Introducing (5) into the second equation of (1) yields

$$\Delta^2 w - [w, -\frac{1}{2} \Delta^{-2} [w, w]] + pw_{xx} = 0 \tag{6}$$

with the boundary conditions for w either given by case A or case B.

3. Bifurcation equations.

Equation (6) with the boundary conditions (2) or (4) is a bifurcation problem: When increasing the parameter p quasistatically from zero we reach a critical value p_k for which the flat configuration of the plate becomes unstable. Physically this means that for a small disturbance of the flat configuration the plate will pass into a buckled equilibrium position. Mathematically it means that nontrivial solutions bifurcate from the trivial solution. The critical value p_k of the parameter is obtained from the linear eigenvalue problem (with boundary

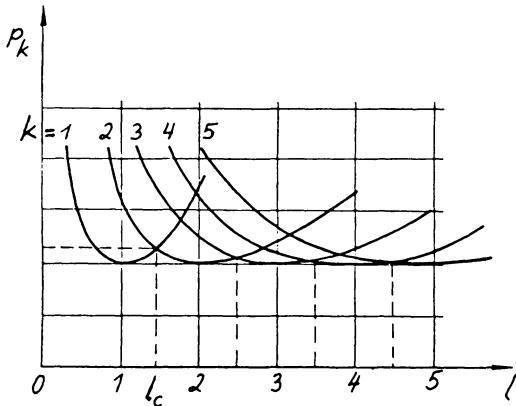


Fig.2: Eigenvalue curves p_k versus l with double eigenvalues occurring for $l = l_c$.

conditions (2) or (4))

$$\Delta^2 h + ph_{xx} = 0, \quad (7)$$

resulting from the linearization of (6) at the flat plate $w=0$. (7) is solved for the boundary conditions of case A in [1] and of case B in [3].

In Fig. 2 eigenvalue curves are shown for case B, where the critical load p_k is plotted as a function of the length of the plate. It can be seen that generically simple eigenvalues occur. However, for special critical values of the length of the plate

$$l_c = (k(k+2))^{\frac{1}{2}} \quad \text{in case A} \quad (8)$$

$$l_c = (k(k+1))^{\frac{1}{2}} \quad \text{in case B}, \quad (9)$$

double eigenvalues occur. The existence of these double eigenvalues is responsible for the mode jumping phenomenon because a small perturbation of the system in the neighborhood of a double eigenvalue can lead to qualitatively different buckled states of the plate.

By means of the Liapunov-Schmidt-method ([1]) with the ansatz

$$w = r\phi_1(x,y) + s\phi_2(x,y) + W(r,s;x,y) \quad (10)$$

where ϕ_1 and ϕ_2 are the eigenfunctions of (7), we can reduce (6) to a system of two nonlinear algebraic equations ([1])

$$\begin{aligned} r^3 + \alpha rs^2 - \lambda r &= 0 \\ s^3 + \beta r^2 s - \lambda s &= 0 \end{aligned} \quad (11)$$

in the two unknowns r and s . α and β are constants depending on the boundary conditions and λ is a new loading parameter, which is zero for $p = p_k$. The function $W(r,s;x,y)$, which is important to give a correct splitting between critical and noncritical variables ([3]) can be neglected in the case of the plate. This follows from the fact that (11) is three determinate ([3]). (See also the discussion in [2]).

(11) are the one parameter bifurcation equations at the double eigenvalue $r = s = 0$. They describe the behavior of the perfect plate, which is highly non generic, as perfect structures are exceptional cases. Nevertheless they are of great practical importance because they form the organizing center for the physically relevant imperfect situations.

4. Restricted Generic Bifurcation.

In practical situations one not only wants to solve (11) but one also wants to obtain the solutions of (11) under all or at least under physically meaningful perturbations of (11). For the first case the answer can be found in the theory of universal unfoldings ([3]), whereas in the second case the concept of Restricted Generic Bifurcation ([4]) is applicable. In [3] it is shown that (11) corresponds to a Double Cusp catastrophe the universal unfolding of which requires an eight parameter family. Such a high dimensional family of parameters is quite difficult to realize physically ([3]) and secondly the study of a bifurcation diagram in an eight dimensional parameter space seems to be a formidable task. By a bifurcation diagram we understand a partition of the parameter space into open and dense sets of systems with qualitatively similar behavior. These difficulties can be circumvented to some extend by applying the concept of Restricted Generic Bifurcation as developed in [4,8,9], where fewer parameters are considered, than would be required for a universal unfolding. Typically all these parameters have a distinct physical meaning. In our case we have: (i) the in plane loading $\lambda = p - p_k$, (ii) the deviation τ of the length of the plate from the values given by (8) or (9) (iii) two cases of transversal loads μ and ν (Fig. 1). These parameters enter into (11) in the following way:

$$\begin{aligned} r^3 + ars^2 - \lambda r &+ \mu + h.o.t. = 0 \\ s^3 + Bsr^2 - \lambda s - \tau s + \nu + h.o.t. &= 0, \end{aligned} \quad (12)$$

where h.o.t. denotes higher order terms and where the scaling constants in front of λ, τ, μ, ν are set to 1. μ and ν also can be considered as the projections of a continuous loading $q(x,y)$ on the corresponding eigenfunctions ϕ_1 and ϕ_2 , i.e. $\mu = \epsilon h_1$ and $\nu = \epsilon h_2$ where $h_1 = \langle q, \phi_1 \rangle$ and $h_2 = \langle q, \phi_2 \rangle$ and $\langle u, v \rangle = \int_{\Omega} uv d\Omega$.

The problem (12) is analyzed in [3,7,8]. In addition, our own numerical results show that the only type of singularity occurring when λ is varied and τ, μ, ν are kept at fixed nonzero values is that of a limit point. From the engineering standpoint this is a very satisfactory situation because then additional small perturbations of the system (small compared to $|\tau|, |\mu|, |\nu|$) do not lead to qualitative changes in its behavior during the loading process, as would be the case if bifurcation points were present.

5. Numerical results.

We shall treat the two cases A and B as mentioned in the Introduction separately.

Let us start with case A. Here mode jumping is obtained without a transversal loading, just by increasing the thrust p alone. So we set $\mu = \nu = 0$. The only imperfection which we retain in (12) is the deviation of the length of the plate l from the critical value l_c given by (8). Then (12) reduces to

$$\begin{aligned} r^3 + \alpha r s^2 - \lambda r &= 0 \\ s^3 + \beta s r^2 - \lambda s - \tau s &= 0. \end{aligned} \tag{13}$$

It is shown in [1] that, due to the clamped boundary conditions, the "modal parameters" satisfy

$$1 < \alpha, \quad 1 < \alpha\beta, \quad \beta < 1. \tag{14}$$

Then the solutions of (13) are

$$(1) \quad r = 0, \quad s = 0$$

$$(2,3) \quad r = 0, \quad s = \pm (\lambda + \tau)^{\frac{1}{2}} \quad \lambda \geq -\tau \tag{15}$$

$$(4,5) \quad r = \pm \lambda^{\frac{1}{2}}, \quad s = 0, \quad \lambda \geq 0$$

$$(6,7,8,9) \quad r = \pm \left[\frac{-\tau\alpha + \lambda(1-\alpha)}{1 - \alpha\beta} \right]^{\frac{1}{2}}, \quad s = \pm \left[\frac{\tau + \lambda(1-\beta)}{1 - \alpha\beta} \right]^{\frac{1}{2}}$$

$$\frac{\tau\alpha}{1 - \alpha} \leq \lambda \leq \frac{\tau}{\beta - 1}, \quad \tau < 0, \tag{16}$$

with the following bifurcation points

$$B_1: \quad \lambda = 0 \quad r = 0 \quad s = 0$$

$$B_2: \quad \lambda = -\tau \quad r = 0 \quad s = 0$$

$$B_{3,4}: \quad \lambda = \frac{\tau\alpha}{1 - \alpha} \quad r = 0, \quad s = \pm \left[\frac{\tau\alpha}{1 - \alpha} \right]^{\frac{1}{2}} \tag{17}$$

$$B_{5,6}: \quad \lambda = \frac{\tau}{\beta - 1} \quad r = \pm \left[\frac{\tau}{\beta - 1} \right]^{\frac{1}{2}} \quad s = 0$$

The corresponding bifurcation diagram in the $\lambda - \tau$ plane indicating the number of solutions for a given pair of these parameters is Fig. 3.

To obtain a geometric understanding of these solutions, the variables r and s , which are the amplitudes of the corresponding mode shapes (Fig. 4) are represented as functions of λ in Fig. 5 and in an axonometric representation in Fig. 6. We pick a negative value of τ , i.e. the length of the plate is shorter than the critical value l_c for the first double eigenvalue. For $\lambda < 0$ ($p < p_k$) we see from Fig. 3, that there exists only one solution

namely the stable trivial solution $r = s = 0$ (Figs. 5,6). Increasing λ we reach at $\lambda = 0$ the first bifurcation point B_1 (Figs. 3,5,6) where the trivial solution becomes unstable and two stable r -solutions (Fig. 4 a) bifurcate which represent the physical configuration of the plate. Increasing λ further we reach B_2 on the trivial (unstable branch) where two s -solutions (Fig. 4 b) bifurcate. These are unstable until we arrive

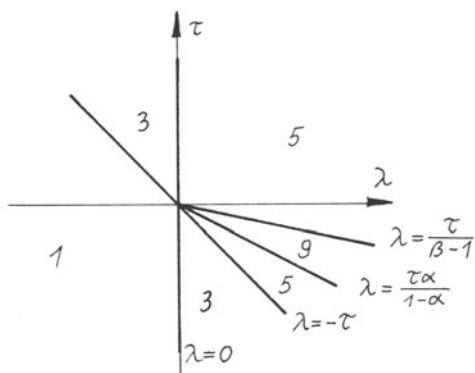


Fig.3: Bifurcation diagram giving the number of solutions of (13).



Fig.4: Eigenmodes of the first double eigenvalue for case A.

at $B_{3,4}$. At the bifurcation points B_3 and B_4 where secondary bifurcations occur four new solutions show up. This is due to the fact that we have set $\mu = v = 0$ which allows for the number of solutions to change by 4 rather than by 2. After $B_{3,4}$ the s -solutions are stable. However the plate is still in the r -mode until the bifurcation points $B_{5,6}$ are reached, where again secondary bifurcations occur. Now the r -solutions become unstable and the buckled plate has to move into a new stable equilibrium position, which, as can be seen from Figs. 5,6, is the s -mode. The transition will take place in form of jumps if $\lambda_{3,4} < \lambda_{5,6}$ as it is shown in our figures. If, however, $\lambda_{3,4} > \lambda_{5,6}$ the transition from r to s will be a smooth one. In Figs. 7 and 8 we show numerical results for the imperfect case, which is a plate with small transversal loads μ and v (Fig.1). We see from these figures that now only limit points are present which indicates that the problem is insensitive to small perturbations. This is not the case for the system of Figs. 5 and 6 where the bifurcation points indicate that the used mechanical model will still be sensitive to small perturbations.

Let us consider now case B, where we have simply supported boundaries at all edges of the plate. Now the modal parameters no longer satisfy

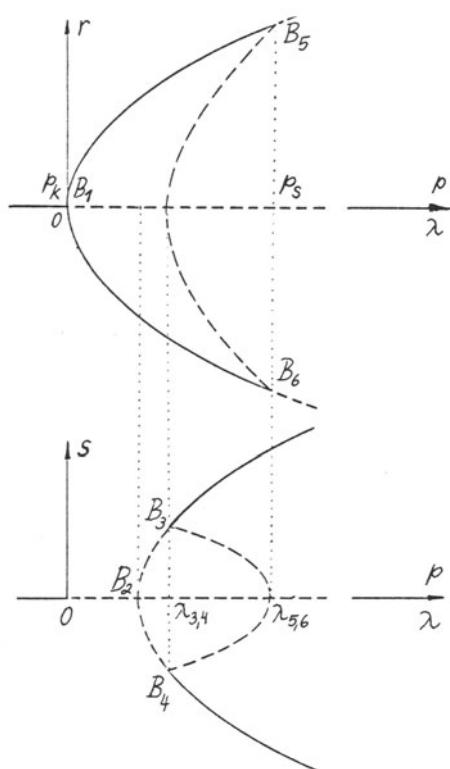


Fig.5: Amplitudes r and s of the eigenmodes of Fig.4 versus $p(\lambda)$ for $\tau < 0$ and $\mu = \nu = 0$ (case A). Mode jumping at p_s .

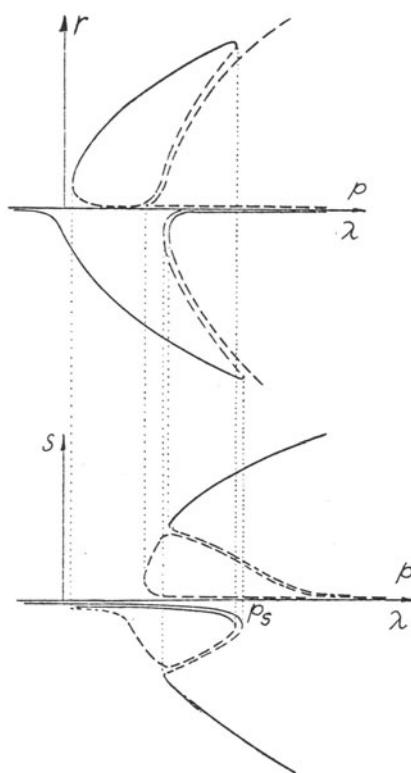


Fig.7: As Fig.5 but with transversal loading ($\mu \neq 0$, $\nu \neq 0$).

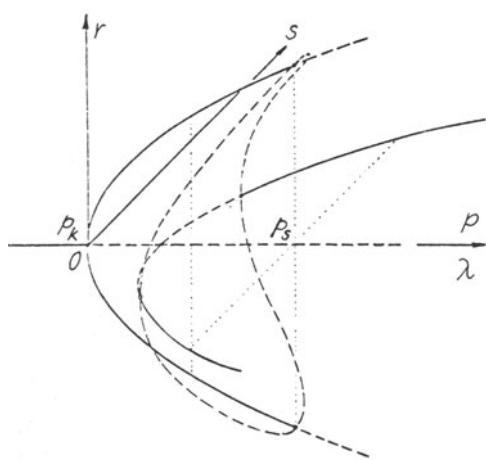


Fig.6: Axiometric representation of Fig.5.

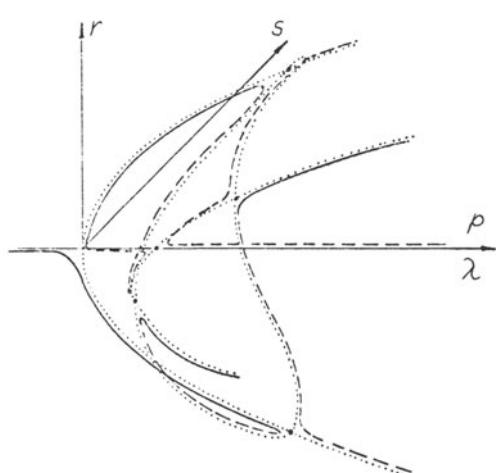


Fig.8: Axiometric representation of Fig.7.

(14) but instead the conditions $\alpha > 1$, $\beta > 1$ hold. The representations (15) for the case $\mu = v = 0$ are still valid provided (16) is replaced by $-\tau\alpha \leq \lambda(\alpha-1)$, and only the bifurcation points B_1 to B_4 occur. The solutions are depicted in

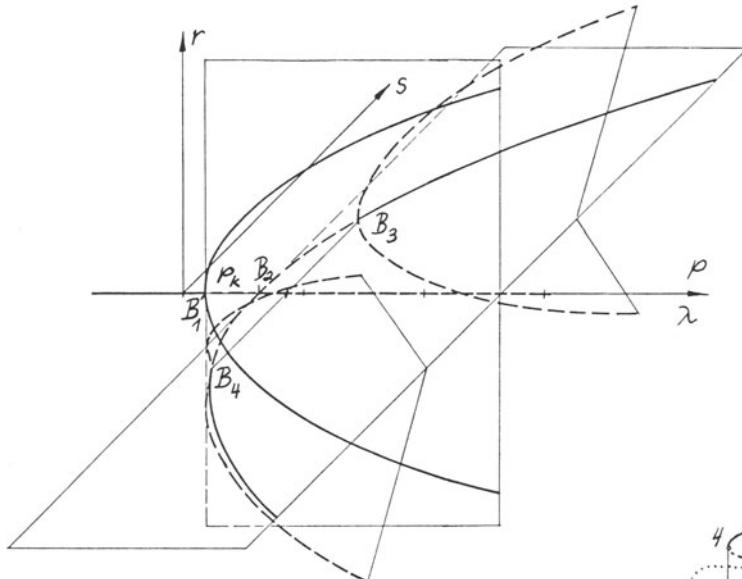
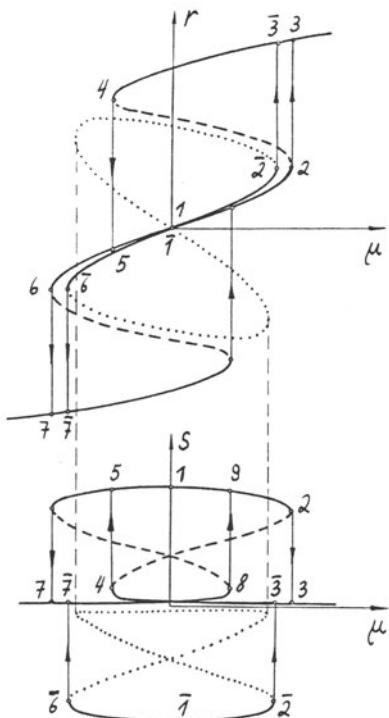


Fig. 9: Amplitudes r and s of the eigenmodes for case B ([2]). No mode jumping by variation of p .

fixed values of λ , $\tau < 0$ and v the force μ in the middle of the plate is varied. We see that by varying μ we can obtain buckled plate configurations which either have a dominating r - or s -mode or which are a superposition of both. Again only limit points are found in Fig. 10, which is a strong evidence for a meaningful engineering model. A thorough description of mode jumping in case B is given in [2], which also contains a short description of the path-following method ([10]) which was used to compute the curves presented here ([12]).

Fig. 9, for $\tau < 0$. The main difference to Fig. 6 is, that on the stable r -branches which bifurcate at B_1 no secondary bifurcations occur and therefore mode jumping only can happen if transversal loads are applied. This is the case in Fig. 10 where for



6. Summary.

Numerical results illustrating the mode jumping phenomenon of rectangular plates for two different cases of boundary conditions are presented.

Fig. 10 Mode jumping in case B by variation of μ . ($p > p_k$, $\tau < 0$, $v \neq 0$ are fixed).

From the diagrams in Figs. 6 and 9 one can see that a theoretical analysis of the problem as is given in [1] is of great importance because certain qualitative differences hardly could be explained from numerical results only. This is even more important if FEM-calculations of secondary bifurcations are made ([11]). The application of the concept of Restricted Generic Bifurcation proved to be quite successful because it requires only a simple model whereas a universally unfolding parameter family would lead to a rather complicated model of the problem.

References

- 1 Schaeffer D., M.Golubitsky, Boundary Conditions and Mode Jumping in the Buckling of a Rectangular Plate, Commun. Math. Phys. 69 (1979) 209 - 236.
- 2 Suchy H., H.Troger, R.Weiss, Numerical Study of Mode Jumping of Rectangular Plates, ZAMM submitted for publication.
- 3 Poston T., I.Stewart, Catastrophe Theory and its Applications, Pitman, London 1978.
- 4 Hale J.K., Restricted generic bifurcation, in: Nonlinear Analysis, Academic Press, New York, 1978, p. 83 - 98.
- 5 Ciarlet P.G., P.Rabier, Les Equations de von Karman, Lect. Notes in Math. 826, Springer, Berlin, Heidelberg, New York 1980.
- 6 Magnus R., T.Poston, On the full unfolding of the von Karman equations at a double eigenvalue, Math. Report 109, Battelle Advanced Stud. Cent., Geneva (Switzerland) 1977.
- 7 Schaeffer D., General introduction to steady state bifurcation, Lecture Notes in Math. 898, Springer, Berlin, Heidelberg, New York 1981, p. 13 - 47.
- 8 Chow S.N., J.K.Hale, J.Mallet-Paret, Applications of Generic Bifurcation I and II, Archive Rat. Mech. Anal. 59 (1975) 159 - 188 and 62 (1976) 209 - 235.
- 9 List S.E., Generic Bifurcation with Application to the von Karman Equations, J. Diff. Equ. 30 (1978) 89 - 118.
- 10 Schwetlick H., Numerische Lösung nichtlinearer Gleichungen, Oldenburg Verlag, München, Wien 1979.
- 11 Carnoy E.G., T.J.R.Hughes, Finite element analysis of the secondary buckling of a flat plate under uniaxial compression, Int.J.Non-Linear Mechanics 18 (1983) 167 - 175.
- 12 Steinrück H., Mode-jumping von Rechteckplatten, Diplomarbeit am Institut für Angewandte und Numerische Mathematik, TU-Wien, 1983.

H. Steinrück, H. Troger, R. Weiss, Institut für Mechanik, TU Wien, Karlsplatz 13, A-1040 Wien

Application of Bifurcation Theory to the Solution of
Nonlinear Stability Problems in Mechanical Engineering.

H. Troger (Wien)

1. Introduction.

A great deal of problems in engineering, but for example also in physics, biology or sociology deals with equilibrium solutions of nonlinear equations and their stability. By equilibrium solutions one can understand steady state solutions or time periodic solutions. However to be more specific we want to concentrate on the case which seems to be of greatest practical importance namely the case of loss of stability of a stable steady state equilibrium position. Some arbitrarily choosen examples in the above mentioned fields are in engineering buckling phenomena in structural mechanics ([24]) or the onset of a hunting motion of a railvehicle ([20]). In biology the change in the behavior of an animal from backing up into an attacking mode ([35]) or in sociology the sudden readiness of the whole population of a democratically governed country to go to war. In physics a good example is the behavior of a fluid layer heated from below ([11]). To all these examples it is common that a distinguished parameter is varied and that at a critical parameter value a qualitative change in the behavior of a system takes place, i.e. the original equilibrium looses its stability. The special parameter is for the buckling problem the magnitude of loading, for the railvehicle the speed, for the animal it could be a measure of the closeness of an invader to its nest, for the country getting ready to start a war it could be the amount of humiliation suffered from a dictatorial regime and for the fluid layer it is the temperature difference between the lower and the upper surface of the layer.

So far in engineering applications of stability theory mostly only the critical parameter value has been calculated. This in general only requires a linear investigation. Further information on the behavior of the system after having reached the critical parameter value requires a nonlinear stability analysis. Nonlinear stability investigations mostly done by means of the use of Liapunov functions ([15]) are not a straight forward matter and in many cases where a nonlinear stability analysis is indispensable it cannot be carried out. Numerical simulation which is then sometimes used is not only very expensive if it is done for realistic systems with many degrees of freedom but there is mostly also a certain uncertainty in the interpretation of qualitative changes in the numerical results.

Bifurcation theory, however, supplies us for many practically important cases with the necessary means not only to obtain the critical parameter value for the loss of stability but also to evaluate the post-bifurcation behavior of the system in a straight forward way. We shall see from practical-

ly relevant examples that the behavior of the system after the bifurcation had occurred can be of such a way that the evaluation of the stability limit (linear analysis) only is practically meaningless (see Fig. 18). Furthermore bifurcation theory not only allows to study the stability behavior of a fixed system but also to evaluate the influence of small changes (perturbations) of the system on its stability behavior. This is the practically important question of imperfection sensitivity. This also touches the important question how to set up a mathematical model which is robust against small perturbations ([34]).

The only limitation in the application of bifurcation theory is that the degeneracy of the system at the critical parameter value must not be too large. The degeneracy of a problem is expressed by its codimension ([24]), which is equal to the minimum number of parameters of a family in which the singular system must be embedded in order to be brought into a general (nondegenerate or generic) position. The codimension depends on the number of eigenvalues with realpart zero of the linearized system and the lowest order of the nonlinear terms. It is intuitively clear that in practically occurring situations low degeneracy and therefore also low codimension will prevail.

The important fact now is that the original problem of dimension n or infinity can be reduced to a problem of bifurcation equations of dimension k , where k is the number of eigenvalues with realpart zero of the linearized system. This reduction can be done by means of the theory of invariant manifolds ([7]) and will be shortly explained in chapter 2. These bifurcation equations completely describe the stability behavior of the original system locally around the bifurcation point. This also explains why, for example, in [1] there is such a detailed study of these low dimensional nonlinear systems. They in general govern the stability behavior of infinite or n dimensional systems, if only the degeneracy is small enough!

Furthermore it must be mentioned that up to a certain codimension, the value of which is different, for different mathematical descriptions (functions, differential equations) there exist only few cases of qualitatively different bifurcations which are more or less completely studied and classified.

It must however be admitted that there also exist engineering problems where higher degeneracies occur. These are, for example, due to optimization procedures ([32]) or due to special symmetries ([12,26]). Normally the treatment of singular cases of higher codimension makes sense only in parameterized families because for a special degenerate system a small perturbation of the system can make the singular case disappear whereas in a parameter family the singular (degenerate) case can appear in a robust way.

In treating a nonlinear stability problem one has in general to do three steps. The first is, as already mentioned the reduction of the original system to a system of bifurcation equations. Secondly one must transform the bifurcation system into Normal Form ([8]) and then embed it into a universal

parameter family which gives all qualitatively possible solutions. In chapter 3 some of these cases are listed which have been classified so far and therefore are immediately applicable for a stability analysis. Finally one tries to construct a bifurcation diagram in the parameter space. By a bifurcation diagram we understand a partition of the parameter space according to qualitatively different properties of the system. For many unfolded cases being classified such bifurcation diagrams exist ([1,4,24]).

Finally it should be mentioned that there exist many quite different bifurcation phenomena in many fields of science but on the other hand there are only very few, at least for low codimension, mathematically different forms which describe them. Therefore a classification of bifurcation phenomena in the sciences quite naturally is done by their mathematical description.

2. Reduction to bifurcation equations.

An important question to be answered at the beginning of every stability investigation is: How complicated must be the mechanical model and consequently the mathematical model in order to describe the physical phenomenon properly. For mechanical systems and many others as well we have quite naturally as mathematical models differential equations. These are ordinary, partial, hybrid or integro differential equations. Sometimes, however, differential equations are unnecessarily complicated and can be replaced by mappings. Especially this case is quite often found in structural mechanics where it gives a completely satisfactory description of the problem if one has static loadings that can be derived from a potential and secondly if one is more interested in the equilibrium positions of the system and their stability than in the transient behavior taking place during the transition of the system from one equilibrium position to another.

We shall assume now that the mathematical model is either given by differential equations or by a functional describing the behavior of the system.

Suppose now, in order to be more specific, the states of a mechanical system are determined as the solutions of the functional equation

$$G(u, \lambda) = 0, \quad (1)$$

where u is an element of a Banach space E , $\lambda \in L \subseteq \mathbb{R}^1$ is a parameter and G is a nonlinear mapping from $E \times L$ to another Banach space H . Now suppose that (u_0, λ_0) is a solution of (1), i.e. $G(u_0, \lambda_0) = 0$. We now calculate the Frechet derivative $G_u(u_0, \lambda_0)$, which is a linear mapping from E to H . From the Implicit Function theorem ([26]) follows that if G_u has an inverse then for $|\lambda - \lambda_0|$ sufficiently small there exists a smooth unique curve of solutions $u(\lambda)$ through (u_0, λ_0) and therefore a bifurcation cannot occur. Bifurcation, however, can occur if $G_u(u_0, \lambda_0)$ is not invertible. For applications it is sufficient to restrict to problems, where G_u is a Fredholm operator of index zero ([26]). If G_u at $\lambda = \lambda_c$ is not invertible, then one or several eigenvalues

cross the imaginary axis. Suppose now we have a k -fold eigenvalue for G_u the kernel of which is spanned by the elements $\{\phi_1, \dots, \phi_k\}$. We now represent u by the following ansatz

$$u = \sum_{i=1}^k x_i \phi_i + U(x_1, \dots, x_k), \quad (2)$$

where x_1, \dots, x_k are the essential variables and $U(x_1, \dots, x_k)$ has to be determined from the Liapunov-Schmidt procedure ([11, 12, 26]) after introducing (2) into (1) and gives the correct splitting of the infinite dimensional problem into a finite dimensional problem with k essential variables ([24]). As result we obtain a set of k nonlinear algebraic equations

$$F_i(x_1, \dots, x_k, \lambda) = 0 \quad i = 1, \dots, k \quad (3)$$

the solutions of which are in one-to-one correspondence with the solutions of the original system, if we stay sufficiently close to the bifurcation point. An important property of U in (2) is that

$$U(x_1, \dots, x_k) = O(|x_1 + \dots + x_k|^2), \quad (4)$$

from which follows that in many problems U can be set to zero for reasons of determinacy ([24]).

If we have a dynamic description

$$u_t = M(u, \lambda) \quad (5)$$

of the physical phenomenon we would have to use Center Manifold Theory ([3, 7, 8, 26, 33]) in order to reduce (5) to a set of k nonlinear ordinary differential equations

$$\dot{x}_i = f_i(x_1, \dots, x_k, \lambda). \quad i = 1, \dots, k. \quad (6)$$

The underlying geometric idea is that the flow of system (5) is a contraction in the direction of the eigenvectors which correspond to the negative eigenvalues to an invariant manifold namely the center manifold on which the interesting dynamics occur. Again for many practical problems a property similar to (4) holds also for the center manifold which then reduces the necessary computations quite considerably ([8, 33]).

3. Normal Forms, Universal Unfoldings and Bifurcation Diagrams.

Thanks to the reduction process we only have to deal with either (3) or (6). First one tries to simplify them by using the Normal Form Theorem ([8]). The idea is, to make a, in general, nonlinear transformation of variables which transforms the nonlinear terms into their simplest form. For low degenerate cases this is not too difficult to do, as the simple normal forms are known and are the classified cases. Mostly also a linear transformation does the job. If we have simple eigenvalues then the transformation to normal form is trivial.

The next step is to find a versal parameter family of bifurcation equations with the minimum number of parameters which includes all qualitatively possible cases of the given class. This is called a universal family or a universal unfolding. For the two cases of mathematical descriptions (3) and (6) we give now for the practically important cases the universal unfoldings.

Instead of working with the equilibrium equations (3) it is more convenient to introduce the potential $V(x_1, \dots, x_k, \lambda)$ from which (3) can be derived. It follows

$$\frac{\partial V}{\partial x_i} = F_i(x_1, \dots, x_k, \lambda) = 0. \quad (7)$$

Up to codimension $c = 4$ there exists a complete classification, which includes only seven different potentials in one or two variables, which are the socalled "Elementary Catastrophes" ([24]). They are

$$c = 1: \quad \frac{x_1^3}{3} + \varepsilon_1 x_1 \quad (8)$$

$$c = 2: \quad \frac{x_1^4}{4} + \varepsilon_1 \frac{x_1^2}{2} + \varepsilon_2 x_1 \quad (9)$$

$$c = 3: \quad \frac{x_1^5}{5} + \varepsilon_1 \frac{x_1^3}{3} + \varepsilon_2 \frac{x_1^2}{2} + \varepsilon_3 x_1 \quad (10)$$

$$x_1^3 - 3x_1 x_2^2 + \varepsilon_1 (x_1^2 + x_2^2) + \varepsilon_2 x_1 + \varepsilon_3 x_2 \quad (11)$$

$$x_1^2 x_2 + x_2^3 + \varepsilon_1 (x_1^2 + x_2^2) + \varepsilon_2 x_1 + \varepsilon_3 x_2 \quad (12)$$

$$c = 4: \quad \frac{x_1^6}{6} + \varepsilon_1 \frac{x_1^4}{4} + \varepsilon_2 \frac{x_1^3}{3} + \varepsilon_3 \frac{x_1^2}{2} + \varepsilon_4 x_1 \quad (13)$$

$$x_1^2 x_2 + x_2^3 + \varepsilon_1 x_1^2 + \varepsilon_2 x_2^2 + \varepsilon_3 x_1 + \varepsilon_4 x_2. \quad (14)$$

The ε_i are the parameters and x_i the variables. If the parameters are set to zero we have the degenerate situation and consequently the singular bifurcation equations. It is important to note that higher order terms in the potentials (8 - 14) do not influence qualitatively the behavior of the system locally around the bifurcation point because of determinacy ([24]) of the above listed potentials. It further should be noted that for other potentials with two variables which are important in applications and which are not among those listed the codimension increases rapidly. For example the singular potential

$$V = \frac{x_1^4}{4a} + \frac{1}{2} x_1^2 x_2^2 + \frac{x_2^4}{4b} \quad (15)$$

which will be found for the plate buckling problem in chapter 4 has codimension $c = 8$. A universal unfolding for this case is given in [24]. As the study of an eight parameter family is quite complicated we shall use in this case an other concept namely the concept of Restricted Generic Bifurcation ([6,17]), which for engineering purposes seems to be more appropriate. By Restricted

Generic Bifurcation we understand an unfolding of the bifurcation equations with less parameters than would be necessary for a universal unfolding and the condition is that at a bifurcation point the number of solutions is allowed to change only by two.

A different type of complete unfolding with great practical importance is given in [4], where a distinguished parameter, which is in general λ in equations (3) or (6) is kept separated from the unfolding parameters. This has the advantage that the physically meaningful parameter λ does not mix with variables or imperfection parameters. We give an application in chapter 4 for the driving behavior of a tractor semitrailer.

Let us turn now to the case where we have dynamic bifurcation equations in the form of (6). Here up to codimension two a more or less complete classification can be given ([1,8,9])

$c = 1$: (a) One zero root:

$$\dot{x}_1 = ax_1^2 + \epsilon_1 + O(x_1^3) \quad (16)$$

(b) One purely imaginary pair:

$$\dot{z} = z(i\omega + \epsilon_1 + az\bar{z} + O(z^4)) \text{ with } z = x_1 + ix_2 \quad (17)$$

$c = 2$: (a) One zero root:

$$\dot{x}_1 = ax_1^3 + \epsilon_1 x_1 + \epsilon_2 + O(x_1^4) \quad (18)$$

(b) One purely imaginary pair:

$$\dot{z} = z(i\omega + \epsilon_1 + \epsilon_2 z\bar{z} + az^2\bar{z}^2 + O(z^6)), z = x_1 + ix_2 \quad (19)$$

(c) One double zero root:

$$\dot{x}_1 = x_2 \quad (20)$$

$$\dot{x}_2 = \epsilon_1 x_1 + \epsilon_2 x_2 + ax_1^3 + bx_1^2 x_2 + O(|x_1 + x_2|^4)$$

(d) One zero root and one purely imaginary pair:

$$\dot{x}_1 = \epsilon_1 x_1 + ax_1^3 + bx_1 x_2^2 + O(|x_1 + x_2|^4) \quad (21)$$

$$\dot{x}_2 = \epsilon_2 x_2 + dx_1^2 x_2 + x_2^3 + O(|x_1 + x_2|^4)$$

(e) Two purely imaginary pairs:

$$\dot{x}_1 = x_1(\epsilon_1 + x_1 + ax_2 + B_1(x) + O(|x_1 + x_2|^3))$$

$$\dot{x}_2 = x_2(\epsilon_2 + bx_1 + x_2 + B_2(x) + O(|x_1 + x_2|^3)) \quad (22)$$

with $B_1(x) = m_{11}x_1^2 + m_{12}x_1 x_2 + m_{13}x_2^2$. a, b, d and m_{ij} are given

constants. In the cases (d) and (e) of $c = 2$ the azimuthal components of the vector fields have not been written down. For a first analysis they can be ignored. However in cases where, for example, chaotic motions can occur they are important.

For the case of codimension three except for simple eigenvalues practically nothing is known.

The practical use of the universally unfolded singular cases (8 - 14) and (16 - 22) is given by Bifurcation Diagrams. They are a partition of the parameter space according to different qualitative properties of the system. In Fig. 1 the bifurcation diagram for the cusp catastrophe (9) is given in the lower part of the figure by the semicubical parabola. It separates the regions in the parameter space with one and three solutions. We shall make frequently use of it later. The cusp catastrophe (Fig. 1) is the most complicated singular potential for which a complete representation of the behavior and the parameter space can be given. If for example we have (12), which is the hyperbolic umbilic catastrophe we can represent only the bifurcation diagram in three space, which is shown in Fig. 13. Again the different number of solutions in different regions are indicated. An application in shell buckling is given in chapter 4. In Fig. 24 a bifurcation diagram for a dynamic system with a double zero root (42) is given.

4. Some applications in mechanical engineering.

A) Systems governed by potentials.

We restrict the application of this class of systems to the important class of buckling problems of elastic structures under conservative loading.

A natural classification is given whether we have problems with simple, double or multiple eigenvalues.

a) Simple eigenvalues.

In this class of problems fall the buckling problems of simple rods. There are basically two different types of bifurcation points: (i) symmetric ones, which are related to the Cusp catastrophe (9) and (ii) nonsymmetric ones corresponding to the Fold catastrophe (8) (Fig. 6). Equation (1) for a rod can be given by ([33])

$$\frac{d^2u}{ds^2} + P \sin u = 0 \quad (23)$$

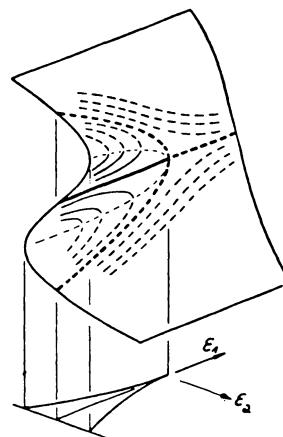


Fig. 1: Cusp Catastrophe (9).



Fig. 2: Simply supported rod.

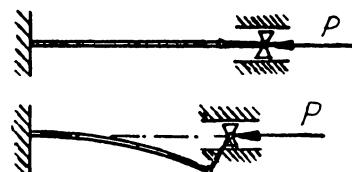


Fig. 3: Rod system with critical post buckling behavior.

with boundary conditions (Fig. 2)

$$u'(0) = u'(1) = 0. \quad (24)$$

(24) mean that we have simply supported ends of the rod. Application of the Liapunov Schmidt method yields the one dimensional bifurcation equation

$$x_1^3 + \lambda x_1 + \text{h.o.t.} = 0, \quad (25)$$

because we have a simple zero eigenvalue and u can be represented by

$$u = x_1 \cos \pi s + U(x_1, P), \quad (26)$$

where $\cos \pi s$ spans the kernel of the linearized operator obtained by taking the Frechet derivative of (23). λ in (25) is proportional to the loading measured from the critical value P_c . The full unfolding is given by the potential (9), which is related to (25). From Fig. 1 we can see, that there are only two different possibilities for a symmetric bifurcation point depending on the direction of variation of the loading parameter ϵ_1 . Either we have noncritical postbuckling behavior (negative direction of ϵ_1) as it is given by the common rod (Fig. 2) or we have a critical postbuckling behavior (positive direction of ϵ_1) as it could be given by the rod shown in Fig. 3 ([31]). The essential difference between these two cases is that in the case of the critical postbuckling behavior the bifurcation point is unstable and therefore we have a strong imperfection sensitivity whereas in the case of the noncritical postbuckling behavior the bifurcation point is stable and the system is not very sensitive against imperfections. Nonsymmetric bifurcation points are found for the loss of stability of frames ([13]) (Fig. 4) where the imperfection e as can be seen from Fig. 5 has a decisive influence on the qualitative behavior of the system. This behavior is given by the Fold Catastrophe (8) (Fig. 6).

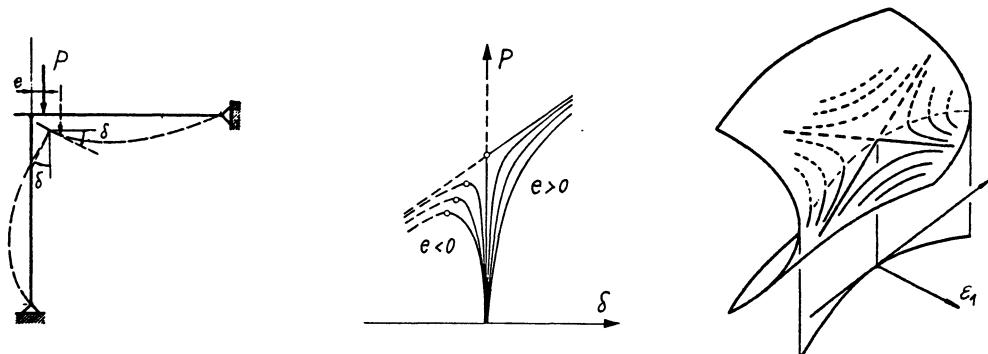


Fig. 4: Simple frame, e gives imperfect positions of the loading.

Fig. 5: P - δ diagram corresponding to Fig. 4.

Fig. 6: Fold-Catastrophe.

Among engineers there is sometimes some confusion about the influence of different imperfections and consequently about the question which is the

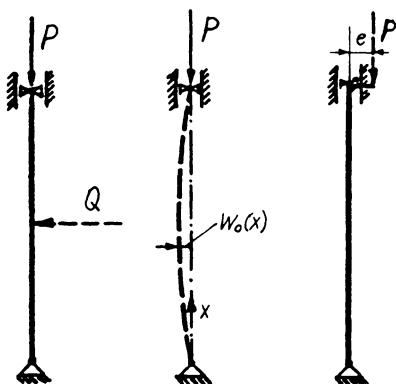


Fig. 7: Physically different imperfections Q , w_0 and e taken care of by ϵ_2 in (9).

most damaging one. It seems therefore to be appropriate to comment on this here in connection with (25). As (9) is the universally unfolded potential related with (25) all qualitatively different cases must be included in (9) and therefore all examples of different imperfections shown in Fig. 7 and also possible others are described by the parameter ϵ_2 in (9) for small variations of the parameters and variables in the neighborhood of the bifurcation point for the given class of buckling problems. To know this is a psychologically important fact because due to the universal unfolding one can be sure not to have overseen some, perhaps, important physical quantities.

b) Double eigenvalues.

They are found in a number of nongeneric situations. For example in buckling problems of arches ([23]), elastically supported rods ([10]) and plates. We want to explain their occurrence for plate problems. Let us consider a rectangular plate loaded by a thrust p and which is simply supported at its edges (Fig. 8). Then for certain critical values of the ratio of the length l of the plate to its width 1 double eigenvalues occur (Fig. 9) ([24]). Another example is an annular plate under uniform compression ([18]). Again at special

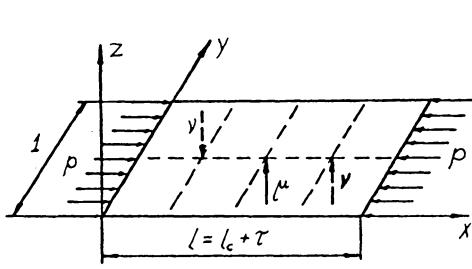


Fig. 8: Simply supported plate, close to critical length $l_c = \sqrt{2}$ for a double eigenvalue.

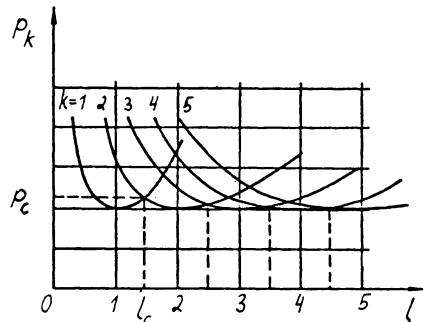


Fig. 9: Eigenvalue curves showing double eigenvalues for $l = l_c$

values of the ratio of the inner radius to the outer radius double eigenvalues occur. However these are pretty nongeneric cases and the question arises whether it makes sense to study the complicated case of bifurcation at double eigenvalues if those can be avoided mathematically by a small perturbation. The answer, nevertheless, is definitively yes, because of, at least, the

following two reasons. Firstly, even if the system is not precisely at a double eigenvalue but close to one then its behavior will be strongly affected by this closeness by the occurrence of secondary bifurcations ([5]). These secondary bifurcations change the postbuckling behavior and also the imperfection sensitivity considerably and therefore the theoretical investigation of the bifurcation problem in the neighborhood of a double eigenvalue is inevitable. A well known example is the phenomenon of mode jumping of plates ([5]) which basically depends on the effect of secondary bifurcations. Secondly there also exist problems where double eigenvalues occur inevitably by optimisation processes. A nice example is given in [32] where the buckling of a stiffened plate is investigated (Fig. 10). The two modes of failure are Euler buckling and local plate buckling. For a minimum weight design the critical values of the load will be identical for both types of failure. This yields a structure strongly sensitive against imperfections which can be handled mathematically by a hyperbolic umbilic ([11]) ([24,32]).

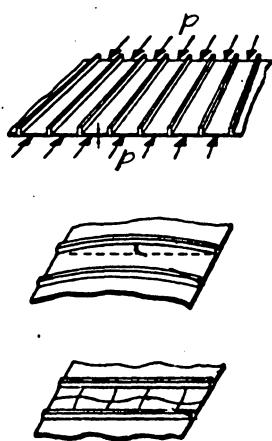


Fig. 10: Stiffened plate and different buckling failure modes.

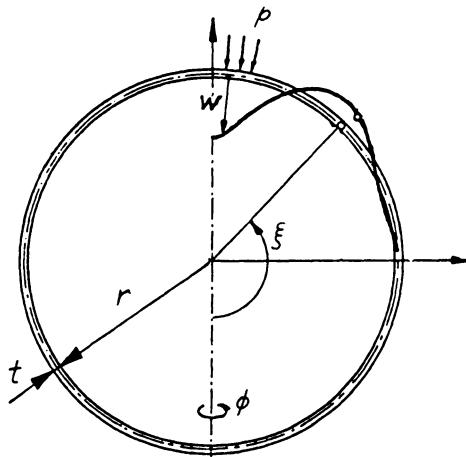


Fig. 11: Thin walled spherical shell.

Another practically important case for the occurrence of double eigenvalues is given in the buckling of thin axisymmetric spherical shells (Fig. 11) ([27]). From Fig. 12 we see that generically, as it was also found for the plate (Fig. 9), simple eigenvalues occur, but that there are certain special values of the ratio t/r for which double eigenvalues are found. The analysis of such a case ([27]) leads to a two dimensional system of bifurcation equations, which are related to a hyperbolic umbilic catastrophe (Fig. 13). In [27] the transformation to Normal Form and the physical interpretation of the universal unfolding are given. From Fig. 12 we can also see another phenomenon, which is important for shell structures, namely that for such

ratios of t/r , which are meaningful for flexible thin walled shells closely spaced eigenvalues are obtained. For these a classical bifurcation analysis has only a limited range of applicability because of the impractically small domain of admissible parameter variations. A better way to handle such problems is given by a singular perturbation approach ([14]).

All those cases mentioned so far could be traced back to one of the potentials of Elementary Catastrophe Theory. However, we mentioned before, that already for the case of a rectangular plate at a double eigenvalue the corresponding bifurcation equations correspond to a singularity of codimension eight and are therefore quite complicated concerning their unfolding ([24]). In such a case the theory of Restricted Generic Bifurcations is practically of more importance. The one parameter bifurcation equations which can be obtained from the von Karman plate equations ([24]) by means of the Liapunov Schmidt method are of the following form (3) (see also [30])

$$\begin{aligned} x_1^3 + \alpha x_1 x_2^2 - \lambda x_1 + \text{h.o.t.} &= 0 \\ x_2^3 + \beta x_1^2 x_2 - \lambda x_2 + \text{h.o.t.} &= 0. \end{aligned} \tag{27}$$

They are the analogous set of equations for the plate at a double eigenvalue as is (25) for the rod or a plate at a simple eigenvalue. α and β are constants depending on the boundary conditions ([5]). If one wants to study mode-jumping ([5]) the following parameters besides the thrust λ are of physical significance: the deviation τ of the length of the plate from the length l_c at which the double eigenvalue occurs and two types of transversal loads μ and ν which act on the two eigenfunctions corresponding to the double eigenvalue (Fig. 8). It is proved in [17] that the following unfolded set of bifurcation equations

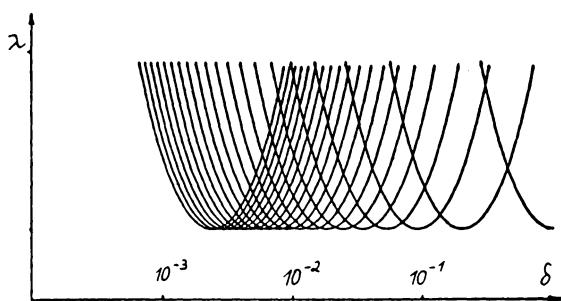


Fig. 12: Eigenvalue curves with double and closely spaced eigenvalues ($\delta \approx t/r$, $\lambda \approx$ load).

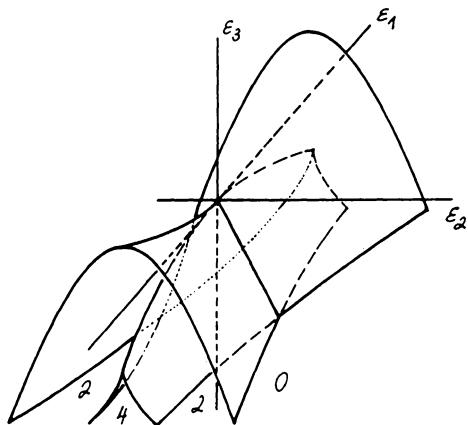


Fig. 13: Bifurcation diagram of the Hyperbolic Umbilic Catastrophe (11).

$$\begin{aligned} x_1^3 + \alpha x_1 x_2^2 - \lambda x_1 + \mu + \text{h.o.t.} &= 0 \\ x_2^3 + \beta x_1^2 x_2 - \lambda x_2 + \nu + \text{h.o.t.} &= 0 \end{aligned} \quad (28)$$

gives Restricted Generic Bifurcations, i.e. that at bifurcation points the number of solutions changes by two. In [30] it is also shown that for (28) only limit points occur. This shows that we are in a generic situation, because small perturbations of the system do not lead to qualitative changes in its behavior.

c) Multiple eigenvalues.

They mostly occur in shell buckling problems, where due to great symmetries the multiplicity can be very high. As a representative example for such a case and the related problems we shall consider buckling of a complete spherical shell under external pressure (Fig. 11), but without any restriction to axisymmetric deformations. In [12] a derivation of the governing equations is given which for the case of a spherical shell can be reduced to a set of two nonlinear partial differential equations in two variables f and w , being a stress and a displacement quantity respectively. By elimination of the stress-function f one can reduce this set of equations to one functional equation (1)

$$G(w, \lambda) = 0 \quad (29)$$

in the displacement variable $w(\xi, \phi)$ only ([12]). Calculating the Frechet derivative G_w at the perfect spherical but contracted sphere and solving the corresponding linear eigenvalue problem, one finds that the kernel of G_w is spanned by spherical harmonics, which allow a representation of $w(\xi, \phi)$ in the form

$$w(\xi, \phi) = \sum_{i=0}^m (x_i \cos i\phi + y_i \sin i\phi) P_m^i(\cos \xi) \quad (30)$$

where the P_m^i are Legendre polynomials and the x_i and y_i will be the variables in the bifurcation equations. From (30) one can see that there exist

$$k = 2m + 1 \quad (31)$$

different buckling modes. m depends on the ratio t/r and is given by the following relationship

$$m(m+1) = (12(1 - v_e^2))^{\frac{1}{2}} (r/t) \quad (32)$$

where v_e is a material constant namely the Poisson number ($0 \leq v_e \leq 0.5$). For flexible thin walled shells the order of magnitude of k calculated from (32) and (31) is in the range of 40 - 200. This means that after application of the Liapunov Schmidt method one obtains a system of k nonlinear algebraic equations in k unknowns. A further reduction of the order is therefore essential. We only want to indicate how one could proceed in this direction. The major tool is group theory ([12,26]). Unfortunately the application of group theory is not straight forward, because it is necessary to know the symmetry proper-

ties of the buckling pattern emerging on the shell before group theory can be used. Such information about the symmetry properties can be taken from experiments. As group action quite naturally the group of orthogonal transformations of R^3 $O(3)$ will be used. If T_g is a representation of $O(3)$ one says that (29) is invariant under the group action if ([12,26])

$$G(T_g w, \lambda) = T_g G(w, \lambda). \quad (33)$$

Having given a certain symmetry of the buckling pattern one now is able to select those modes from (30) which are invariant under the group action. This results in a strong reduction of the order of the problem. The buckling problem still remains quite complicated as the order of the reduced system, for example for hexagonal symmetry, will be about 5 to 10 for thin walled shells.

Analogous properties are found in a physically quite different problem, namely the convection problem (Benard problem) in a spherical fluid shell heated at the inner boundary. However in one respect this problem is simpler than the shell buckling problem, as also thick fluid shells are physically meaningful. For thick fluid shells the dimension of the kernel is small and then the application of group theory and Elementary Catastrophe Theory is possible [2,33]. For the ratio $t/r \rightarrow 0$ one obtains a similar result as it is given by (32) namely that one has an infinite dimensional kernel, which means that we have the well known degeneracy of the plane Benard problem ([33]).

B) Systems modelled by differential equations.

We give now some applications of the classified cases of chapter 3.

a) Simple eigenvalues.

We restrict to some problems in vehicle dynamics.

a) Tractor-semitrailer dynamics ([36]).

For the stability analysis of the steady state straight and cornering motion of such a vehicle (Fig. 14) a mechanical model consisting of rigid bodies which are coupled by hinges, springs and dampers with a finite number of degrees of freedom can be used. The equations of motion can be transformed into a set of n ordinary, nonlinear differential equations, which can be written

$$\dot{x} = H(x, V) \quad (34)$$

where $x, H \in R^n$ and $V \in R^1$ is the speed. $x \equiv 0$ is a solution of (34). Calculating

$$A = \frac{\partial H_i}{\partial x_j}(0) \quad (35)$$

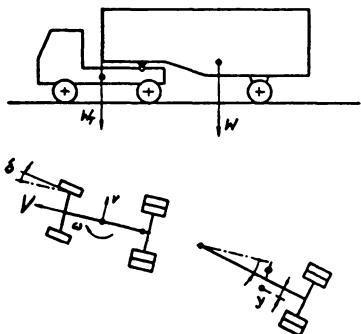


Fig. 14: Mechanical model of a tractor-semitrailer.

we find that for increasing V quasistatically from zero the system for low speeds behaves asymptotically stable ([15]) and reaches a critical state of behavior at $V = V_c$ when a single eigenvalue of A becomes zero. We obtain therefore a single bifurcation equation ([33, 36])

$$\dot{\xi} = h(\xi) \quad (36)$$

with $\xi \in \mathbb{R}^1$ being a linear transformation of the old variables x and $h(\xi)$ is a polynomial the lowest order term of which is cubic. The universal unfolding is given by (18).

However as already mentioned in chapter 3 a physically better way of unfolding is given by the theory developed in [4]. We shortly want to indicate this in connection with the tractor semi-trailer. As distinguished parameter we have the speed of the vehicle. However there are two other parameters namely the steering angle δ and an eccentric loading y (Fig. 14). As the singular system to be unfolded now

$$\dot{\xi} = \xi^3 + \lambda_1 \xi \quad (37)$$

is considered, where

$$\lambda_1 = f_1(V). \quad (38)$$

The full unfolding of (37) is ([4])

$$\dot{\xi} = \xi^3 + \lambda_3 \xi^2 + \lambda_1 \xi + \lambda_2 \quad (39)$$

with

$$\lambda_2 = r - s, \quad \lambda_3 = f_2(s), \quad r = f_3(\delta), \quad s = f_4(y). \quad (40)$$

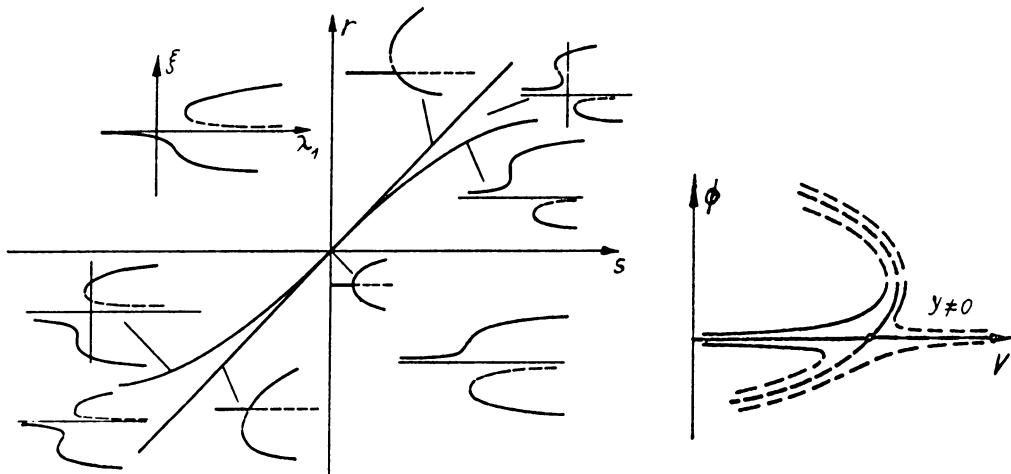


Fig. 15: Bifurcation diagram for (39) giving an unfolding of (37).

Fig. 16: Solutions of a non-symmetrically loaded trailer.

(39) easily could be transformed to the Normal Form (18), but this would lead to a mixing of variables with parameters and also of the distinguished parameter V with the imperfection parameters δ and y . The bifurcation diagram for (39) is given in Fig. 15. The physical interpretation is quite easy. For $y = 0$ we obtain an unstable symmetric bifurcation point (Fig. 1). The system therefore is sensitive to small perturbations which for example could result from steering actions. For the nonsymmetrically loaded system we obtain the bifurcation solutions of Fig. 16, which of course also can be obtained from Fig. 1 by sections with planes not parallel to the coordinate axis.

8) Hunting motion of a railway bogie ([20]).

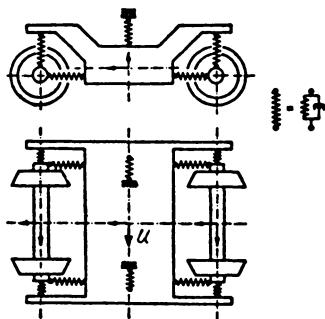


Fig. 17: Mechanical model of a railway bogie.

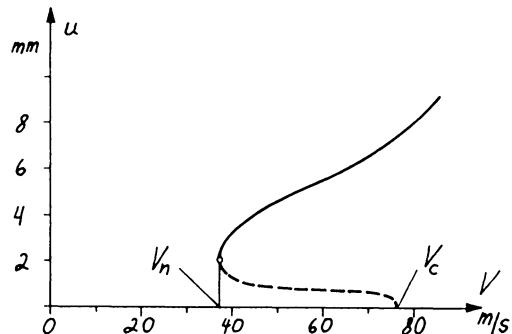


Fig. 18: Amplitudes of the unstable (dotted line) and stable limit cycles of the bogie.

Similar to the tractor semitrailer a mechanical model consisting of rigid bodies (Fig. 17) can be used to derive the equations of motion. The simplest system which is considered to reflect all the properties necessary for a practically meaningful stability analysis has eight degrees of freedom. Again we obtain a set of equations (34) with the speed V as distinguished parameter. At the critical speed V_c , however, we obtain now a purely imaginary pair of eigenvalues which gives rise to a Hopf-bifurcation. The bifurcation equation is after the introduction of polar coordinates one dimensional and of codimension one. We have

$$\dot{r} = r^3 + \lambda_1 r \quad (41)$$

with the notation of (38). For the practical behavior of the system it is important that at the bifurcation point the system is unstable and we have a critical post-bifurcation behavior which is of such an unpleasant form (Fig. 18) that the value of V_c is practically meaningless as already small perturbations of the system at speeds V with $V_n < V < V_c$ lead to hard self excitations and the system goes into a stable oscillatory motion. Hence V_n is the practi-

cally important critical speed which can be obtained only by a nonlinear stability analysis. After loss of stability of the steady state (trivial) solution, which is a flutter instability the system runs into a stable limit cycle. The nonlinear stability analysis of this limit cycle can be done in an analoguos way as we did it for the steady state solution if one introduces a Poincarè mapping and studies the stability of the fixed point of the Poincarè mapping ([1]).

y) Recovery of an aircraft from flat spin motion ([19]).

For high angle of attack flight conditions flat spin motions can occur. Here a basic question is whether it is possible by controlling only the rudder to bring the plane back into a trim positon. In [19] a bifurcation analysis is given which shows that only limit points or Hopf bifurcations occur and that it is crucial to the asked question whether there exists a limit point in the relationship between the spin motion and the rudder angle. If not the pilot will have to use parachutes or rockets launched at the tips of the wings in order to regain control of the aircraft, as it was the case for certain fighter-planes.

b) Multiple eigenvalues.

We want to give examples of self excited systems due to flowing fluids as applications of the more complicated Codimension Two cases listed in chapter 3. All of them are pretty nongeneric cases, but as already mentioned even the closeness to one of these exceptional situations makes the study of these complicated cases necessary, as nonlinear coupling between the different modes of instability leads to a very complicated behavior. Before we shortly discuss these cases we want to mention that the classical example of a failure due to an excitation by flowing fluid is the Tacoma bridge desaster of 1940 where oscillations resulting from a Hopf bifurcation led to the complete destruction of the bridge.

a) Oscillations of pipes conveying fluid ([22,29]).

Besides the every days experience with flutter instabilities occurring for the free end of a hose used to water a lawn there are also problems of significant technical relevance like oscillations and buckling instabilities of pipelines. The first problems of importance in this respect occurred with the Trans Arabian pipeline in 1950 ([22]). A simple discrete mechanical model for a pipe conveying fluid is given by a double pendulum with nonconservative follower force loading P and an elastic end support with stiffness c (Fig. 19). P and c are the two essential parameters. If c is zero or very small it is clear that we get flutter instability because we have a purely imaginary pair of roots at criticality. On the other hand if the spring is very stiff we obtain divergence bifurcation (a zero root) at the critical value of P . Therefore it is intuitively clear that there must be a critical value c_c in between

(Fig. 20) for which coupling between flutter and divergence instability occurs. In the bifurcation diagram of Fig. 21 one can see how small perturbations of the system around the critical point in the parameter plane influence the behavior of the system. As the bifurcation system obtained in this case is three-dimensional even chaotic motions can occur as it is shown in [29].

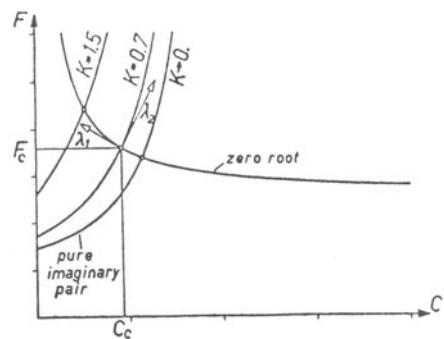


Fig. 20: Eigenvalue curves showing a double eigenvalue.

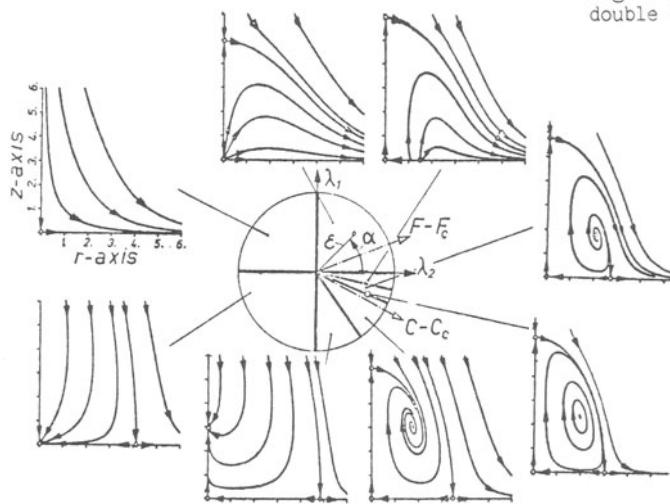


Fig. 21: Bifurcation diagram of (21) with notations of Fig. 20.

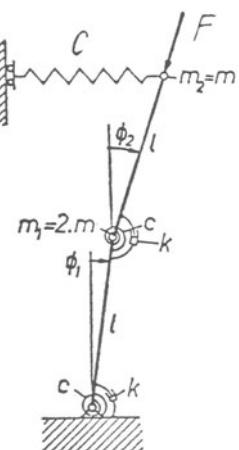


Fig. 19: Simple mechanical model of a pipe conveying fluid.

b) Oscillations of pipes in heat exchangers.

Such oscillations are highly unwelcome. They can occur for example due to galloping instabilities. As it is possible that some pipes in a row can be under tension and some other pipes in the same row under compression (Fig. 22) due to an inhomogeneous temperature distribution one can

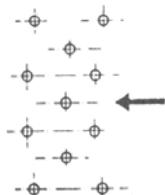
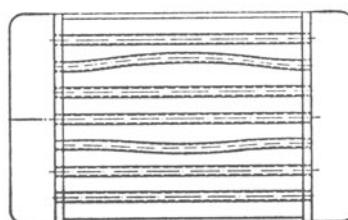


Fig. 22: Part of a heat exchanger.

get a system with a double zero eigenvalue. A simple mechanical model for a pipe is given in Fig. 23 from which the equation of motion

$$\ddot{x} + \epsilon_1 \dot{x} + \epsilon_2 x + x^3 + \dot{x}^3 = 0 \quad (42)$$

easily can be derived. The corresponding bifurcation diagram is given in Fig. 24. (42) also could be obtained from Center Manifold theory applied to an infinite dimensional model of the pipe.

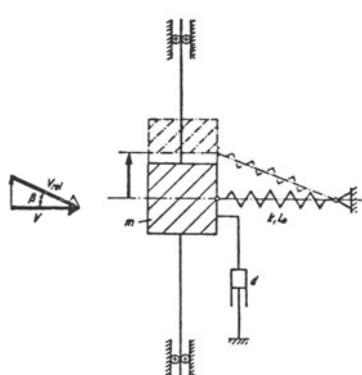


Fig. 23: Simple mechanical model of one pipe in Fig. 22.

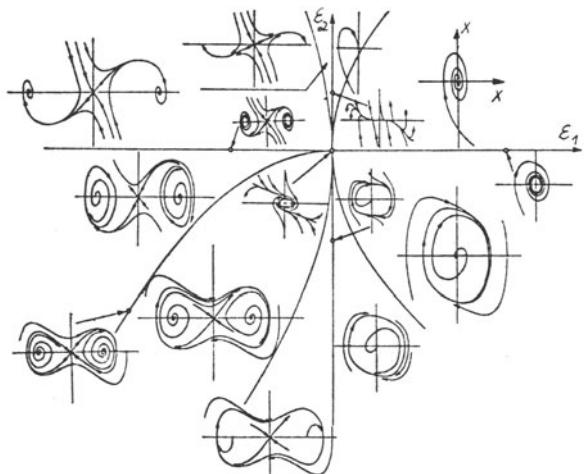


Fig. 24: Bifurcation diagram for (42) and Fig. 23.

Finally we shortly discuss a case where two purely imaginary pairs can occur. We take an oscillating pipe and couple it by a spring with an elastic wall which is often the case in heat exchangers. Fig. 25 gives a simple mechanical model, where the index 1 refers to the pipe and the index 2 to the wall. K is the stiffness of the spring coupling the two oscillators and F is the exciting force resulting from the flowing fluid. Qualitatively it is now, similar to the system of Fig. 19, quite easy to explain how two pairs of purely imaginary eigenvalues can occur. As parameters we take F which is proportional to the velocity of the fluid and c_2 , the stiffness of the wall. Again we have easily understandable extremal situations. A very stiff almost rigid wall ($c_2 \approx \infty$) and on the other hand a very soft wall ($c_2 \approx 0$). In either case Hopf bifurcations occur with different frequencies ω_1 and ω_2 for the resulting oscillation. Of course there exists now a certain critical value of the stiffness c_2 for which these two purely imagi-

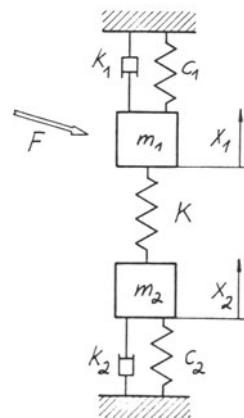


Fig. 25: Simple mechanical model for (22).

nary roots occur simultaneously. The bifurcation system is now four dimensional. By introduction of polar coordinates and by ignoring the azimuthal variables one ends up with a two dimensional bifurcation system like (22). However it has to be checked whether the case of strong resonance ($n\omega_1 - m\omega_2 = 0$, $|n| + |m| \leq 4$) which occurs in our case ($n=m=1$) is included in the versal deformation or not.

5. Hamiltonian Systems ([16]).

We only make few remarks on this class of systems, the stability theory of which is quite complicated because they do not possess asymptotically stable attracting solutions. For integrable Hamiltonian systems one obtains a set of tori in phase space covered by trajectories which are in the generic case densely distributed (quasi-periodic solution) whereas in the non generic case one also can have periodic motions. An important question now is: what happens to this picture if the system is slightly perturbed. For systems with at most two degrees of freedom the answer will be given by KAM (Kolmogorov, Arnold, Moser) theory ([16]). It roughly says that if the perturbation is small enough the phase space can be divided into two regions of non vanishing volume, one of which is small compared to the other and shrinks to zero as the perturbation strength goes to zero. The larger region has the familiar structure of embedded tori covered in the generic case with dense trajectories. But there is still the small region (i.e. a certain set of initial conditions) for which wildly erratic trajectories are found. This region is not of measure zero and is the domain of instability.

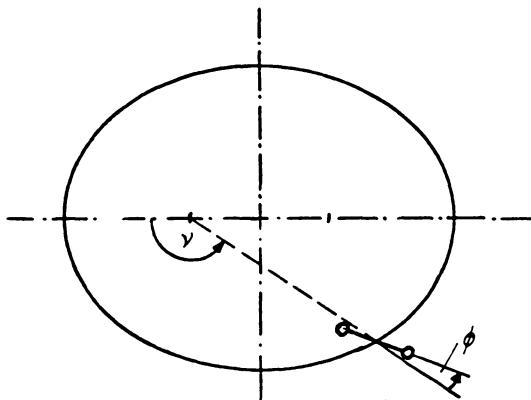


Fig. 26: Dumbbell satellite on an elliptic orbit.

As an example we shortly discuss the motion of a dumbbell satellite on an elliptic orbit ([37]) (Fig. 26). This concerns the practically important problem of the passive orientation of telecommunication satellites. It basically is the question which initial conditions must be given to the system such that the satellite has a prescribed orientation in certain points of its orbit. The equation of motion is

$$(1+e \cos \nu) \frac{d^2 \phi}{d\nu^2} - 2e \sin \nu \frac{d\phi}{d\nu} + a \sin \phi \cos \phi = 2e \sin \nu. \quad (43)$$

α is a constant describing the mass distribution of the satellite and will be kept at a fixed value. e is the eccentricity of the ellipse and is the parameter to be varied. For $e = 0$ we have an integrable Hamiltonian system. For small e we can see from Fig. 27 ([25]) where a cross section of the tori in three space is shown that there is a big

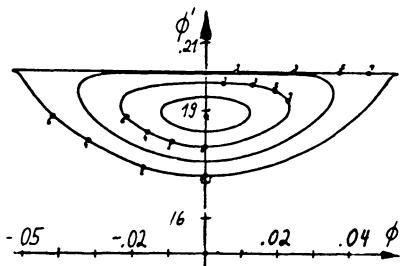


Fig. 28: As Fig. 27 for $e = 0.35$. Notice the different scale compared to Fig. 27 !

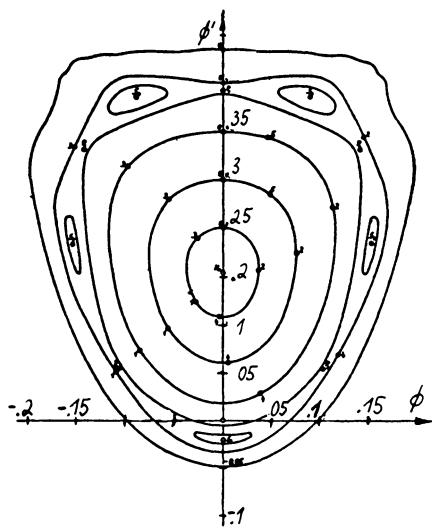


Fig. 27: Cross-section through the tori for $e = 0.3$.

region in the phase space from which regular solutions start. Especially there are certain values of initial conditions for which we obtain solutions of period 5. If we increase e further we see from 28 that the domain of initial conditions for which we have regular solutions shrinks considerably.

Another class of practically important problems where the stability theory of Hamiltonian systems is of relevance is the dynamics of elementary particles in particle accelerators ([21]).

It must however be pointed out that for systems with more than two degrees of freedom a new phenomenon, namely Arnold diffusion can occur which eventually leads to instability of all trajectories ([16]).

References

- 1 Arnold V.I., Geometrical Methods in the Theory of Ordinary Differential Equations, Springer-Verlag, New York, Heidelberg, Berlin 1982
- 2 Busse F.H., Patterns of convection in spherical shells, J.Fluid Mech. 72 (1975) 67 - 85.
- 3 Carr J., Application of Center Manifold Theory, Appl.Math.Sciences 35, Springer-Verlag, New York, Heidelberg, Berlin 1981
- 4 Golubitsky M., D.Schaeffer, A Theory for Imperfect Bifurcation via Singularity Theory, Comm. Pure Appl. Math. 32 (1979) 21 - 98.

- 5 Golubitsky M., D.Schaeffer, Boundary Conditions and Mode Jumping in the Buckling of a Rectangular Plate, Comm. Math. Phys. 69 (1979) 209 - 236.
- 6 Hale J.K., Restricted generic bifurcation, in: Nonlinear Analysis, Academic Press, New York 1978, p. 83 - 98.
- 7 Holmes P.J., J.E.Marsden, Dynamical Systems and Invariant Manifolds, in P.J.Holmes (ed.), New Approaches to Nonlinear Problems in Dynamics, SIAM, Philadelphia 1980, p. 3 - 26.
- 8 Holmes P.J., Center Manifolds, Normal Forms and Bifurcations of Vector Fields with Application to Coupling between Periodic and Steady Motions, Physica 2D (1981) 449 - 481.
- 9 Horozov E.I., Bifurcations of a vector field near a singular point in the case of two pairs of imaginary eigenvalues, I and II, Comptes Rendues de l'Academie Bulgare des Sciences 34 (1981) 1221 - 1224 and 35 (1982) 149 - 152.
- 10 Hui D., J.S.Hansen, Two Mode Buckling of an Elastically Supported Plate and its Relation to Catastrophe Theory, J.Appl.Mech. 47 (1980) 607 - 612.
- 11 Kirchgässner K., Bifurcation in Nonlinear Hydrodynamic Stability, SIAM Review 17 (1975) 652 - 683.
- 12 Knightly G.H., D.Sather, Buckled States of a Spherical Shell under Uniform External Pressure, Arch.Rat.Mech.Anal. 72 (1980) 315 - 380.
- 13 Koiter W.T., Post-Buckling Analysis of a Simple Two-Bar Frame, in:(Broberg, Hult, Niordson eds.) Recent Progress in Applied Mechanics, Almqvist and Wiksell, Stockholm 1967, p.337 - 354.
- 14 Lange Ch.G., G.A.Kriegsmann, The Axisymmetric Branching Behavior of Complete Spherical Shells,Quart.Appl.Math. 39 (1981) 145 - 178.
- 15 La Salle J., S.Lefschetz, Stability by Liapunov's Direct Method with Applications, Academic Press, New York 1961.
- 16 Lichtenberg A.J., M.A.Liebermann, Regular and Stochastic Motion, Applied Math.Sciences 38, Springer-Verlag, Berlin, Heidelberg, New York 1983.
- 17 List S.E., Generic Bifurcation with Application to the von Karman Equations, J.Diff.Equ. 30 (1978) 89 - 118.
- 18 Majumdar S., Buckling of a Thin Annular Plate under Uniform Compression, AIAA-J. 9 (1971) 1701 - 1707.
- 19 Mehra R.K., J.V.Carroll, Bifurcation Analysis of Aircraft High Angle of Attack Flight Dynamics, in: P.J.Holmes (ed.), New Approaches to Nonlinear Problems in Dynamics, SIAM, Philadelphia 1980, p.127 - 146.
- 20 Moelle D., R.Gasch, Nonlinear Bogie Hunting, in: A.H.Wickens (ed.), Proc. 7th IAVSD-Symposium in Cambridge UK 1981, Swets and Zeitlinger B.V., Lisse 1982, 455 -467.
- 21 Moser J., Is the Solar System Stable?, The Mathematical Intelligencer 1 (1978) 65 - 71.
- 22 Paidoussis M.P., N.T.Issid, Dynamic Stability of Pipes Conveying Fluid, J. Sound Vibr. 33 (1974) 267 - 294.

- 23 Plaut R.H., Buckling of Shallow Elastic Structures, in P.J.Holmes (ed.), New Aproaches to Nonlinear Problems in Dynamics, SIAM, Philadelphia 1980, p. 361 - 376.
- 24 Poston T., I.Stewart, Catastrophe Theory and its Applications, Pitman, London 1978.
- 25 Reindorf W., Ebene Schwingungen eines Satelliten auf einer elliptischen Umlaufbahn, Diplomarbeit, TU-Wien 1982.
- 26 Sattinger D.H., Bifurcation and Symmetry Breaking in Applied Mechanics, Bull. Amer. Math. Soc. 3,2 (1980) 779 - 819.
- 27 Scheidl R., H.Troger, Buckling and Postbuckling of Complete Spherical Shells, Proc. Euromech. Coll. on Flexible Shells, Munich 1983, Springer-Verlag to appear.
- 28 Scheidl R., H.Troger, Verzweigungsverhalten eines nichtlinearen Schwingers mit einem doppelten Eigenwert Null, ZAMM 62 (1982) T 72 - T 74.
- 29 Scheidl R., H.Troger, K.Zeman, Coupled Flutter and Divergence Bifurcation of a Double Pendulum, Int.J.Non-linear Mech. in print.
- 30 Steinrück H., H.Troger, R.Weiss, Mode Jumping of Imperfect Buckled Rectangular Plates, this volume.
- 31 Stern J., Der Gelenkstab bei großen elastischen Verformungen, Ing.Archiv 48 (1979) 173 - 184.
- 32 Thompson J.M.T., G.W.Hunt, Towards a Unified Bifurcation Theory, ZAMP 26 (1975) 581 - 603.
- 33 Troger H., Verzweigungstheorie - eine Herausforderung für Mathematiker und Ingenieure, GAMM-Nachrichten 2 (1982) 47 - 82.
- 34 Troger H., K.Zeman, Application of bifurcation diagrams to the modelling of stability problems, X.Avula (ed.), Proceedings of the IV.Int.Conf. Math.Modelling, Zürich 1983, Pergamon Press 1983.
- 35 Zeeman E.C., Catastrophe Theory, Scientific American April 1976, 65 - 83.
- 36 Zeman K., H.Troger, R.Scheidl, Bifurcation Theory and Vehicle Dynamics - With Application to the Tractor Semitrailer, in: A.H.Wickens (ed.), Proc. 7th IAVSD - Symposium in Cambridge UK 1981, Swets and Zeitlinger B.V., Lisse 1982, 97 - 110.
- 37 Zlatoustov V.A., D.E.Okhotsimsky, V.A.Sarychev, A.P.Torzhevsky, Investigation of satellite oscillations in the plane of an elliptic orbit, Proc.IUTAM-Conf. Munich 1964 (H.Görtner ed.) Springer-Verlag, Berlin, Heidelberg 1965, p. 436 - 439.

Hans Troger, Institut für Mechanik, Karlsplatz 13, A-1040 Wien

A SINGULAR MULTI-GRID ITERATION METHOD FOR BIFURCATION PROBLEMS

Helmut Weber

We propose an efficient technique for the numerical computation of bifurcating branches of solutions of large sparse systems of nonlinear, parameter-dependent equations. The algorithm consists of a nested iteration procedure employing a multi-grid method for singular problems. The basic iteration scheme is related to the Lyapounov-Schmidt method and is widely used for proving the existence of bifurcating solutions. We present numerical examples which confirm the efficiency of the algorithm.

1. Introduction

In this paper we analyze an efficient algorithm for the iterative solution of large, sparse systems of nonlinear, parameter-dependent equations which arise from discretizations of nonlinear elliptic eigenvalue problems of the form

$$(1) \quad Lu = f(\lambda, u) \quad \text{in } \Omega, \quad Bu = 0 \quad \text{on } \partial\Omega$$

Here $\Omega \subset \mathbb{R}^n$ is a bounded domain and L a linear elliptic differential operator, B a homogeneous boundary operator, $B0 = 0$. $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ is a smooth nonlinearity satisfying $f(\lambda, 0) = 0$. We are interested in the fast computation of nontrivial solution branches of (1) that bifurcate from the trivial solution branch at certain points $\lambda_0 \in \mathbb{R}$, cf. [5], [10].

Finite-dimensional approximations to (1) of the form

$$(2) \quad L_h u_h - f_h(\lambda, u_h) = 0$$

have been analyzed under different assumptions, e.g. by Beyn [1], Brezzi et al. [3], Kikuchi [11] and Weiss [22]. On the other hand there has been some progress in developing fast algorithms for solving the finite-dimensional problems (2) in the context of continuation and bifurcation, see Chan and Saad [4], Mackens and Jarausch [14], Mittelmann and Weber [15] or Weber [19].

Our work is based on an iterative algorithm proposed by Keller and Langford [10] and Demoulin and Chen [6] which consists in fact of a preliminary transformation of the problem and a certain simplified Newton iteration. It has been used often for proving the existence of bifurcating branches and

has also been investigated in the context of numerical computations, see e.g. Kikuchi [11], Rheinboldt [16], Weber [18], Weber and Werner [21] and Weiss [22].

Before beginning to describe the algorithm we want to point out some goals important for the design of fast and efficient algorithms in the present case:

1. Optimality: linear or almost linear growth of the computational effort with the number N of discrete unknowns.
2. Storage demands proportional to N .
3. No difficulties with the parametrization of the solution branch.
4. Wide applicability with respect to nonlinearities and domains.
5. Easy implementation.
6. Theoretically well founded convergence.

This list is of course not very complete but nevertheless helpful.

We come back to a short description of the algorithm. A certain multi-grid method is used for solving the linear, (nearly) singular problems that arise in every iteration step of the basic method. The class of multi-grid methods that we use here is based on work by Brandt [2] and Hackbusch [7,8]. We want to mention some very desirable properties of multi-grid methods: for certain linear elliptic operators on an n by n grid the multi-grid method is able to compute an approximate solution to truncation error in $O(n^2)$ arithmetic operations, using $O(n^2)$ storage places. Combining this approach with the nested iteration technique [7,17] we obtain a very fast and secure algorithm for computing bifurcating solutions of (1) which is not subject to limitations with respect to the parametrization - for pathological examples see [5], or with respect to the stability of the solutions involved.

2. The Basic Iterative Algorithm

We consider the finite-dimensional bifurcation problem

$$(3) \quad Lu - f(\lambda, u) = 0$$

where $L \in \mathbb{R}^{n,n}$, $f: \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ smooth, $f(\lambda, 0) = 0$ and $L_0 := L - f_u(\lambda_0, 0)$ is symmetric. Assume that

$$(4) \quad N(L_0) = \text{span}\{\phi\}, \quad 0 \neq \phi \in \mathbb{R}^n, \quad \langle \phi, f_{u\lambda}^0 \phi \rangle = a_0 \neq 0,$$

$\langle \cdot, \cdot \rangle$ denoting the Euclidean inner product in \mathbb{R}^n , $f_{u\lambda}^0 := f_{u\lambda}^0(\lambda_0, 0)$. Setting $Z := \{x \in \mathbb{R}^n \mid \langle \phi, x \rangle = 0\}$ we make the following Ansatz for the nontrivial bifurcating solutions

$$(5) \quad u(\varepsilon) = \varepsilon\phi + \varepsilon^2 v(\varepsilon), \quad v(\varepsilon) \in Z, \quad \lambda(\varepsilon) = \lambda_0 + \varepsilon\gamma(\varepsilon),$$

ε being a real parameter. The method of Keller and Langford [10] and Demoulin and Chen [6] has the form

$$(6) \quad \begin{aligned} & v^0 := 0, \quad \gamma^0 := 0, \quad u^0 := \varepsilon\phi, \quad \lambda^0 := \lambda_0; \\ & \text{for } i := 0, 1, 2, \dots \text{ do} \\ & \quad \left\{ \begin{array}{l} \gamma^{i+1} := \gamma^i - (\varepsilon^2 a_0)^{-1} \langle \phi, f(\lambda^i, u^i) - f_u^0 u^i \rangle \\ \text{solve} \\ L_0 v^{i+1} = \frac{1}{\varepsilon^2} \{ f(\lambda^i, u^i) - f_u^0 u^i \} + f_{u\lambda}^0 \phi(\gamma^{i+1} - \gamma^i), \quad v^{i+1} \in Z, \\ u^{i+1} := \varepsilon\phi + \varepsilon^2 v^{i+1}, \quad \lambda^{i+1} := \lambda_0 + \varepsilon\gamma^{i+1} \end{array} \right. \end{aligned}$$

Theorem 1 Let (4) be valid for equation (3). Then, for $|\varepsilon| < \varepsilon_0$, ε_0 sufficiently small, the iteration process (6) is convergent. The limit $(\bar{\gamma}(\varepsilon), \bar{v}(\varepsilon))$ of the sequence (γ^i, v^i) forms a smooth curve in $\mathbb{R} \times Z$ for $|\varepsilon| < \varepsilon_0$. The unique bifurcating family of nontrivial solutions of (3) near $(\lambda_0, 0)$ is given by $(\bar{\lambda}(\varepsilon), \bar{u}(\varepsilon)) = (\lambda_0 + \varepsilon\bar{\gamma}(\varepsilon), \varepsilon\phi + \varepsilon^2 \bar{v}(\varepsilon))$.

The proof consists mainly of an application of Banach's fixed point theorem, cf. [10], [6].

In the case of a multiple eigenvalue zero of the linearization L_0 a similar, somewhat more involved iteration scheme exists, see e.g. [6], [10] or [21]. There are also various modifications of the above scheme for similar applications and situations, e.g. for perturbed bifurcation, see [9], [21].

The characteristic feature of all of these methods is the presence of a singular operator or matrix which has to be inverted. Several techniques have been proposed therefore, see [4], [11], [18], [21]. These depend considerably on the class of problems to be treated. The approach to be presented here is especially well suited for large, sparse systems of equations arising from finite element or finite difference discretizations of semi-linear elliptic equations of the form (1).

The particular implementation of (6) by a singular multi-grid method which is discussed in the next paragraph is applied to the discrete

version (2) of the original problem (1). For simplicity we shall assume that the differential operator consisting of L and B in (1) is symmetric, furthermore we assume that problem (1) exhibits simple bifurcation in the sense of Crandall and Rabinowitz [5] at $(\lambda_0, 0)$. It is necessary to inspect the approximating problems (2), $L_h u_h - f_h(\lambda, u_h) = 0$, somewhat more in detail. We have $L_h \in \mathcal{L}(X_h, Y_h)$, $f_h : \mathbb{R} \times X_h \rightarrow Y_h$ smooth and $f_h(\lambda, 0) = 0$ for all λ . X_h and Y_h are finite-dimensional normed vector spaces with $\dim X_h = \dim Y_h$, $\|\cdot\|_h$ denotes the norm in X_h . h is a real discretization parameter, tending to zero. $\Delta_h \in \mathcal{L}(X_h, X_h)$ is the usual matching operator, $\lim \|\Delta_h u\|_h = \|u\|, u \in X$, where X is the space, in which the problem (1) is posed, e.g. $C_0^{2,\alpha}(\Omega)$ or $H_0^1(\Omega)$ for second order problems with zero boundary conditions.

Under appropriate consistency and stability conditions it has been shown for certain classes of problems and certain discretizations that the discrete problems (2) also exhibit simple bifurcation from the trivial solution at points $(\lambda_{0h}, 0) \in \mathbb{R} \times X_h$, $\lambda_{0h} \rightarrow \lambda_0$ for $h \rightarrow 0$, cf. e.g. [1], [3], [11], [22]. The main tool for proving this is the iteration scheme (6). Particularly, if k is the order of consistency of the discretization method, usually the following error estimates are valid:

$$(7) \quad |\lambda_0 - \lambda_{0h}| \leq c_1 h^k, \quad \|\Delta_h \phi - \phi_h\|_h \leq c_2 h^k, \quad \|\phi_h\|_h = \|\phi\| = 1,$$

$$|\bar{\gamma}(\epsilon) - \bar{\gamma}_h(\epsilon)| \leq c_3 h^k, \quad \|\Delta_h \bar{v}(\epsilon) - \bar{v}_h(\epsilon)\|_h \leq c_4 h^k, \quad |\epsilon| < \epsilon_1 \leq \epsilon_0$$

for the discrete bifurcating solutions $\bar{u}_h(\epsilon) = \epsilon(\phi_h + \epsilon \bar{v}_h(\epsilon))$, $\bar{\lambda}_h(\epsilon) = \lambda_{0h} + \epsilon \bar{\gamma}_h(\epsilon)$. ϕ (ϕ_h) is the linearized (discrete) eigenfunction. The c_i do not depend on h and ϵ . Thus we have

$$(8) \quad \|\Delta_h \bar{u}(\epsilon) - \bar{u}_h(\epsilon)\|_h \leq |\epsilon|(c_2 + |\epsilon|c_4)h^k$$

$$|\bar{\lambda}(\epsilon) - \bar{\lambda}_h(\epsilon)| \leq (c_1 + |\epsilon|c_3)h^k$$

for $|\epsilon| < \epsilon_1$, $(\bar{\lambda}(\epsilon), \bar{u}(\epsilon))$ being the nontrivial branch of solutions of the original problem (1).

3. A Multi-Grid Method for Singular Problems

Consider the discrete problems (2) for a sequence

$$h_0 > h_1 > h_2 > \dots > h_{e-1} > h_e > \dots \quad (e: \text{level number})$$

of stepstizes. For simplicity we assume $h_\ell = 2^{-\ell} h_0$ ($\ell \in \mathbb{N}$) and we write (2) in the form

$$(9) \quad L_\ell u_\ell - f_\ell(\lambda, u_\ell) = 0,$$

setting $x_\ell = x_{h_\ell} = y_{h_\ell}$. $\| \cdot \|_\ell$ denotes the norm in X_ℓ . The levels ℓ and $\ell-1$ are connected by the restriction $r_\ell: X_\ell \rightarrow X_{\ell-1}$ and by the prolongation $p_\ell: X_{\ell-1} \rightarrow X_\ell$.

The characteristic feature of the multi-grid method for solving large linear systems is the combination of smoothing steps and coarse grid corrections. During the smoothing steps - usually Gauss-Seidel or Jacobi iterations - the defect is reduced only slightly but smoothed. By the following correction step the discrete solution is improved by solving an auxiliary equation on a coarser grid. In fact this equation has to be of the same structure and sparsity pattern and is solved again by an application of the algorithm on a lower level. This leads to the recursive structure of the multi-grid algorithm which can be well described by an ALGOL-like program. Introductions to multi-grid methods are [2] or [17].

Having in mind the iteration scheme (6) we consider here a multi-grid method for solving the linear singular problem

$$(10) \quad L_\ell^0 u_\ell = f_\ell$$

where L_ℓ^0 is a symmetric matrix having a simple eigenvalue zero. By $\langle \cdot, \cdot \rangle_\ell$ we denote the inner product on level ℓ . The eigenfunction of L_ℓ^0 corresponding to the eigenvalue zero is denoted by $\phi_\ell: L_\ell^0 \phi_\ell = (L_\ell^0)^T \phi_\ell = 0$. Problem (10) is solvable if and only if $\langle \phi_\ell, f_\ell \rangle_\ell = 0$. Define the orthogonal projection onto the range of L_ℓ^0 by $Q_\ell x_\ell = x_\ell - \langle \phi_\ell, x_\ell \rangle_\ell \phi_\ell / \langle \phi_\ell, \phi_\ell \rangle_\ell$.

We describe the multi-grid process for solving (10) by the following ALGOL-like program which performs one step of this multi-grid iteration.

```

PROCEDURE mgsing(l,u,f);
  INTEGER l; ARRAY u,f;
  COMMENT l: actual level number, u: input: u^(i), output: u^(i+1),
        f: right hand side of equation (10) to be solved;
  IF l=0 THEN u:=(L_0^0)^T*f ELSE
  BEGIN INTEGER j; ARRAY v,d;
  FOR j:=1 STEP 1 UNTIL v DO u:=G_l(u,f);

```

```

d:=Qℓ-1*rℓ*(Lℓ0*u - f); v:=0;
FOR j:=1 STEP 1 UNTIL γ DO mgsing(ℓ-1,v,d);
u:=u - pℓ*v;
FOR j:=1 STEP 1 UNTIL μ DO u:=Gℓ(u,f);
u:=Qℓ*u;
END mgsing;

```

In the above program G_ℓ symbolizes the smoothing procedure $u_\ell \rightarrow G_\ell(u_\ell, f_\ell)$, which usually consists of a step of the Gauss-Seidel, Jacobi or SOR iteration, cf. [2]. Other choices are possible, too. γ and μ are the numbers of such smoothing steps before and after the coarse grid correction step which is performed by γ calls of mgsing on level $\ell-1$.

The convergence of the singular multi-grid method has been shown under reasonable assumptions in [8].

There is a useful simplification of mgsing without projections Q_ℓ . Under the conditions

$$(11) \quad p_\ell \phi_{\ell-1} = s \phi_\ell, \quad r_\ell \phi_\ell = s' \phi_{\ell-1} \quad (\ell \geq 1, s, s' \in \mathbb{R})$$

also this multi-grid iteration - let us call it mgsing1 - converges provided the equation (10) is solvable and the original method mgsing is convergent. The convergence is modulo the kernel of L_ℓ^0 . mgsing1 has considerable advantages with respect to the computational effort. We have used it in practice without encountering deviations from the performance of mgsing.

For theoretical purposes a strategy with fixed γ , μ and γ (for example $\gamma=\mu=1$, $\gamma=2$) as in the above ALGOL-program is useful. For the design of efficient programs an adaptive strategy might be sometimes preferable. An adaptive strategy which was used alternatively in our program is discussed in [2].

In our applications the matrix L_ℓ^0 is frequently of the form $L_\ell - \lambda_\ell I_\ell$ or $L_\ell - \lambda_\ell C_\ell$ with some sparse matrix C_ℓ . It is shown in [8] that the eigenvalues λ_ℓ may be perturbed by $O(h_\ell^2)$ for second order problems and that also the projections need not to be exact. So it is possible to use results of foregoing numerical calculations.

The above procedure is easily generalized to (almost) singular problems with multiple eigenvalues zero, cf. [8].

4. The Nested Multi-Grid Bifurcation Iteration

We want to compute a bifurcating branch of solutions of (9) on the highest level, say $\ell = m$. Let ϕ_ϵ be the linearized discrete eigenfunction on level ℓ defined by $(L_\ell - a_\ell)\phi_\ell = 0$, $\|\phi_\ell\|_\ell = 1$, where $a_\ell = \frac{\partial^2}{\partial u_\ell} f_\ell(\lambda_{0\ell}, 0)$. The $\lambda_{0\ell}$ are the discrete bifurcation points. A multi-grid version of the iteration scheme (6) in the case of a symmetric matrix $L_\ell^0 := L_\ell - a_\ell$, which we consider here only, has the form

$$(12) \quad \begin{cases} \epsilon \neq 0 \text{ given; } v_\ell^0 := 0, \gamma^0 := 0, u_\ell^0 := \epsilon \phi_\ell, \lambda^0 := \lambda_{0\ell}; \\ a_{0\ell} := \langle \phi_\ell, \frac{\partial^2}{\partial u_\ell} f_\ell(\lambda_{0\ell}, 0) \phi_\ell \rangle; \\ \text{for } i := 0, 1, 2, \dots \text{ do} \\ \left\{ \begin{array}{l} b_\ell^i := \frac{1}{\epsilon^2} [f_\ell(\lambda^i, u_\ell^i) - a_\ell u_\ell^i]; \\ \gamma^{i+1} := \gamma^i - a_{0\ell}^{-1} \langle \phi_\ell, b_\ell^i \rangle; \quad \tilde{b}_\ell^i := Q_\ell b_\ell^i; \\ \text{solve the equation } L_\ell^0 v_\ell^{i+1} = \tilde{b}_\ell^i, v_\ell^{i+1} \in Q_\ell X_\ell \\ (\text{by } s \text{ iterations of mgsing or mgsing1 on level } \ell); \\ u_\ell^{i+1} := \epsilon(\phi_\ell + \epsilon v_\ell^{i+1}), \lambda^{i+1} := \lambda_{0\ell} + \epsilon \gamma^{i+1} \end{array} \right. \end{cases}$$

The complement of $N(L_\ell^0)$ is here $Q_\ell X_\ell = \{x_\ell \in X_\ell \mid \langle \phi_\ell, x_\ell \rangle_\ell = 0\}$, where $Q_\ell x_\ell = x_\ell - \langle x_\ell, \phi_\ell \rangle_\ell \phi_\ell / \langle \phi_\ell, \phi_\ell \rangle_\ell$.

This algorithm, applied on level m , works well if the discrete problem on level m exhibits simple bifurcation and if the multi-grid method for singular problems is convergent. However, for relatively large values of ℓ , for which the iterative algorithm (6) still converges, the computational effort may become prohibitively large, even if one linear, singular problem is solved very fast by mgsing or mgsing1.

Thus we make use of the idea of the nested iteration (cf. [2,17]) that makes it possible to need only a small number of calls of mgsing or mgsing1 on the highest - and most expensive - level $\ell = m$.

We introduce an interpolation operator

$$(13) \quad q_\ell: X_{\ell-1} \longrightarrow X_\ell$$

which may be different from the prolongation p and of higher order, cf. [2], [17]. Furthermore we introduce a mapping

$$(14) \quad \bar{q}_\ell: R \longrightarrow R$$

which transforms the result γ^* of the iteration (12) on level $\ell-1$ into an

initial approximation for γ on level ℓ . A trivial choice is $\bar{q}_\ell = \text{id}$. The nested iteration then has the form

$$(15) \quad \begin{aligned} & \ell_0 > 0 \text{ fixed, } \epsilon \neq 0 \text{ given;} \\ & a_{0\ell} := \left\langle \phi_\ell, \frac{\partial^2}{\partial x^2} f_\ell(\lambda_{0\ell}, 0) \phi_\ell \right\rangle, \ell = \ell_0, \dots, m; \\ & \text{for } \ell := \ell_0, \ell_0 + 1, \dots, m \text{ do:} \\ & \left\{ \begin{array}{l} v_\ell^0 := \begin{cases} 0, & \ell = \ell_0 \\ q_\ell v_{\ell-1}^*, & \ell > \ell_0 \end{cases}; \quad \gamma_\ell^0 := \begin{cases} 0, & \ell = \ell_0 \\ \bar{q}_\ell(\gamma_{\ell-1}^*, \gamma_{\ell-2}^*), & \ell > \ell_0 \end{cases}; \\ u_\ell^0 := \epsilon(\phi_\ell + \epsilon v_\ell^0), \quad \lambda_\ell^0 := \lambda_{0\ell} + \epsilon \gamma_\ell^0; \\ \text{for } i := 0, 1, 2, \dots \text{ do:} \\ \left\{ \begin{array}{l} b_\ell^i := \frac{1}{\epsilon^2} [f_\ell(\lambda_\ell^i, u_\ell^i) - a_{0\ell} u_\ell^i]; \\ \gamma_\ell^{i+1} := \gamma_\ell^i - a_{0\ell}^{-1} \langle \phi_\ell, b_\ell^i \rangle; \quad \tilde{b}_\ell^i := Q_\ell b_\ell^i; \\ \text{solve equation } L_\ell^0 v_\ell^{i+1} = \tilde{b}_\ell^i, \quad v_\ell^{i+1} \in Q_\ell X_\ell \text{ (by}} \\ \text{s iteration steps of mgsing or mgsing1 on level } \ell); \\ u_\ell^{i+1} := \epsilon(\phi_\ell + \epsilon v_\ell^{i+1}), \quad \lambda_\ell^{i+1} := \lambda_{0\ell} + \epsilon \gamma_\ell^{i+1}; \\ \text{result of this iteration: } v_\ell^*, \gamma_\ell^*; \\ u_m^* := \epsilon(\phi_m + \epsilon v_m^*), \quad \lambda_m^* := \lambda_{0m} + \epsilon \gamma_m^* \text{ (final results).} \end{array} \right. \end{array} \right. \end{aligned}$$

By S_ℓ we denote the mapping $(\gamma_\ell^{i+1}, v_\ell^{i+1}) = S_\ell(\epsilon, (\gamma_\ell^i, v_\ell^i))$, defined by the iteration (6) on level ℓ . For the convergence of our algorithm we have the following result.

Theorem 2 Assume that

- (i) the continuous problem (1) exhibits simple bifurcation in the sense of [5] at $(\lambda_0, 0)$;
- (ii) the discrete problems (2) approximating (1) exhibit simple bifurcation at $(\lambda_{0h}, 0)$, the estimates (7) and (8) being valid for $|\epsilon| < \epsilon_1 \leq \epsilon_0$ and all $h_\ell, \ell_0 \leq \ell \leq m$;
- (iii) the mapping S_ℓ is contracting on $V_\ell^\ell := \{(\gamma_\ell, v_\ell) \mid v_\ell \in Q_\ell X_\ell, |\gamma_\ell| + \|v_\ell\| \leq \varsigma\}$ for $|\epsilon| < \epsilon_1, \ell_0 \leq \ell \leq m$, and some $\varsigma > 0$;
- (iv) $|\bar{q}_\ell \bar{\gamma}_{\ell-1}(\epsilon) - \bar{\gamma}_\ell(\epsilon)| = o(1)$ for $h \rightarrow 0$, where $\bar{\lambda}_\ell(\epsilon) = \lambda_{0\ell} + \epsilon \bar{\gamma}_\ell(\epsilon)$, $\bar{\lambda}_\ell$ exact solution of (9) on level ℓ ;
- (v) for all $u \in X$ we have $\|q_\ell \Delta_{\ell-1} u - \Delta_\ell u\| \leq d_1 \|u\| h_\ell^\alpha, \alpha > 0, \ell_0 + 1 \leq \ell \leq m$, d_1 independent of ℓ ;

- (vi) $\|q_\epsilon u\|_e \leq d_2 \|u\|_{e-1}$ for all $u \in X_{e-1}$, $e_0 + 1 \leq e \leq m$, d_2 independent of ϵ , u ;
 (vii) $h = 2^{-\ell} h_0$, $\ell = 1, 2, \dots, m$.

Then, for $|\epsilon|$ and h_m sufficiently small, the nested iteration process (15) with exact solution of $L_e^0 v_e^{i+1} = b_e^i$ converges on all levels $\ell = \ell_0, \dots, m$ and we have $\lim_{i \rightarrow \infty} u_m^i(\epsilon) = \bar{u}_m(\epsilon)$, $\lim_{i \rightarrow \infty} \lambda^i(\epsilon) = \bar{\lambda}_m(\epsilon)$.

Proof. Since the mapping S_ϵ is contracting, Banach's fixed point theorem implies the convergence for $|\epsilon|$ sufficiently small, see [6], [10], [21].

At level ℓ_0 the iteration converges to $(\bar{\gamma}_\epsilon^0(\epsilon), \bar{v}_\epsilon^0(\epsilon))$ for $|\epsilon| < \epsilon_1$ by (i)-(iii).

Due to Ostrowski's theorem the iteration process converges at all levels

$\ell = \ell_0 + 1, \dots, m$ towards $(\bar{\gamma}_\epsilon(\epsilon), \bar{v}_\epsilon(\epsilon))$ for $|\epsilon| < \epsilon_1$ provided the starting values $(\gamma_\epsilon^0(\epsilon), v_\epsilon^0(\epsilon))$ are sufficiently good. We have

$$\begin{aligned} \|v_e^0(\epsilon) - \bar{v}_e(\epsilon)\|_e &= \|q_\epsilon \bar{v}_{e-1}(\epsilon) - \bar{v}_e(\epsilon)\|_e \|q_\epsilon (\bar{v}_{e-1}(\epsilon) - \Delta_{e-1} \bar{v}(\epsilon))\|_e \\ &+ \|q_\epsilon \Delta_{e-1} \bar{v}(\epsilon) - \Delta \bar{v}(\epsilon)\|_e + \|\Delta \bar{v}(\epsilon) - \bar{v}_e(\epsilon)\|_e \leq d_2 \|\bar{v}_{e-1}(\epsilon) - \Delta_{e-1} \bar{v}(\epsilon)\|_{e-1} \\ &+ d_1 \|\bar{v}(\epsilon)\| h_e^\alpha + C_4 h^k \leq C_4 (1 + 4d_2) h^k + d_1 \|\bar{v}(\epsilon)\| h_e^\alpha, \text{ due to hypotheses (v),} \end{aligned}$$

(vi) and (vii). Furthermore (iv) implies

$$|\gamma_\epsilon^0(\epsilon) - \bar{\gamma}_\epsilon(\epsilon)| = |\bar{q}_\epsilon \bar{\gamma}_{e-1}(\epsilon) - \bar{\gamma}_\epsilon(\epsilon)| \rightarrow 0 \text{ for } h \rightarrow 0.$$

Thus, for $|\epsilon|$ and h_m sufficiently small, the assertion follows inductively.

Remarks

- The assumptions (ii) and (iii) of Theorem 2 are related since the existence of $\bar{u}(\epsilon)$ and $\bar{\lambda}(\epsilon)$ is usually proved by means of (6), see e.g. [22]. So it is natural to impose (iii).
- By (7) hypothesis (iv) is fulfilled for $\bar{q}_\epsilon = \text{id}$. However, if \bar{q}_ϵ has an asymptotic expansion with respect to h , it is possible to obtain better choices of \bar{q}_ϵ by extrapolation..
- If equation $L_e^0 v_e^{i+1} = \tilde{b}_e^i$ is solved by s steps of the multi-grid iteration using (mgsing1) standard perturbation results valid for contracting mappings imply that (15) will yield iteration sequences slightly perturbed if compared with the above discussed. The distances between the corresponding iterates can be made arbitrarily small by choosing s sufficiently large. However, note that it is not useful to compute the discrete solutions more accurately than to the level of the discretization error. In practice $s=1$ or $s=2$ turned out to be enough.
- Let n be the number of space dimensions and N_ϵ denote the number of grid

points (discrete unknowns) at level ℓ . Then our algorithm needs $3N_m/(1-2^{-n})$ storage units. This is 1.5 times the number of storage units needed for the multi-grid solution of Poisson's equation. If the ϕ_ℓ are explicitly known we obviously have the same storage demand as for solving Poisson's equation.

5. The algorithm (15) may be combined with an ε -continuation on level ℓ_0 .
6. For non-symmetric matrices L_ℓ^0 the algorithm has to be modified accordingly by introducing the projections $Q_\ell^* : X_\ell \rightarrow \text{span}\{\phi_\ell^*\}^\perp$, where ϕ_ℓ^* spans the nullspace of $(L_\ell^0)^T$. See also [20].
7. In algorithm (15) we have assumed that ϕ_ℓ and $\lambda_{0\ell}$ are known. For domains with simple geometries this is often true, but in general one has to compute these quantities numerically. There are different ways to do this in the context of multi-grid methods, too. In the present case it seems natural to use mgsing (mgsing1) also for this purpose. A technique for solving eigenvalue problems of the form $L_\ell \phi_\ell - K_\ell(\lambda) \phi_\ell = 0$ which is based on mgsing was analyzed by Hackbusch [8]. We used Hackbusch's method, that does not need to be described here, throughout. It should be pointed out, however, that the additional implementation of this eigenvalue algorithm into a computer program for (15) does not impose a great programming effort.
8. So far we dealt with discretizations of problem (1). A prototype of such problems is $-\Delta u = f(\lambda, u)$ in $\Omega \subset \mathbb{R}^2$, $u=0$ on $\partial\Omega$ with $f(\lambda, 0) = 0$. Immediate generalizations cover the case of $Lu = f(\lambda, x, u, Du, \dots)$, $Bu=0$, where L is a $2m$ -th order linear elliptic differential operator and f depends also on derivatives of u up to order $2m$, cf. [5]. It is also possible to treat von Kármán's equations for the buckling of a thin elastic plate. Another class of extensions is given by the problems of perturbed bifurcation, see [21]. In section 2 we mentioned bifurcation from multiple eigenvalues. If a multiple eigenvalue of the linearized continuous problem is preserved by the discretization, there will be no difficulties in generalizing (15) for such problems. But note that in general multiple eigenvalues will split up due to the influence of discretization. Here additional considerations are necessary. In [11] Kikuchi discussed an iteration scheme for computing solution branches through simple turning points. It is natural to extend our algorithm to this case. Another type

of problems which could be treated by a generalization of (15) has the form $Lu = \lambda f(u)$ in Ω , $u = -1$ on $\partial\Omega$, where $f(u)(x) = u(x)$ if $u(x) \geq 0$ and $f(u)(x) = 0$ if $u(x) < 0$, see Kikuchi [12]. Let us finally mention secondary bifurcation from a non-trivial solution in the case of simple eigenvalues. Rheinboldt's algorithm [16] closely related to (6) uses a singular chord iteration. It should be possible to extend (15) into this direction. However, there may be difficulties due to the influence of discretization which destroys true bifurcation in many cases, cf. [1], [3].

5. Numerical Results

Semilinear elliptic differential equations of the form (1) which exhibit bifurcation phenomena arise in many disciplines of the applied sciences, for example in the theory of elasticity, in the context of reaction-diffusion equations in chemistry and biology, in plasma physics and statistical mechanics, cf. e.g. Lions [13].

The class of problems we have solved numerically by (15) has the form

$$(16) \quad -\Delta u = f(\lambda, u) \quad \text{in } \Omega, \quad u=0 \text{ on } \partial\Omega,$$

where Ω is a rather general, bounded domain in \mathbb{R}^2 . One-dimensional problems of the form $-u'' = f(\lambda, u)$, $u(0) = u(1) = 0$, were treated, too. For results see [20].

The author has written three programs: one for solving the above mentioned one-dimensional problem, one for solving (16) on rectangular domains $\Omega \subset \mathbb{R}^2$ (BIFMG0) and one for solving (16) on a "general" domain (BIFMG1). BIFMG1 is able to solve also the linearized eigenvalue problem by Hackbusch's method [8]. It makes use of several subroutines of Stüben's program MG01, see [17] and of the organization of grids introduced there. All programs are written in portable FORTRAN 66. The discretization is the usual five-point star with Shortley-Weller boundary approximation in BIFMG1.

Using these routines the user has to specify only FORTRAN functions defining the boundaries of Ω and the right hand side $f(\lambda, u)$ as well as $f_u(\lambda, 0)$. An approximation of λ_0 is required, too. It is possible to create graphical output, including contour plots and tree-dimensional representations of the solutions. Some details of mgsing1 are: smoothing by (checkered) Gauss-Seidel relaxation, linear interpolation for prolongation, half injection for

restriction, $\gamma = 1$, $\psi = 1$ or 2 , $\mu = 1$ (or adaptive strategy [2]). In algorithm (15) we used quadratic (BIFMGO) or cubic (BIFMG1) interpolation for q , for \bar{q} we chose $\bar{q}_\ell = \gamma_{\ell-1}^*$ for $\ell = 1$ and $\bar{q}_\ell = \frac{5}{4}\gamma_{\ell-1}^* - \frac{1}{4}\gamma_{\ell-2}^*$ for $\ell \geq 2$, corresponding to $k = 2$. The number s of calls of mgsing1 was $s=1$ or $s=2$ throughout.

In the following tables we denote by it_m the number of iteration steps of (15) on level m . By time we denote the CPU time in seconds. All computations were performed on the Honeywell Bull 66/80 computer of the Rechenzentrum of the Johannes Gutenberg University at Mainz in single precision FORTRAN (27 bit mantissa).

a) Examples on rectangular domains

On $\Omega = (0,1)^2$ we consider $-\Delta u = \lambda(u(1 - \sin u) + u^3)$ with zero boundary conditions, cf. Lions [13]. For $m=6$, $\ell_0=1$, $h=h_6=1/128$, adaptive strategy we summarize some results for the first

Table 1

branch emanating at $\lambda_0 = 4\pi^2 \approx 19.739\dots$

with linearized eigenfunction

$\phi(x,y) = \sin \pi x \sin \pi y$. The discrete eigenvalue is $\lambda_0 = 19.738217$. The shape of the bifurcating curve is sketched in Fig. 1. It has a turning point at $\epsilon \approx 0.55$, $\lambda \approx 25.09$ which was passed without difficulties due to the parametrization with respect to ϵ .

ϵ	$\lambda(\epsilon)$	$u(\frac{1}{2}, \frac{1}{2})$	it_6	time
0.05	20.444869	0.0049827	1	16.5
0.1	21.137387	0.0993465	1	16.5
0.25	23.039527	0.2467528	1	16.3
0.5	24.991605	0.4947046	1	16.4
0.75	24.387155	0.7608215	1	20.9
1.0	21.027946	1.0638886	3	30.6
1.25	16.312800	1.4082507	4	39.1
1.5	12.018802	1.7746194	5	50.7

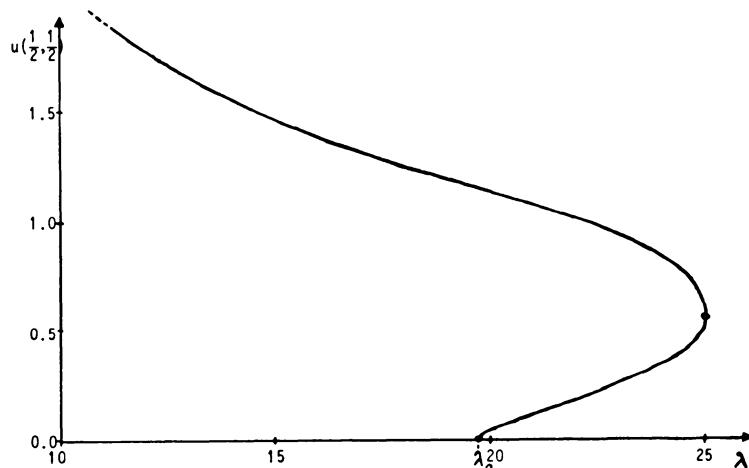


Fig. 1

As a second example of this class consider $-\Delta u = \lambda u - u^3$, $u=0$ on $\partial\Omega$ on the unit square Ω . We are interested in computing the unstable branch emanating at $\lambda_0 = 8\pi^2 = 78.956835\dots$, the third (simple) eigenvalue of the linearization with eigenfunction $\phi(x,y) = \sin 2\pi x \sin 2\pi y$. For $m=4$, $\ell_0=0$, $h=h_4=1/80$, $\lambda_{04}=78.916366$ some results

are given in Table 2. Note that for the computation of branches bifurcating at eigenvalues of higher order with linearized eigenfunctions having a more complicated nodal behavior the discretization on the coarsest level has to be fine enough to represent the solution properly.

Table 2

ϵ	$\lambda(\epsilon)$	$u(\frac{1}{4}, \frac{1}{4})$	it ₄	time
0.1	78.92199	0.099999	1	3.9
0.25	78.95152	0.249980	1	4.0
0.5	79.05695	0.499839	1	4.0
0.75	79.23255	0.749456	2	6.1
1.0	79.47818	0.998712	3	9.3
1.5	80.17853	1.495670	4	12.0

b) Examples on non-rectangular domains

Consider $-\Delta u = \lambda \sinh u$ on Ω , $u=0$ on $\partial\Omega$, where $\Omega = \{(x,y) \in \mathbb{R}^2 \mid \sqrt{x^2+y^2} < \frac{1}{2}\}$. We treated bifurcation from the first simple eigenvalue $\lambda_0 = 23.132745\dots$. In Table 3 we present some results for

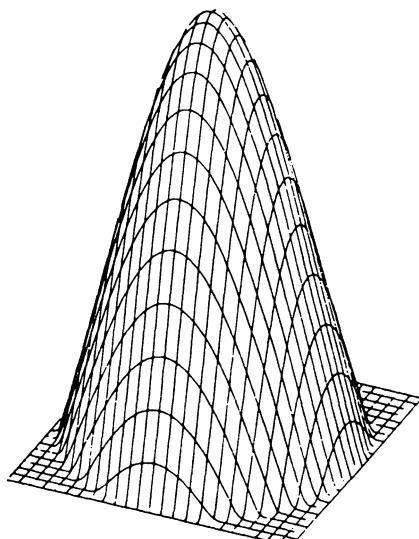


Fig. 2

Table 3

ϵ	$\lambda(\epsilon)$	$u(0,0)$	it ₅	time
0.1	23.10301	0.100016	1	2.9
0.25	22.98894	0.250257	1	3.3
0.5	22.58586	0.502065	2	4.9
0.75	21.92864	0.757000	3	6.6
1.0	21.03817	1.016704	4	9.1
1.5	18.67120	1.557423	6	12.5

$m=5$, $\ell_0=0$, $h=h_5=1/64$, $\lambda_{05}=23.124794$.

In Fig. 2 we present the shape of the bifurcating solution for $\epsilon = 1.5$. For more details we refer to [20].

Another example of this class, $-\Delta u = \lambda \sin u - u^4$ with zero boundary conditions on a "banana-shaped" domain $\Omega \subset \mathbb{R}^2$ can be found in [20], too.

In Table 4 we list some computing times for the first example in b) for different values of m . N_m is the number of

discrete unknowns. They confirm a nearly linear growth of the computational effort with N_m .

Finally let us make some remarks on the fulfillment of the conditions 1.-6. from the introduction. Points 1.-4. are easily seen to be fulfilled, due to our considerations and the numerical examples. The implementation of the algorithm (15), point 5, is "easy" if one possesses a multi-grid program for the corresponding linear problem $Lu = f$, $Bu = g$, e.g. the program in [2] or MG01, see [17]. The convergence, point 6, is well established, too.

Table 4

m	N_m	time $\epsilon = 0.5$	time $\epsilon = 1.0$
1	9	0.06	0.09
2	45	0.18	0.24
3	193	0.66	0.85
4	793	2.07	2.67
5	3205	6.64	9.06

6. References

1. W.-J. Beyn, On discretizations of bifurcation problems, in: Bifurcation problems and their numerical solution, H.D. Mittelmann, H. Weber (eds.), ISNM Vol.54, Birkhäuser-Verlag, Basel 1980.
2. A. Brandt, Multi-level adaptive solutions to boundary value problems, *Math. Comput.* 31(1977), 333-390.
3. F. Brezzi, J. Rappaz, P.A. Raviart, Finite dimensional approximation of nonlinear problems, part iii: Bifurcation points, *Numer. Math.* 38(1981), 1-30.
4. T.F. Chan, Y. Saad, Iterative methods for solving bordered systems with applications to continuations methods, preprint no. 235, Dept. Comp. Sci. Yale University, New Haven 1982.
5. M.G. Crandall, P.H. Rabinowitz, Bifurcation from simple eigenvalues, *J. Funct. Anal.* 8(1971), 321-340.
6. Y.-M.J. Demoulin, Y.M. Chen, An iteration method for solving nonlinear eigenvalue problems, *SIAM J. Appl. Math.* 28(1975), 588-595.
7. W. Hackbusch, On the convergence of multi-grid methods, *Beitr. Numer. Math.* 9(1981), 213-239.
8. W. Hackbusch, On the computation of approximate eigenvalues and eigenfunctions of elliptic operators by means of a multi-grid method, *SIAM J. Numer. Anal.* 16(1979), 201-215.
9. J.P. Keener, H.B. Keller, Perturbed bifurcation theory, *Arch. Rational Mech. Anal.* 50(1974), 159-174.
10. H.B. Keller, W.F. Langford, Iterations, perturbations and multiplicities for nonlinear bifurcation problems, *Arch. Rational Mech. Anal.* 48(1972), 83-108.
11. F. Kikuchi, Finite element approximations to bifurcation problems of turning point type, in: R. Glowinski, J.L. Lions (eds.), Computing methods in applied sciences and engineering, Lecture Notes in Math. Vol. 704, pp.243-266, Springer-Verlag, Berlin 1979.

12. F. Kikuchi, An iteration scheme for a nonlinear eigenvalue problem, *Theoretical and Applied Mechanics* 29(1981), 319-333, University of Tokyo Press.
13. P.L. Lions, On the existence of positive solutions of semilinear elliptic equations, *SIAM Rev.* 24(1982), 441-467.
14. W. Mackens, H. Jarausch, CSNP a fast, globally convergent scheme to compute stationary points of elliptic variational problems, Bericht Nr.15, Inst. f. Geom. u. Prakt. Math., RWTH Aachen, Aug. 1982.
15. H.D. Mittelmann, H. Weber, Multi-grid solution of bifurcation problems, manuscript, Dortmund 1983, submitted for publication.
16. W.C. Rheinboldt, Numerical methods for a class of finite dimensional bifurcation problems, *SIAM J. Numer. Anal.* 15(1978), 1-11.
17. K. Stüben, U. Trottenberg, Multigrid methods: fundamental algorithms, model problem analysis and applications, in: *Multigrid methods*, W. Hackbusch, U. Trottenberg (eds.), Lecture Notes in Math. Vol.960, pp. 1-176, Springer-Verlag, Berlin 1982.
18. H. Weber, Numerische Behandlung von Verzweigungsproblemen bei gewöhnlichen Differentialgleichungen, *Numer. Math.* 32(1979), 17-29.
19. H. Weber, An efficient technique for the computation of stable bifurcation branches, *SIAM J. Sci. Stat. Comput.* 1983, in press.
20. H. Weber, Multi-grid bifurcation iteration, Report. No.3(1983), Rechenzentrum der Johannes-Gutenberg Univ. Mainz, submitted for publication.
21. H. Weber, W. Werner, On the numerical solution of some finite-dimensional bifurcation problems, *Numer. Funct. Anal. Optimiz.* 3(1981), 341-366.
22. R. Weiss, Bifurcation in difference approximations to two-point boundary value problems, *Math. Comput.* 29(1975), 746-760.

Helmut Weber
Rechenzentrum der
Johannes Gutenberg-Universität
Postfach 3980
D-6500 Mainz 1

REGULAR SYSTEMS FOR BIFURCATION POINTS WITH UNDERLYING SYMMETRIES

Bodo Werner

For the computation of bifurcation points the regularity of determining systems is a desirable property. We will present regular systems for simple and double bifurcation points of one-parameter dependent problems $g(x, \lambda) = 0$ which satisfy certain symmetry-invariance conditions. Besides of the numerical relevance we will show that regular systems can be a rather powerful tool to obtain analytical bifurcation results.

1. INTRODUCTION

For the computation of bifurcation points (or more general of singular points) of

$$g(x, \lambda) = 0 \quad (g: X \times \mathbb{R} \rightarrow X, \quad X \text{ a Banachspace}) \quad (1.1)$$

various types of (*extended, inflated, augmented, defining, determining, ...*) systems of equations are used. See the papers of Beyn, Caluwaerts, Cliffe, Griewank, Jepson, Kubicek, Reddien, Roose, Seydel and Spence in these proceedings (Küpper/Mittelmann/Weber, Eds.) and Werner/Spence (83) for further references.

We will use the notation *regular system* if the singular point (x_0, λ_0) of (1.1) corresponds to an isolated solution (the Frechet derivative has a bounded inverse) of the system. The number of additional (unfolding) parameters in the regular system is related to the codimension of the singularity in the sense of Golubitsky/Schaeffer (79a). For example see the regular system for "nonsimple" (better: simple cubic) turning points in Spence/Werner (82) and the system for simple bifurcation points in Moore (80) where one unfolding parameter is involved.

We will mainly be concerned with the simple system

$$G(x, \lambda, \phi) := \begin{bmatrix} g(x, \lambda) \\ g_x(x, \lambda)\phi \\ 1\phi - 1 \end{bmatrix} = 0 \quad (G: X \times X \times \mathbb{R} \rightarrow X \times X \times \mathbb{R}) \quad (1.2)$$

where $l \in X^*$ is some scaling functional ($g_x := \partial g / \partial x$, $l\phi := \langle l, \phi \rangle$).

The use of (1.2) seems to be limited to simple quadratic turning points since (1.2) is regular only for those singular points.

In many numerical treatments of non-artificial bifurcation problems it has been noticed that bifurcation phenomena of non-turning point type are mostly due to symmetry- or other invariance conditions for $g(x, \lambda)$ (see Sec.2). Golubitsky/Schaeffer (79b) proved that the codimension of the singularity is reduced considerably in the presence of symmetry. Under this background it is not surprising that the restriction of x and ϕ in (1.2) to symmetric and anti-symmetric elements respectively, leads to a regular system for simple symmetry-breaking (pitchfork) bifurcation points (Werner/Spence (83)).

As a first slight extension we notice (Th. 3.1) that (1.2) is still a regular system if only x is restricted to be symmetric (or more general, to be an element of an invariant subspace X_0). This might explain why Seydel (79) succeeded numerically using (1.2) for the computation of bifurcation points which are not turning points.

In Sec.4 we investigate the regularity (and singularity) of certain systems of type (1.2) for double singular S-points (x_0, λ_0) of (1.1) where x_0 is symmetric and the kernel $N(g_x^0)$ is spanned by a symmetric and an anti-symmetric element (Th.4.1). For certain double bifurcation points it turns out (Th.4.4) that they correspond to simple bifurcation points of a corresponding extended system. This is related (Remark 4.5) to the hypothesis of Bauer/Keller/Reiss (75) "multiple eigenvalues lead to secondary bifurcation" which has been justified by Shearer (80).

This is one example that the use of determining systems for singular points is not limited to computational aspects. As a further example we show (Sec.3) that the regularity of certain systems can easily be used to obtain analytical bifurcation results of the type in Crandall/Rabinowitz (71).

Several symmetry invariances are due to geometrical symmetries of underlying domains of boundary value problems. We will give two examples which are not of this type where double singular S-points occur in a quite natural way (complexification of a real turning point problem and a 3-cell reaction problem). Other examples for symmetries in static bifurcation problems (1.1) can be found in these proceedings (Bohl, Raugel, Steinrück/Troger/Weiss).

2. SYMMETRY - INVARIANCE

The most common underlying symmetry for (1.1) is the \mathbb{Z}_2 -symmetry:

$$(I_S) \quad \begin{cases} \text{There exists a linear operator } S \in L(X) \text{ with } S^2 = I \text{ such that} \\ g(Sx, \lambda) = Sg(x, \lambda), \quad x \in X, \lambda \in \mathbb{R}. \end{cases}$$

The subspaces $X_S := \{x \in X : x = Sx\}$ and $X_A := \{x \in X : x = -Sx\}$ contain the *symmetric* and *anti-symmetric* elements respectively. A more general condition is

$$(I_\gamma) \quad \begin{cases} \text{There exists a finite group } \gamma \subset L(X) \text{ with} \\ g(Tx, \lambda) = Tg(x, \lambda), \quad T \in \gamma, \quad x \in X, \lambda \in \mathbb{R}. \end{cases}$$

Observe that (I_S) implies (I_γ) with $\gamma = \{I, S\}$. On the other hand γ might contain a "reflection" S such that (I_S) holds. The subspace

$$X_O := \{x \in X : x = Tx \text{ for each } T \in \gamma\}$$

contains the γ -symmetric elements. We have $X_O \subset X_S$ for each $T \in \gamma$ and $X_O = X_S$ if $\gamma = \{I, S\}$. It follows that X_O (and X_S) are invariant under $g(\cdot, \lambda)$. This leads to the following invariance condition:

$$(I_O) \quad \text{There exists a subspace } X_O \text{ of } X \text{ such that } g(x, \lambda) \in X_O \text{ for } x \in X_O, \lambda \in \mathbb{R}.$$

The best studied case is $X_O = \{0\}$.

Bifurcation problems with underlying symmetries have been investigated by Sattinger (79), Golubitsky/Schaeffer (79b), Shearer (80) and Raugel (82) (see also the references therein). We will be concerned with bifurcation points (x_O, λ_O) where $x_O \in X_S$ or $x_O \in X_O$. With respect to the system (1.2) we state the following

LEMMA 2.1

Let (I_S) and (I_O) be satisfied with $X_O \subset X_S$. Then the following subspaces of $X \times X \times \mathbb{R}$ are invariant under G (for the definition of G see (1.2)):

$$\left. \begin{array}{l} Y_{SA} := X_S \times X_A \times \mathbb{R}, \quad Y_{SS} := X_S \times X_S \times \mathbb{R}, \quad Y_S := X_S \times X \times \mathbb{R}, \quad Y_O := X_O \times X \times \mathbb{R} \\ Y_{OS} := X_O \times X_S \times \mathbb{R}, \quad Y_{OA} := X_O \times X_A \times \mathbb{R}, \quad Y_{OO} := X_O \times X_O \times \mathbb{R} \end{array} \right\} \quad (2.1)$$

Notation: By $G_{SA}, G_{SS}, G_S, G_O, G_{OS}, G_{OA}, G_{OO}$ we denote the restrictions of G to the subspaces in (2.1). The corresponding systems are $G_{SA} = 0$, $G_{SS} = 0$, etc. Observe that only for G_{OS} and G_{OA} (I_S) and (I_O) have to be satisfied. In the other cases it is sufficient to assume (I_S) or (I_O) .

3. SIMPLE S-BREAKING AND X_o -BREAKING BIFURCATION POINTS.

Let (x_o, λ_o) be a simple singular point of (1.1):

$$N(g_x^o) = \text{span } \{\phi_o\}, \quad R(g_x^o) = \{y \in X : \psi_o x = o\}, \quad \phi_o \in X \setminus \{o\}, \quad \psi_o \in X^* \setminus \{o\}.$$

A simple (quadratic) turning point is characterized by

$$\psi_o g_\lambda^o \neq o, \quad (\psi_o g_{xx}^o \phi_o \neq o), \quad (3.1)$$

a simple bifurcation point satisfies

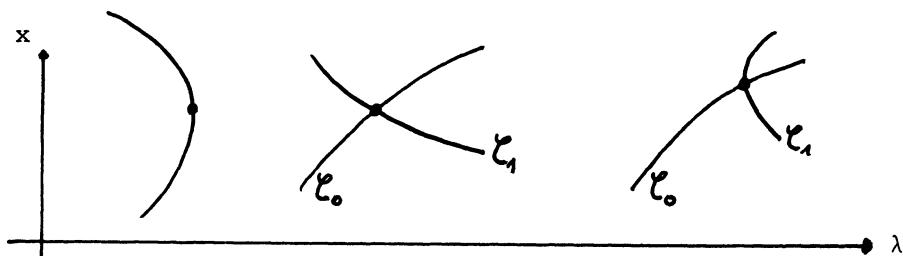
$$\psi_o g_\lambda^o = o, \quad B_o^2 - A_o C_o > o, \quad \text{where} \quad (3.2)$$

$$A_o := \psi_o g_{xx}^o \phi_o, \quad B_o := \psi_o (g_x^o + g_{xx}^o v_o) \phi_o, \quad (3.3)$$

$$C_o := \psi_o (g_{\lambda\lambda}^o + 2g_{x\lambda}^o v_o + g_{xx}^o v_o v_o), \quad g_\lambda^o + g_x^o v_o = o. \quad (3.4)$$

We call (x_o, λ_o) a simple X_o -breaking singular point if (I_o) holds, $x_o \in X_o$ and $\phi_o \notin X_o$. For those points it follows that $\psi_o x = o$ for $x \in X_o$ and $g_\lambda^o, g_{\lambda\lambda}^o, v_o, g_{x\lambda}^o v_o, g_{xx}^o v_o v_o \in X_o$. Hence $\psi_o g_\lambda^o = o$, $C_o = o$ and (x_o, λ_o) is a simple (X_o -breaking) bifurcation point iff $B_o \neq o$. A bifurcation analysis (see below) shows that there is a branch \mathcal{E}_o of solutions (x, λ) with $x \in X_o$ and a branch \mathcal{E}_1 intersecting \mathcal{E}_o in (x_o, λ_o) such that locally $\mathcal{E}_1 \cap (X_o \times \mathbb{R}) = \{(x_o, \lambda_o)\}$. That is the reason why we say X_o -breaking.

If $X_o = X_s$ and (I_s) holds, then $\phi_o \notin X_s$ implies that $\phi_o \in X_a$ and additionally $A_o = o$. Then \mathcal{E}_o is a branch of solutions (x, λ) with symmetric x , and (ϕ_o, o) is a tangent vector of \mathcal{E}_1 at the pitchfork bifurcation point (x_o, λ_o) - the direction of bifurcation is anti-symmetric. Instead of X_o -breaking we say S-breaking or symmetry-breaking. See figure 1.



Simple quadratic turning point.

Simple X_o -breaking bifurcation point.

Simple S-breaking bifurcation point.

Figure 1

The following theorem about the regularity of systems $G_o = 0$ and $G_{sa} = 0$ is a slight extension of Th.3.1 in Werner/Spence (83). No new ideas for the proof are needed.

THEOREM 3.1

Assume (I_o) or (I_s) . Let (x_o, λ_o) be a simple X_o -[or S-breaking singular point of (1.1)]. Then the system $G_o = 0$ [or the system $G_{sa} = 0$] is regular for (x_o, λ_o) if and only if (x_o, λ_o) is a simple X_o -[or S-breaking bifurcation point ($B_o \neq 0$ in (3.3))].

The regularity implies that Newton's method applied to $G_o = 0$ [or $G_{sa} = 0$] is locally quadratically convergent. Details about the implementation are in Werner/Spence (83). We will draw three further conclusions from the regularity of the systems $G_o = 0$ and $G_{sa} = 0$. These conclusions are based on the implicit function theorem applied to $(|\varepsilon| < \varepsilon_o)$

$$F_o(., \varepsilon) = 0, \text{ where } F_o(., \varepsilon) : Y_o \rightarrow Y_o \text{ and } F_o(., 0) = G_o,$$

or

$$F_{sa}(., \varepsilon) = 0, \text{ where } F_{sa}(., \varepsilon) : Y_{sa} \rightarrow Y_{sa} \text{ and } F_{sa}(., 0) = G_{sa} :$$

1. The perturbation of $g(x, \lambda) = 0$ by

$$f(x, \lambda, \varepsilon) = 0, \text{ where } f(., 0) = g$$

leads to

$$F(x, \phi, \lambda, \varepsilon) := (f(x, \lambda, \varepsilon), f_x(x, \lambda, \varepsilon)\phi, 1\phi - 1) = 0. \quad (3.5)$$

This allows the possibility to follow a path of X_o - or S-breaking bifurcation points by ε -continuation. Observe that $f(., \varepsilon)$ has to satisfy the same invariance condition (I_o) or (I_s) as g (for each ε).

2. (a) Replace $g_x(x, \lambda)\phi$ by a one-sided difference quotient and set

$$F_o(x, \phi, \lambda, \varepsilon) := (g(x, \lambda), \varepsilon^{-1}(g(x + \varepsilon\phi, \lambda) - g(x, \lambda)), 1\phi - 1), \quad \varepsilon \neq 0. \quad (3.6)$$

Then it follows the existence of smooth $x(\varepsilon)$, $\phi(\varepsilon)$, $\lambda(\varepsilon)$ such that $x(0) = x_o$, $\phi(0) = \phi_o$, $\lambda(0) = \lambda_o$ and

$$g(x(\varepsilon), \lambda(\varepsilon)) \equiv 0 \text{ (branch } \mathcal{E}_o), \quad g(x(\varepsilon) + \varepsilon\phi(\varepsilon), \lambda(\varepsilon)) \equiv 0 \text{ (branch } \mathcal{E}_1).$$

It follows easily that $\lambda'(0) = -A_o/2B_o$, $x'(0) = \lambda'(0)v_o$.

This is essentially the method of Crandall/Rabinowitz (71) who did not explicitly use the regularity of the system $G_o = 0$.

(b) Assume (I_s) , replace $g_x(x, \lambda)\phi$ by a central difference quotient and $g(x, \lambda)$ by a certain mean value and set

$$F_{sa}(x, \phi, \lambda, \varepsilon) := \begin{cases} (g(x+\varepsilon\phi, \lambda) + g(x-\varepsilon\phi, \lambda))/2 \\ (g(x+\varepsilon\phi, \lambda) - g(x-\varepsilon\phi, \lambda))/2\varepsilon \\ 1\phi-1 \end{cases}, \quad \varepsilon \neq 0 \quad (3.7)$$

One gets an analytical expression of the branch \mathcal{C}_1 of non-symmetric solutions:

$$g(x(\varepsilon) \pm \varepsilon\phi(\varepsilon), \lambda(\varepsilon)) \equiv 0, \quad \phi(\varepsilon) = -\phi(-\varepsilon).$$

3. Introduce an eigenvalue ε of $g_x(x, \lambda)$ and set

$$F_o(x, \phi, \lambda, \varepsilon) := (g(x, \lambda), g_x(x, \lambda)\phi - \varepsilon\phi, 1\phi-1).$$

One obtains a parametrization of \mathcal{C}_o by an eigenvalue ε of $g_x(x, \lambda)$, and it is now not difficult to prove the known result that for algebraic simple eigenvalues $(\psi_o, \phi_o \neq 0)$, on \mathcal{C}_o an eigenvalue ε crosses zero with nonzero velocity.

4. DOUBLE SINGULAR S-POINTS.

Let (x_o, λ_o) be a double singular point of (1.1):

$$\dim N(g_x^o) = 2 = \text{codim } R(g_x^o).$$

Assume that (I_s) holds and that $x_o \in X_s$. Then S maps $N(g_x^o)$ into itself. Since $S^2 = I$, it follows that $N(g_x^o) \subset X_s$, $N(g_x^o) \subset X_a$ or

$$N(g_x^o) = \text{span } \{\phi_s^o, \phi_a^o\}, \quad \phi_s^o \in X_s, \quad \phi_a^o \in X_a \quad (4.1)$$

$$R(g_x^o) = \{y \in X: \psi_s^o y = 0 \text{ and } \psi_a^o y = 0\}, \quad (4.2)$$

$$\psi_s^o x = 0 \text{ for } x \in X_a, \quad \psi_a^o x = 0 \text{ for } x \in X_s. \quad (4.3)$$

We will call (x_o, λ_o) a *double singular S-point* if $x_o \in X_s$ and (4.1)-(4.3) are satisfied. There are several applications where those points occur: the Brusselator model in Schaeffer/Golubitsky (81), the rod-spring model in Bauer/Keller/Reiss (75), the buckling problem of rectangular plates in Chow/Hale/Mallet-Paret (76) and the two-cell reaction model in Werner/Spence (83). Two further examples are given below.

Bifurcation results can be found in Golubitsky/Schaeffer (79b), Shearer (80) and Raugel (82). The Lyapunov-Schmidt method reduces (1.1) to a problem

$g(x, \lambda) = 0$, $x \in X := \mathbb{R}^2$, where

$$g_1(x_1, -x_2, \lambda) = g_1(x_1, x_2, \lambda), g_2(x_1, -x_2, \lambda) = -g_2(x_1, x_2, \lambda) \quad (4.4)$$

$$g_i(0, 0, 0) = 0, (\partial g_i / \partial x_j)(0, 0, 0) = 0, i, j = 1, 2. \quad (4.5)$$

Here $(0, 0, 0)$ is a double singular S-point with the reflection $S := \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$.

A double singular S-point (x_0, λ_0) of (1.1) is a simple singular point of $g_S(x, \lambda) = 0$, where g_S is the restriction of g to $X_S \times \mathbb{R}$. We consider the following two types of double singular S-points:

TYPE I (x_0, λ_0) is a simple quadratic turning point of $g_S(x, \lambda) = 0$:

$$\psi_{S\lambda}^O \neq 0, A_S^O := \psi_S^O g_{xx}^O \phi_S^O \neq 0.$$

TYPE II (x_0, λ_0) is a simple x_0 -breaking bifurcation point of $g_S(x, \lambda) = 0$:

$$(I_O) \text{ holds with } x_0 \in X_S, \phi_S^O \neq x_0, \psi_S^O g_\lambda^O = 0,$$

$$B_S^O := \psi_S^O (g_x^O + g_{xx}^O v_0) \phi_S^O \neq 0, \text{ where } g_\lambda^O + g_{xx}^O v_0 = 0, v_0 \in X_0.$$

For both types it is clear how branches of symmetric solutions near (x_0, λ_0) look like (see figure 1). Moreover, the systems $G_{ss} = 0$ (type I) and $G_{os} = 0$ (type II) are regular for (x_0, λ_0) . But what about the system $G_{sa} = 0$?

THEOREM 4.1

(a) Let (x_0, λ_0) be a double singular S-point of type I. Then $G_{sa} = 0$ is regular for (x_0, λ_0) if and only if

$$D_0^O := \psi_a^O g_{xx}^O \phi_S^O \neq 0.$$

(In that case we call (x_0, λ_0) a double S-breaking (quadratic) turning point).

(b) Let (x_0, λ_0) be a double singular S-point of type II and let

$$B_a^O := \psi_a^O (g_x^O + g_{xx}^O v_0) \phi_a^O \neq 0.$$

Then $G_{sa} = 0$ is not regular for (x_0, λ_0) and it is $\dim N(DG_{sa}^O) = 1$, $\text{codim } R(DG_{sa}^O) = 1$. But $G_{oa} = 0$ is regular for (x_0, λ_0) .

Th.4.1(a) is identical with Th.3.2 in Werner/Spence (83). Part (b) can be proved with the techniques in Spence/Werner (82) and Werner/Spence (83).

The Brusselator model in Schaeffer/Golubitsky (81), e.g.

$$x \in \mathbb{R}^2, g_1(x_1, x_2, \lambda) = x_1^2 + x_2^2 + \lambda x_1, g_2(x_1, x_2, \lambda) = 0.8x_1^2 + \lambda x_2$$

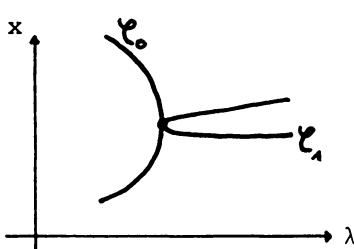
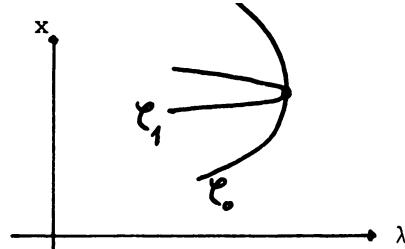
(c. (4.4), (4.5)) shows that there are not necessary non-symmetric solutions

near a bifurcation point of type II. But at double S-breaking turning points it can be shown with the method based on (3.7) that non-symmetric solutions bifurcate. The two intersecting branches are:

$$\xi_0 = \{(\lambda_0 + \alpha^2 \gamma_s + O(\alpha^3), x_0 + \alpha \phi_s^0 + O(\alpha^2)\} \subset X_s \times \mathbb{R} \quad (4.6)$$

$$\xi_1 = \{(\lambda_0 + \beta^2 \gamma_a + O(\beta^3), x_0 + \beta \phi_a^0 + O(\beta^2)\} \not\subset X_s \times \mathbb{R}, \text{ where} \quad (4.7)$$

$$\gamma_s := -A_s^0 / (2\psi_s^0 g_\lambda^0), \quad \gamma_a := -A_a^0 / (2\psi_a^0 g_\lambda^0), \quad A_a^0 := \psi_s^0 g_{xx}^0 \phi_a^0 \phi_a^0.$$

figure 2a: $\gamma_a \gamma_s < 0$ figure 2b: $\gamma_a \gamma_s > 0$

Double S-breaking turning points

Remark 4.2 For (4.4), (4.5) we have

$$\psi_s^0 g_\lambda^0 = \partial g_1^0 / \partial \lambda, \quad A_s^0 = \partial^2 g_1^0 / \partial x_1^2, \quad D_0 = \partial^2 g_2^0 / (\partial x_1 \partial x_2), \text{ etc.}$$

Remark 4.3 In general an isolated solution of $G_{sa} = 0$ will correspond to a simple S-breaking bifurcation point. But if on the symmetric branch ξ_0 , a simple turning point coalesce with a simple S-breaking bifurcation point to a double S-breaking turning point, Th.4.1(a) says that $G_{sa} = 0$ is still regular. Hence the numerical computation of simple S-breaking bifurcation points based on $G_{sa} = 0$ does not suffer under neighboured turning points.

Example I (Complexification of a real turning point problem, see Allgower/Georg (83) and Allgower in these proceedings)

Let $h: \mathbb{R}^n \times \mathbb{R} \rightarrow \mathbb{R}^n$ possess a complex analytic extension

$$H: \mathbb{C}^n \times \mathbb{R} \rightarrow \mathbb{C}^n, \quad H(\bar{z}, \lambda) = \overline{H(z, \lambda)}, \quad z \in \mathbb{C}^n, \quad \lambda \in \mathbb{R}.$$

Identifying $z=u+iv \in \mathbb{C}^n$ with $x=(u, v) \in \mathbb{R}^{2n}$, H is transformed into $g: \mathbb{R}^{2n} \times \mathbb{R} \rightarrow \mathbb{R}^{2n}$, where

$$g(u, v, \lambda) := \begin{pmatrix} g^r(u, v, \lambda) \\ g^i(u, v, \lambda) \end{pmatrix} := \frac{1}{2} \begin{pmatrix} H(u+iv, \lambda) + H(u-iv, \lambda) \\ -i(H(u+iv, \lambda) - H(u-iv, \lambda)) \end{pmatrix}.$$

(I_s) is satisfied since

$$g^r(u, -v, \lambda) = g^r(u, v, \lambda), \quad g^i(u, -v, \lambda) = -g^i(u, v, \lambda).$$

(u, o) are the symmetric (real) and (o, v) the anti-symmetric (imaginary) elements. From the Cauchy-Riemann differential equations

$$g_u^r = g_v^i, \quad g_v^r = -g_u^i$$

it follows that each (real) singular point (u_o, o, λ_o) of $g(x, \lambda) = o$ has even multiplicity $2k$ and that (u_o, λ_o) is a k -multiple singular point of $h(u, \lambda) = o$. Let $k=1$ and let $\phi_o(\psi_o)$ be the kernel (cokernel) vector of $h_u(u_o, \lambda_o)$. Then (u_o, o, λ_o) is a double singular S-point of $g(x, \lambda) = o$ with

$$\begin{aligned} \phi_s^o &= (\phi_o, o), \quad \phi_a^o = (o, \phi_o), \quad \psi_s^o = (\psi_o, o), \quad \psi_a^o = (o, \psi_o) \\ \psi_s^o \phi_a^o &= \psi_o h_\lambda^o, \quad A_s^o = -A_a^o = D_o = \psi_o h_{uu}^o \phi_o \phi_o. \end{aligned}$$

Moreover, (u_o, o, λ_o) is a double S-breaking quadratic turning point of $g(x, \lambda) = o$ if and only if (u_o, λ_o) is a simple quadratic turning point of $h(u, \lambda) = o$.

Observe that $\gamma_s = -\gamma_a \neq 0$ in (4.6), (4.7) (c. figure 2a).

To make use of Th.4.1(b) consider a two-parameter problem

$$f(x, \lambda, \mu) = o,$$

where $f: X \times \mathbb{R}^2 \rightarrow X$ is smooth, $f(., \mu_o) = g$ and where $f(., \mu)$ satisfies the same invariance conditions as g does. The corresponding extended systems will be denoted by $F=o$, $F_{sa}=o$, $F_{oa}=o$, etc. F_{sa} satisfies also an invariance condition:

$$(I_o)' \quad F_{sa} \text{ maps } Y_{oa} \times \mathbb{R} \text{ into } Y_{oa} (= X_o \times X_a \times \mathbb{R}).$$

It is $(x_o, \phi_a^o) \in Y_{oa}$, and it can be shown that the kernel vector of $N(DG_{sa}^o)$ in Th.41(b) is not in Y_{oa} . Hence $(x_o, \phi_a^o, \lambda_o, \mu_o)$ is a simple Y_{oa} -breaking singular point of $F_{sa}=o$ (w.r.t. μ). Moreover, we have

THEOREM 4.4

A double singular S-point (x_o, λ_o) of type II which satisfies $B_a^o \neq 0$ corresponds under the above assumptions to a simple Y_{oa} -breaking bifurcation point $(x_o, \phi_a^o, \lambda_o, \mu_o)$ of $F_{sa}=o$ (w.r.t. μ) if and only if the following unfolding condition is satisfied:

$$\det \begin{bmatrix} \psi_s^o (f_{x\lambda}^o + f_{xx}^o v_o) \phi_s^o & \psi_s^o (f_{x\mu}^o + f_{xx}^o w_o) \phi_s^o \\ \psi_a^o (f_{x\lambda}^o + f_{xx}^o v_o) \phi_a^o & \psi_a^o (f_{x\mu}^o + f_{xx}^o w_o) \phi_a^o \end{bmatrix} \neq 0, \quad \text{where} \quad (4.8)$$

$w_o \in X_o$ is defined by $f_\mu^o + f_{x\mu}^o w_o = o$. (Observe that the elements in the first

column are B_s^O and B_a^O , both nonzero).

Proof: One has to show that (4.8) is equivalent to

$$\psi_O(F_{sa,y\mu}^O + F_{sa,yY_O}^O)\phi_O \neq 0, \text{ where}$$

$\phi_O (\psi_O)$ spans the kernel (cokernel) of $DG_{sa,y}^O = F_{sa,y}^O$ and where $Y_O \in Y_{oa}$ is given by

$$F_{sa,y}^O V_O + F_{sa,\mu}^O = 0.$$

This is rather technical and will be omitted.

Remark 4.5 For fixed μ , each zero of $F_{sa}(.,\mu)=0$ corresponds in general to a simple S-breaking bifurcation point of $f(x,\lambda,\mu)=0$ (w.r.t. λ). The non- Y_{oa} -branch of $F_{sa}=0$ corresponds to simple "secondary" S-breaking bifurcation points (x,λ) of $f(x,\lambda,\mu)=0$ for fixed $\mu \neq 0$ ($x \in X_s$, but $x \notin X_O$). See also Shearer (80) for a similar result if $X_O=\{0\}$. His condition (D₁) is then identical with the condition (4.8).

Remark 4.6 Th.4.1 does not provide regular systems which find systematically double singular S-points of $g(x,\lambda)=0$. From codimension arguments it follows that one has to use at least another unfolding parameter μ . Determining systems for double singular S-points of type I and II are

$$F(x, \phi_s, \phi_a, \lambda, \mu) := \begin{bmatrix} f(x, \lambda, \mu) \\ f_x(x, \lambda, \mu) \phi_s \\ f_x(x, \lambda, \mu) \phi_a \\ l_s \phi_s^{-1} \\ l_a \phi_a^{-1} \end{bmatrix} = 0, \quad F: X_O \times X_s \times X_a \times \mathbb{R}^2 \rightarrow X_O \times X_s \times X_a \times \mathbb{R}^2 \quad (4.9)$$

(Set $X_O:=X_s$ in (4.9) for singular points of type I). l_s and $l_a \in X^*$ are again suitable scaling functionals ($l_s \phi_s^O=1$, $l_a \phi_a^O=1$). We do not state the regularity condition of (4.9) for singular points of type I. We only mention that - under the condition $B_a^O \neq 0$ - (4.9) is regular for singular points of type II if and only if the unfolding condition (4.8) is satisfied. In the interesting case $X_O=\{0\}$, (4.9) reduces to

$$f_x(0, \lambda, \mu) \phi_s = 0, \quad f_x(0, \lambda, \mu) \phi_a = 0, \quad l_s \phi_s = 1, \quad l_a \phi_a = 1.$$

In (4.8), one can set $v_O=0$ and $w_O=0$.

Example II (3-cell reaction model)

Consider (1.1) in $X=\mathbb{R}^3$, where

$$\begin{aligned}g_1(x_1, x_2, x_3, \lambda) &:= h(x_1, \lambda) + \varepsilon(x_2 + x_3) \\g_2(x_1, x_2, x_3, \lambda) &:= h(x_2, \lambda) + \varepsilon(x_1 + x_3) \\g_3(x_1, x_2, x_3, \lambda) &:= h(x_3, \lambda) + \varepsilon(x_1 + x_2)\end{aligned}$$

and $h: \mathbb{R}^2 \rightarrow \mathbb{R}$ is smooth. $g(x, \lambda)=0$ can be considered as steady state equations for the concentrations or temperatures of 3 chemical cells being in symmetric contact to each other. ε can be regarded as contact coefficient.

Here (I_γ) is satisfied with the symmetry group

$$\gamma := \{I, R, R^2, S, RS, SR\}, \text{ where}$$

$$R := \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix} \quad (\text{rotation}), \quad S := \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (\text{reflection}).$$

Set $S_1 := S$, $S_2 := RS$, $S_3 := SR$. For each reflection S_i , (I_{S_i}) is satisfied ($i=1, \dots, 3$). The γ -symmetric elements are in

$$X_0 := \{x: x = Tx \text{ for each } T \in \gamma\} = \{x: x = Rx\} = \{(u, u, u) : u \in \mathbb{R}\}.$$

(I_0) is satisfied, and we have $X_0 \subset X_i := X_{S_i}$.

Assume that there is a simple quadratic turning point $(\bar{u}, \bar{\lambda})$ of $h(u, \lambda)=0$. Then $(\bar{u}, \bar{u}, \bar{u}, \bar{\lambda})$ is a triple singular point of $g(x, \lambda)=0$ in the uncoupled state $\varepsilon=0$. For small $\varepsilon \neq 0$ this singular point is splitted into a simple quadratic turning point $(\bar{u}(\varepsilon), \bar{u}(\varepsilon), \bar{u}(\varepsilon), \bar{\lambda}(\varepsilon))$ and a double singular point $(u_0(\varepsilon), u_0(\varepsilon), u_0(\varepsilon), \lambda_0(\varepsilon))$ of $g(x, \lambda)=0$. The latter is a double singular S_i -point of type II for each $i=1, 2, 3$. Hence from the γ -symmetric branch

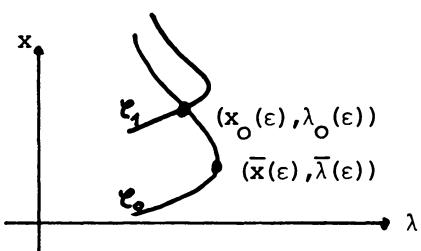


Figure 3

$e_0 \subset X_0 \times \mathbb{R}$, three S_1 -symmetric branches
 $e_i \subset X_i \times \mathbb{R}$ bifurcate in $(x_0(\varepsilon), \lambda_0(\varepsilon))$,
where $x_0(\varepsilon) := (u_0(\varepsilon), u_0(\varepsilon), u_0(\varepsilon))$.
The rotation R maps e_1 on e_2 and e_2
on e_3 . See figure 3, where only
 e_0 and e_1 are shown.

It is interesting to notice that each singular point (x_0, λ_0) of the 3-cell problem with $x_0 \in X_0$ and $N(g_X^0) \cap X_0 = \{0\}$ is necessarily a double singular S_i -point for each $i=1, 2, 3$.

For the exothermic reaction law $h(u, \lambda) = 2u - \lambda e^u$ and for $\varepsilon=0.5$ the following approximate values were obtained (thanks to Henning Wiebers)

$$\begin{aligned}\bar{\lambda}(\varepsilon) &= 0.368, \quad \lambda_0(\varepsilon) = 0.204, \\ \bar{u}(\varepsilon) &= 1, \quad u_0(\varepsilon) = 2.5.\end{aligned}$$

$(u_0(\varepsilon), u_0(\varepsilon), \lambda_0(\varepsilon))$ is a simple transcritical bifurcation point of $g_S(x, \lambda) = 0$, $S=S_i$, $i=1, 2, 3$. The S_i -symmetric branches γ_i which bifurcate from the γ -symmetric branch γ_0 have turning points $(x^i(\varepsilon), \lambda^i(\varepsilon))$, $i=1, 2, 3$. For $\varepsilon=0.5$ and $i=1$ one gets approximately

$$x^1(\varepsilon) = (2.15, 2.15, 2.67), \quad \lambda^1(\varepsilon) = 0.22.$$

5. REFERENCES

- ALLGOWER,E.L & GEORG,K. 1983 Predictor-Corrector and simplicial methods for approximating fixed points and zero points of nonlinear mappings. *To appear.*
- BAUER,L. & KELLER,H.B. & REISS,E.L. 1975 Multiple eigenvalues lead to secondary bifurcation. *SIAM Review* 17, 101-122.
- CHOW,S.N. & HALE,J.K. & MALLET-PARET,J. 1976 Application of generic bifurcation II. *Arch. for Rat. Mech. and Anal.* 62, 209-235.
- CRANDALL,M.G. & RABINOWITZ,P.H. 1971 Bifurcation from simple eigenvalues. *J. Funct. Anal.* 8, 321-340.
- GOLUBITSKY,M. & SCHAEFFER,D. 1979a A theory of imperfect bifurcation via singularity theory. *Comm. Pure Appl. Math.* 32, 21-98.
- GOLUBITSKY,M. & SCHAEFFER,D. 1979b Imperfect bifurcation in the presence of symmetry. *Commun. Math. Phys.* 67, 205-232.
- MOORE,G. 1980 The numerical treatment of non-trivial bifurcation points. *Numer. Funct. Anal. and Optimiz.* 2, 441-472.
- RAUGEL,G. 1982 Finite dimensional approximation of bifurcation problems in presence of symmetries. *Rapport interne No.81, Centre de Math. Appl., Ecole Polytechnique, Palaiseau.*

- SATTINGER,D.H. 1979 Group theoretic methods in bifurcation theory. *Springer Berlin-Heidelberg-New York*
- SCHAFFER,D. & GOLubitsky,M. 1981 Bifurcation analysis near a double eigenvalue of a model biochemical reaction. *Arch. Rat. Mech. and Anal.* 75, 315-347.
- SEYDEL,R. 1979 Numerical computation of branch points in nonlinear equations. *Numer. Math.* 33, 339-352.
- SHEARER,M. 1980 Secondary bifurcation near a double eigenvalue. *SIAM J. Math. Anal.* 11, 365-389.
- SPENCE,A. & WERNER,B. 1982 Nonsimple turning points and cusps. *IMA J. Numer. Anal.* 2, 413-427.
- WERNER,B. & SPENCE,A. 1983 The computation of symmetry-breaking bifurcation points. *To appear in SIAM J. Numer. Anal.*

Bodo Werner
Institut für Angewandte Mathematik
Universität Hamburg
Bundesstr. 55
D-2000 Hamburg

NEWTON ITERATES FOR POSITIVE SOLUTIONS OF A CLASS
OF NONLINEAR EIGENVALUE PROBLEMS

by María A. Astaburuaga, Jaime Figueroa and Jürgen Weyer

Summary:

We consider the nonlinear elliptic equation $F(u) = -\operatorname{div} A(x, \operatorname{grad} u) + b(x, u) = \lambda g(x, u)$ with Dirichlet boundary conditions. The vector field A and the functions b and g are monotone with respect to the second variable. Then, if λ is less than the first eigenvalue of the linearized problem, the Newton iterates given by $v_0 = 0$ and $F(v_{n+1}) = \lambda[g(x, v_n) + g_u(x, v_n)(v_{n+1} - v_n)]$ converge uniformly to the minimal positive solution. If g is convex (concave), then the convergence is from below (above). Concavity implies uniqueness. Picard and Newton approximations have better numerical conditions than the original problem.

I. INTRODUCTION

The starting point of this paper is the time independent diffusion equation $Lu = -\operatorname{div}(A(x)\operatorname{grad} u) = \lambda g(x, u)$ with Dirichlet boundary conditions. Here g denotes a monotone increasing function with respect to u and $A(x)$ is a positive definite matrix, such that L is a linear elliptic operator. In the year 1967 H. Keller and D. Cohen [4] first considered such nonlinear eigenvalue problems. Defining the Picard iterates $u_n(x)$ by $u_0 = 0$ and $Lu_{n+1} = \lambda g(x, u_n)$ they showed that the Picard iterates converge uniformly from below to the minimal positive solution, whenever such a solution exists. Further, they characterized the admissible eigenvalue parameters λ by the eigenvalues of the linearized problem. In the same year D. Cohen [3] studied the above diffusion equation also by the method of Newton iterates defined by $Lv_{n+1} = \lambda[g(x, v_n) + g_u(x, v_n)(v_{n+1} - v_n)]$. These iterates converge uniformly from above to our unique positive solution if g is also concave with respect to u . If g is convex, then the Newton iterates converge uniformly from below to the minimal positive solution.

In the meantime sharper results are due to the authors H. Amann, D. Cohen, H. Keller, T. Laetsch, J.W. Mooney, D.H. Sattinger, L.F. Shampine e.a..

Recently J. Weyer [6] replaced the linear diffusion law by a nonlinear law and considered the "doubly nonlinear" problem

$$F(u) = -\operatorname{div} A(x, \operatorname{grad} u) + b(x, u) = \lambda g(x, u) \quad (1)$$

with Dirichlet boundary conditions. Here A denotes a monotone vector field with respect to the second variable and b and g are monotone increasing functions in u . Defining the Picard iterates

$$u_0 = 0 \text{ and } F(u_{n+1}) = g(x, u_n) \quad (2)$$

the author [6] obtained results analogous to Keller and Cohen [4]. In this paper we consider the Newton iterates

$$v_0 = 0 \text{ and } F(v_{n+1}) = \lambda[g(x, v_n) + g_u(x, v_n)(v_{n+1} - v_n)] \quad (3)$$

and obtain results analogous to Cohen [3] for the "doubly nonlinear" equation (1). Further, we study the uniqueness of the solutions of (1). This leads to special uniqueness results for equations of the form $Lu = \lambda h(x, u)$, where h denotes a function, which is not necessarily monotone with respect to u .

Using Picard or Newton iterates for the solution of (1) we have replaced one nonlinear problem by the sequence of nonlinear problems (2+3). Thus, at the first glance, one has the impression that these iteration procedures are only useful for uniqueness results and a-priori estimates for the admissible eigenvalue parameters λ , while one prefers "direct" methods for the numerical solution of (1). However, if the numerical condition of (1) is bad (for example if the solution of (1) has a singularity near the boundary) a direct numerical integration may fail. In this case, the nonlinear equations represented by the Picard or Newton iterates (2+3) have a better numerical condition and the iterates finally lead to the solution of (1). This will be demonstrated by an example with controlled singularity.

Throughout this paper we make the following assumptions: suppose that $\Omega \subset \mathbb{R}^r$ is bounded and $Lu = -\operatorname{div}(C(x)\operatorname{grad} u) + c_0(x)u$ is an uniformly elliptic, self-adjoint second order operator. The coefficients $c_{ij}(x) = c_{ji}(x)$ of the matrix $C(x)$ are continuously differentiable. Let $c_0(x)$ be a continuous positive function. Define $N(u) = -\operatorname{div} K(x, \operatorname{grad} u)$, where $K(x, p)$ is a continuously differentiable vector field from $\Omega \times \mathbb{R}^r$ into \mathbb{R}^r , which is monotone in the space \mathbb{R}^r with respect to p for any fixed x . That means: $(K(x, p) - K(x, \tilde{p}), p - \tilde{p}) \geq 0$ for every x, p, \tilde{p} . Further, suppose that $b(x, u)$ is continuous and $g(x, u)$ is a C^2 -function from $\Omega \times \mathbb{R}$ into \mathbb{R} ,

such that b is monotone increasing and g is strictly monotone increasing with respect to the variable u . Further, assume $g(x,0) > 0$ and

$$b(x,0) - \sum_{i=1}^r \frac{\partial K}{\partial x_i}(x,0) \leq 0$$

for all $x \in \Omega$. Let us remark, that in the case where b and g are only monotone increasing for $u \geq 0$, we can replace $b(x,u)$ and $g(x,u)$ for $u \leq 0$ by $-b(x,-u)$ and $-g(x,-u)$, respectively. This substitution does not affect the positive solutions of (1). Now define the nonlinear elliptic second order operator F by

$$F(u) = Lu + N(u) + b(x,u)$$

and suppose that the domain of L, N and F is given by

$$D(F) = \{u \in C^2(\Omega) \cap C^0(\bar{\Omega}) \mid u = 0 \text{ on } \partial\Omega\}.$$

Setting $A(x,p) = C(x)p + K(x,p)$ we obtain the notation of [6] and F is given by $F(u) = -\operatorname{div} A(x, \operatorname{grad} u) + b(x,u)$. Vice versa, if F is given as described in [6], we can identify $C(x) = A_p(x,0)$; $Lu = F'_0 u = -\operatorname{div}(C(x) \operatorname{grad} u) + b_u(x,0) u$ and $N(u) = -\operatorname{div}(A(x, \operatorname{grad} u) - C(x) \operatorname{grad} u)$.

II. THE NEWTON ITERATES

In order to show the convergence of Picard or Newton iterates in the case of the equation $Lu = \lambda g(x,u)$ Keller and Cohen had to use a "linear positivity lemma". Studying Picard iterates for equation (1), one has to consider a "nonlinear positivity lemma" [6]. The Newton iterates require the following

Lemma 1 (Monotonicity lemma): Suppose that $r(x) > 0$ is continuous on $\bar{\Omega}$ and u and v belong to $D(F)$. Assume that $0 < \lambda < \mu_1$, where $\mu_1 = \mu_1\{r(x)\}$ denotes the first eigenvalue of the problem

$$\left. \begin{array}{ll} Lh = \mu r(x)h & \text{on } \Omega \\ h = 0 & \text{on } \partial\Omega \end{array} \right\}. \quad (4)$$

Then the pointwise inequality

$$F(u) - \lambda r(x)u > F(v) - \lambda r(x)v \text{ on } \Omega \quad (5)$$

implies $u > v$ on Ω .

Proof: By (5) and by the definition $w = u-v$ there is a continuous function $f(x) > 0$ such that

$$L w - \lambda r(x)w = q(x) \quad (6)$$

where $q(x)$ is given by $q(x) = f(x) - N(u(x)) + N(v(x)) - b(x, u(x)) + b(x, v(x))$.

It is well known, that the solution $w(x)$ of (6) minimizes the functional

$$J(z) = \frac{1}{2} \int_{\Omega} \left\{ \sum_{i,j=1}^r c_{ij}(x) \frac{\partial z}{\partial x_i} \frac{\partial z}{\partial x_j} + c_0(x) z^2 - \lambda r(x) z^2 - 2q(x)z \right\}$$

in the class of all piecewise continuously differentiable functions $z(x)$ which vanish on $\partial\Omega$. Obviously, $|w|$ belongs to this class and one obtains:

$$\begin{aligned} J(w) - J(|w|) &= \int_{\Omega} q(x)(|w| - w) dx \\ &= \int_{\Omega} \{f(x)(|w|-w) + (b(x,u) - b(x,v))(w-|w|)\} dx \\ &\quad + \int_{\Omega} [K(x, \text{grad } u) - K(x, \text{grad } v)] \text{grad}(w-|w|) dx. \end{aligned}$$

Now, if $w(x_0) < 0$ for some $x_0 \in \Omega$, then the set $\Omega_- = \{x \in \Omega | w(x) < 0\}$ has a measure $\neq 0$ and $w - |w| = 2(u-v) < 0$ on Ω_- as well as $w - |w| = 0$ on $\Omega \setminus \Omega_-$. By the monotonicity of b and K we conclude $J(w) - J(|w|) > 0$, which contradicts the minimal property of w . Hence $w(x) \geq 0$ on Ω .

Thus, $F(u) - F(v) = \lambda r(x)w + f(x) > 0$ on Ω .

It was shown in [6] that F is inverse monotone and we obtain $u > v$ pointwise on Ω . Using the fact that $F(0) \leq 0$ the monotonicity lemma leads to the following

Lemma 2 (Positivity lemma): Suppose that the assumption of lemma 1 holds. Then, $F(u) - \lambda r(x)u > 0$ on Ω implies $u > 0$ on Ω .

Remark: This result improves the positivity lemma given in [6] and contains the linear case of Keller and Cohen [4]. By the above proof one can easily see that the statements of lemmata 1 and 2 are also true, if we replace all signs $>$ by \geq . Further, in general, it is not true, that (5) together with assumption $u > v$ implies $\lambda < \mu_1$.

Assumption R-0: Suppose that $0 \leq g_u(x, u) \leq r(x) \in C^0(\bar{\Omega})$ for all $x \in \Omega$, $u \geq 0$. Let $\mu_1\{r(x)\}$ denote the first eigenvalue of the problem (4). Let the operator F have the following properties: (i) If $\lambda < \mu_1\{r(x)\}$, then the Newton iterates $v_n(x)$ given by (3) are well defined; that is,

for any $v_n(x) \in D(F)$ there is a solution $v_{n+1}(x) \in C^2(\bar{\Omega}) \cap C^0(\bar{\Omega})$ satisfying $v_{n+1} = 0$ on $\partial\Omega$. (ii) There is a maximal monotone single-valued extension of F denoted by $\tilde{F} : D(\tilde{F}) \subseteq L^p(\Omega) \rightarrow L^q(\Omega)$ (with $1/p + 1/q = 1$) such that every operator solution $v(x)$ of $\tilde{F}(v) = \lambda g(x, v)$ satisfies $v \in D(F)$.

R-0 is fulfilled if F satisfies one of the following conditions

R-1, 2, 3, 4. This was shown in [6]. (Part (i) of R-0 follows, if we replace the operator $F(u)$ of the paper [6] by $F(u) - \lambda g_u(x, v_n)u$.)

R-1; The dimension is $r=1$ and $|b(x, u)| \leq \alpha u + \beta$ on $\bar{\Omega} \times \mathbb{R}_+$ for some $\alpha, \beta > 0$.

R-2; The dimension is $r=1$. Suppose that the functions $C(x) = C$

$K(x, p) = K(p)$ and $b(x, u) = b(u)$ do not depend on x .

R-3; The dimension r is arbitrary. $K(x, p) = 0$; that means $F = L$.

Assume: $b(x, u) \leq \alpha u^s + \beta$ on $\bar{\Omega} \times \mathbb{R}_+$ for some $\alpha, \beta, s > 0$.

R-4; The dimension is $r \leq 3$. Assume $F(u) = -\Delta u$ and $b(x, u) = b(u)$ does not depend on x .

Theorem 1: Suppose that one of the conditions R-0, 1, 2, 3, 4 holds.

Assume $g_{uu}(x, u) < 0$ on $\Omega \times \mathbb{R}_+$ (g strictly concave). Suppose

$0 < \lambda < \mu_1\{g_u(x, 0)\}$. Then, the Newton iterates v_n satisfy

$0 < u(x) < v_{n+1}(x) < v_n(x)$ pointwise on Ω for $n \geq 1$. Here u denotes the unique positive solution of (!). Further, v_n converges uniformly and monotonically downwards to u .

Proof: The proof of $0 < v_{n+1} < v_n$ is analogous to the proof of the corresponding statement in the "semi-nonlinear" case given by Cohen [3].

Identifying $f(x, u) = g(x, u)$ one only has to replace L by F in the paper of Cohen. If a two-term expression $L(v_n - v_{n+1})$ occurs in the proof of Cohen, one has to replace this expression by $F(v_n) - F(v_{n+1})$. Then the statement follows by the application of our lemmata 1 and 2 with $r(x) = g_u(x, 0)$ instead of the application of the original positivity lemma of Keller and Cohen.

Hence, there is a function $v(x)$ such that $v_n(x) \rightarrow v(x)$, where the limit is pointwise monotonically downwards. By R-0 all $v_n(x)$ are continuous and satisfy

$$\int_{\Omega} v_n^p(x) dx \leq \int_{\Omega} v_1^p(x) dx.$$

Now, Fatou's lemma implies that v_n also converges to v with respect to the topology of the space $L^p(\Omega)$ ($p > 1$). By the monotonicity of g the sequence $g(x, v_n(x))$ converges pointwise monotonically downwards to $g(x, v(x))$.

The inequality

$$\int_{\Omega} (g(x, v_n(x)))^q dx \leq \int_{\Omega} (g(x, v_1(x)))^q dx$$

together with Fatou's lemma yield that this convergence is also in the sense of the space $L^q(\Omega)$ ($1/p + 1/q = 1$). Further, $g_u(x, u)$ is monotone decreasing with respect to u . Thus, the absolute value of $d_n(x) = g_u(x, v_n) \cdot (v_{n+1} - v_n)$ is bounded by $g_u(x, 0) |v_{n+1} - v_n|$. Hence, $d_n \rightarrow 0$ with respect to the topology of $L^q(\Omega)$. Now, let \tilde{F} denote the maximal extension of F , according to R-O. Then \tilde{F} is closed. Summarizing we have $v_n \rightarrow v$ in $L^p(\Omega)$ and $\tilde{F}(v_n) = F(v_n) \rightarrow \lambda g(x, v)$ in $L^q(\Omega)$. The closedness of \tilde{F} together with part (ii) of R-O imply $v \in D(F)$ and v is a positive continuous solution of (1).

By Dini's theorem v_n converges uniformly to v . In section III we will see that the solution of (1) is unique. Hence $u = v < v_n$.

Theorem 2: Suppose that one of the conditions $R = 0, 1, 2, 3, 4$ holds. Assume $g_{uu}(x, u) > 0$ (g strictly convex) and $g_u(x, u) \leq r(x) \in C^0(\bar{\Omega})$ on $\Omega \times \mathbb{R}_+$. Suppose $0 < \lambda < \mu_1\{r(x)\}$. Then, the Newton iterates $v_n = v_n(\lambda; x)$ satisfy

$$u_n < v_n < v_{n+1} < u \tag{7}$$

pointwise on Ω for $n \geq 1$. Here u denotes the minimal positive solution of (1) and the u_n are the Picard iterates given by (2). Further, v_n converges uniformly and monotonically upwards to u .

Proof: By R-O the Picard iterates u_n are well defined and converge uniformly and monotonically upwards to the minimal positive solution u of (1) (see [6]). Thus v_n has the same convergence behavior if (7) is true. The inequalities $0 < v_n < v_{n+1} < u$ can be shown analogously to the corresponding statements in the "semi-nonlinear" case given by Mooney and Roach [5]. For that purpose one has to show

$$\mu_1\{g_u(x, v_n)\} > \mu_1\{r(x)\} \tag{8}$$

by the means of [5, lemma 3.2]. Then, the application of our improved lemmata 1 and 2 instead of the original positivity lemma of Keller and Cohen [4] yields $0 < v_n < v_{n+1} < u$. The statement $u_n < v_n$ follows by induction. From (2) and (3) we get: $F(v_1) - F(u_1) = \lambda v_1 g_u(x, 0) > 0$. Hence, $v_1 > u_1$ by (8) and

lemma 1. Now suppose that $v_n > u_n$. Then we have:

$$F(v_{n+1}) - F(u_{n+1}) = \lambda[g(x, v_n) - g(x, u_n) + g_u(x, v_n)(v_{n+1} - v_n)] > \lambda g_u(x, v_n)(v_{n+1} - v_n).$$

By (8) and lemma 1 we obtain $v_{n+1} > u_{n+1}$ and the proof is complete.

III. UNIQUENESS

Theorem 3: Suppose that one of the conditions $R = 0, 1, 2, 3, 4$ holds.

Assume $g_u(x, u) \leq r(x) \in C^0(\bar{\Omega})$ on $\bar{\Omega} \times \mathbb{R}_+$. Then there is an unique positive solution of (1) provided that $0 < \lambda < \mu_1\{r(x)\}$. (Note that $g_{uu}(x, u) < 0$ implies $g_u(x, u) \leq r(x) = g_u(x, 0)$.)

Proof: By integration we obtain $g(x, u) \leq g(x, 0) + r(x)u$. In [6] it was shown that this implies the existence of a minimal positive solution u of (1) provided that $0 < \lambda < \mu_1\{r(x)\}$. Let v denote another solution of (1). Then we have $u(x) \leq v(x)$ on $\bar{\Omega}$. By the mean value theorem we conclude: $F(v) - F(u) = \lambda[g(x, v) - g(x, u)] \leq \lambda r(x)(v - u)$. Lemma 1 implies $v(x) \leq u(x)$ on $\bar{\Omega}$.

Corollary: Suppose that the monotone functions b and g depend only on u . Assume $b(0) \leq 0$ and $b'(u) - b'(0) \geq 0$; further $g(0) > 0$ as well as $0 < g'(u) \leq r \in \mathbb{R}_+$ for $u > 0$. Then, the equation

$$F(u) = Lu + b(u) = \lambda g(u) \quad \text{on } \Omega \tag{9}$$

with $u = 0$ on $\partial\Omega$ has an unique positive solution if $0 < \lambda g'(0) \leq \mu_1 + b'(0)$ or, in particular, if $0 < \lambda \leq b'(0)/g'(0)$. Here μ_1 denotes the first eigenvalue of the problem $Lh = \mu h$ with Dirichlet boundary condition.

Proof: By theorem 3 the equation $Lu + b'(0)u + b(u) - b'(0)u = \lambda g(u)$ has an unique positive solution, if $0 < \lambda < \bar{\mu}_1$ where $\bar{\mu}_1$ denotes the first eigenvalue of

$$Lh + b'(0)h = \bar{\mu} g'(0) h.$$

Hence, $\bar{\mu}_1 = (\mu_1 + b'(0)) / g'(0)$ which completes the proof.

Example: Studying chemical reactor theory R. Aris [2] suggested the following model:

$$-\Delta v = \gamma(1 + \sigma - v)e^{-k/v} \quad \text{in } \Omega$$

and $v = 1$ on $\partial\Omega$, where γ, σ, k are positive constants. Substituting $u = v - 1$ H. Amann [1] obtained the equivalent model

$$-\Delta u = \gamma(\sigma-u)e^{-k/(1+u)} \quad \text{in } \Omega \quad (10)$$

and $u = 0$ on $\partial\Omega$. Amann studied conditions for multiple solutions. Here, the above corollary allows a description of parameters for which the solution is unique. We define the functions $b(u) = \gamma u e^{-k/(1+u)} + \alpha e^{-1/(1+u)} - \alpha e^{-1}$ and $g(u) = \gamma \sigma e^{-k/(1+u)} + \alpha e^{-1/(1+u)} - \alpha e^{-1}$ and choose $\alpha > \gamma \sigma k^2$. Then, one can easily see that b and g satisfy the assumption of our corollary. Now, (10) has the structure (9) with $Lu = -\Delta$ and $\lambda = 1$. According to the corollary, problem (10) has an unique positive solution if $g'(0) \leq \mu_1 + b'(0)$. Thus, we have uniqueness for $\mu_1 \geq \gamma(\sigma k - 1)e^{-k}$. In particular, the solution of (10) is unique for sufficiently small and sufficiently great parameters k .

IV. NUMERICAL EXPERIENCE

If one is interested in the concrete numerical solution of (1) one will apply a "direct" integration method; for example, a single or multiple shooting method in one dimension. If such methods work, the Picard and Newton iterates are only of theoretical interest aside from the fact that they allow a-priori estimates for the admissible eigenvalue parameters λ . However, if direct integration methods fail in consequence of the bad numerical condition of problem (1), then Picard and Newton iterates have practical importance. We demonstrate this by an example which has a solution with a singularity near the boundary. In the case of a convex right hand side the Picard iterates are smaller than the Newton iterates and these are smaller than the solution. Hence, the slopes at the boundary of the Picard iterates are less than the slopes of the Newton iterates and the slopes of these are less than the slopes of the solution. Thus, the approximating nonlinear equations (2) and (3) have better numerical conditions than (1). Here, we only study the numerical improvement by Newton iterates and consider the following "convex" problem:

$$-(a(x,u'))' + b(x,u) = g(x,u) \quad \text{in } (0,1) \quad (11)$$

with $u(0) = u(1) = 0$. We choose $a(x,p) = (x^2 - x - \delta)p^3 + (x^2 - x - \delta)^2 p$ and $g(x,u) = 2u + e^{-u} - 1 + 6\delta^3(1 - 2x)^2 + 2\delta$ as well as $b(x,u) = (u+1)^3 - (v+1)^3 + 2v + e^{-v} - 1$ with

$$v(x) = \frac{x(x-1)}{(x+\epsilon)(x-1-\epsilon)}$$

and $\delta = \varepsilon + \varepsilon^2$ for $\varepsilon > 0$. It is easy to see that (11) satisfies all assumptions of theorem 2. The exact solution of (11) is $v(x)$, whose singularity near the boundary is controlled by the parameter ε . For $\varepsilon \rightarrow 0$ the function $v(x)$ converges to the "rectangle"

$$\bar{v}(x) = \begin{cases} 0 & \text{if } x \in \{0,1\} \\ 1 & \text{if } x \in (0,1) \end{cases}$$

The numerical treatment is problematical only near the value $x = 1$.

Here we get:

ε	method	$x = 0.90$	$x = 0.92$	$x = 0.94$	$x = 0.95$	$x = 0.98$	$x = 1$
0.04	E	0.683	0.638	0.575	0.480	0.320	0.000
	D	0.681	0.635	0.570	0.474	0.314	0.004
0.03	E	0.744	0.704	0.646	0.554	0.388	0.000
	D	0.741	0.699	0.638	0.541	0.364	- 0.048
	N1	0.714	0.673	0.614	0.520	0.347	- 0.062
	N2	0.740	0.698	0.637	0.543	0.375	0.011
0.02	E	0.815	0.783	0.734	0.653	0.490	0.000
	D	0.805	0.765	0.700	0.572	0.244	- 1.097
	N1	0.752	0.692	0.576	0.316	- 0.447	- 3.270
	N2	0.805	0.770	0.710	0.603	0.365	0.016

Here, E denotes the exact solution while D is the result obtained by the "direct" numerical integration of (1). By N1 and N2 we denote the first and second Newton iterates of the approximating problem (3). Naturally, N1 and N2 are obtained by "direct" integration of (3). For the "direct" integration we always have used a single shooting method combined with a rather poor Runge-Kutta procedure (4 evaluations, fixed step size 1/100). Clearly, by a more sofisticated numerical integration one can calculate the "direct" solution of (1) even for smaller ε . However, any numerical integration method will fail for sufficiently small ε . In this case Picard and Newton

iterates are an alternative.

We note that near the singularity ($x \geq 0,98$) we do not have $0 < N_1 < N_2$ although $0 < v_1(x) < v_2(x)$ is required theoretically. This phenomenon is a consequence of the rounding errors and does not occur on the stable part of the equation. However, in spite of the rounding errors in N_1 , higher Newton iterates may improve the "direct" solution D on the whole domain.

References:

- [1] AMANN H., Existence of multiple solutions for nonlinear elliptic boundary value problems, Indiana Univ. Math. J. 21 (1972), 925-935.
- [2] ARIS R., On stability criteria of chemical reaction engineering, Chemical Engineering Sci. 24 (1969), 149-169.
- [3] COHEN D.S., Positive solutions of a class of nonlinear eigenvalue problems, J. Math. Mech. 17 (1967), 209-215.
- [4] KELLER H.B. & COHEN D.S., Some positone problems suggested by nonlinear heat generation, J. Math. Mech. 16 (1967), 1361-1367.
- [5] MOONEY J.W. & ROACH G.F., Iterative bounds for the stable solutions of convex nonlinear boundary value problems, Proc. Roy. Soc. Edinburgh (A) 76 (1976), 81-94
- [6] WEYER J., Picard iterates for positive solutions of a class of nonlinear eigenvalue problems, J. Appl. Math. Phys. (ZAMP), to appear

Maria A. Astaburuaga and
 Jaime Figueira
 Dpto. Matemática y CC
 Universidad de Santiago
 Casilla 5659 / Correo 2
 Santiago de Chile

Jürgen Weyer
 Mathematisches Institut
 Universität zu Köln
 Weyertal 86-90
 D-5000 Köln 41 / W. Germany