Springer Proceedings in Mathematics & Statistics

Bourama Toni *Editor*

Mathematical Sciences with Multidisciplinary Applications

In Honor of Professor Christiane Rousseau. And in recognition of the *Mathematics for Planet Earth* Initiative

Springer

Springer Proceedings in Mathematics & Statistics

Volume 157

More information about this series at http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Bourama Toni
Editor

Mathematical Sciences with Multidisciplinary Applications

In Honor of Professor Christiane Rousseau. And in recognition of the *Mathematics for Planet Earth* Initiative

Springer

Editor
Bourama Toni
Department of Mathematics and Economics
Virginia State University
Petersburg, VA, USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-319-31321-4 ISBN 978-3-319-31323-8 (eBook)
DOI 10.1007/978-3-319-31323-8

Library of Congress Control Number: 2016943204

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Dedication

The Virginia State University Interdisciplinary Seminar series and its Springer-published proceedings were inspired by Professor Christiane Rousseau's mathematical and scientific endeavors :

- Professor of Mathematics, Département de mathématiques et de statistique, Faculté des arts et des sciences, Université de Montréal, Montréal, Canada
- Member, Scientific Council, UNESCO International Basic Sciences Programme (2015–2017)
- President, Canadian Mathematical Society (2002–2004)
- Vice-President, International Mathematical Union (2011–2014)
- Member, Executive Committee, International Mathematical Union (2015–2018)

Christiane is a great mathematician and indeed passionate about mathematics, convinced of its powerful role as a tool to understand our world. She has always been committed to improving the image of mathematics in the media and in the

society as a whole. We present here a collection of quotes : some of them now well-known, some excerpts from a report by the Québec Science Reporter Mathieu-Robert Sauvé, with my translation and emphasis.

«C'est un outil indispensable pour comprendre le monde. J'explique à mes étudiants que, quand on ne comprend pas un phénomène, on doit mettre ses lunettes mathématiques. Elles permettent de voir différemment le monde qui nous entoure, au même titre que des lunettes 3D vous permettent de voir les reliefs d'une image»

(Mathematics is an indispensable tool for understanding the world. I explain to my students that, when faced with the complexity of any phenomenon, one must put on the mathematical lenses. They indeed allow to see the surrounding world differently, as 3D lenses allow one to see the many aspects of an image.)

«Pour connaître l'effet des changements climatiques, il faut concevoir des modèles à partir d'algorithmes. Il faut faire des prévisions économiques, chiffrer les répercussions sur les populations humaines. Les mathématiques sont au cœur de l'action humaine»

(In order to understand the effects of climate change, models must be designed from algorithms. There is a great need of accurate economic forecast accounting for its impact on human populations. Mathematics is indeed at the heart of all human endeavors !)

«Des enjeux comme les changements climatiques, le développement durable, les désastres créés par l'homme, le contrôle des maladies et épidémies, la gestion des ressources et l'intégration globale sont maintenant à l'avant-scène. Les mathématiques y jouent un rôle clé, ainsi que dans beaucoup d'autres processus affectant la planète Terre, tant comme discipline fondamentale que comme composante essentielle de recherche multidisciplinaire et interdisciplinaire.»

(Challenges such as climate change, sustainable development, man-made disasters, control of diseases and epidemics, management of natural resources and global integration are now at the forefront. As for many other processes impacting the planet Earth, Mathematics play indeed a key role, as a fundamental discipline and as well as an essential component of multidisciplinary and interdisciplinary research.)

«C'est avec de bonnes idées qu'on peut changer le monde»,

(It takes great ideas to positively change the world.)

We dedicate this volume of the STEAM-H series to Professor Christiane Rousseau.

May many generations of mathematicians and scientists continue to be inspired by her mathematical and scientific achievements and visions, her professional probity, and her tireless dedication to mathematics and the sciences.

You, the reader, is hereby kindly invited to share in Professor Rousseau's enthusiasm for Mathematics per se as a human knowledge, her vision and passion for mathematical methods as the ultimate beautiful, elegant, and efficient tool for addressing all the great challenges of human kind !

Bourama Toni

Foreword

It is hard to believe that, although a young and energetic person, Christiane Rousseau is celebrating her jubilee. It is a celebration of decades of creative work, devoted service to the mathematical community, and education at large.

Christiane began her research career under Prof. Dana Schlomiuk at the Université de Montréal. Dana is a highly inspired mathematician, and she shared her inspiration with her student. Christiane started to work independently soon after and became an expert in the theory of planar differential equations and their bifurcations. Her best known early works deal with isochronous centers.

The central problem in the theory of planar differential equations is undoubtedly Hilbert's 16th problem : what can be said about the number and location of limit cycles of a planar polynomial vector field with the components of degree n ? The so-called Hilbert number H(n) is associated to this problem. It is the upper bound of the numbers of limit cycles of the vector field mentioned above. Nobody knows whether H(n) exists, even for $n = 2$.

A fundamental contribution to the study of the existence of H(2) was made by Dumortier, Roussarie, and Rousseau at the beginning of the 1990s when they established and pushed forward the so-called 121-program. A prior result is a finiteness theorem due to R. Bamon (1986) : a quadratic polynomial vector field has but a finite number of limit cycles. In order to prove this theorem, it is necessary to check that limit cycles of a quadratic vector field cannot accumulate to a polycycle (also called graphic or separatrix polygon) of this vector field. There are not that many graphics that a quadratic vector field may have. Bamon studies 20 polycycles and proves that none of them may be an accumulation locus for limit cycles.

A similar idea lies at the basis of the 121-program, called *programme DDR* by Christiane. One should note however that Bamon's polycycles have a first return map, whereas the graphics in the 121-program do not have necessarily such a map, which is why there exists many more. In order to prove the existence of H(2), that is, the statement of the uniform boundedness of the number of limit cycles for quadratic vector fields, one should prove that no polycycle in this family generates an infinite number of limit cycles. That is, any polycycle has finite cyclicity. But the variety of graphics that may generate limit cycles is much wider than that of polycycles

that may be accumulation loci. For quadratic vector fields, this variety contains 121 samples. Up to now, finite cyclicity is proved for almost 80 cases out of 121, and the contribution of Christiane is crucial for this investigation. This supports the widely spread belief that H(2) does exist.

The turn of the millennium was also a turn in Christiane's research interests. She moved to the study of parabolic germs of maps $(\mathbf{C}, 0) \to (\mathbf{C}, 0)$ (the germs that have the identity as linear part) and their unfoldings. The functional invariants of the analytic classification of such germs were discovered by Birkhoff in 1939, but he did not prove that they may be actually realized, and his work remained partly forgotten. In 1981 Ecalle and Voronin proved the realization theorem for these invariants, completed the analytic classification of the parabolic germs, and found numerous applications to geometry. Since then, these invariants are called "Ecalle–Voronin moduli." After that, the unfolding of parabolic germs was studied. Under a perturbation, a parabolic fixed point splits into hyperbolic ones. It is an interesting problem to understand the relation between Sternberg normalizing charts for these hyperbolic points and Ecalle–Voronin moduli of the unperturbed parabolic point.

Partial progress was accomplished by Glutsyuk and Lavaurs in the 1990s, but the final step was taken by Rousseau and Christopher. In their seminal paper, they gave a complete answer to the question stated above. In the early 1980s Arnold discovered that numerous geometric problems contain "hidden dynamics." Consequently, several local classification problems from singularity theory may be solved with the use of Ecalle–Voronin moduli. Christiane solved several problems of the kind, including the classification of germs of couples of tangent analytic curves, the germ being at the tangency point.

The educational activity of Christiane goes far beyond her research domain. Her work reaches a broad audience of undergraduate students and high school teachers and even the public. In 2008, she was invited to deliver a "Regular Lecture" at the ICME11 (International Congress of Mathematical Education) in Monterrey, Mexico.

The grandfathers of those of us who are now in their 70s were born in the last decades of the nineteenth century. The only technical miracle at that time was the steam engine, giving rise to locomotives and steamers. No planes, no cars, no radio ! Within a few generations the world has changed completely. We are now surrounded by technical miracles. To what extent do they use mathematics ? There is a well-known opinion that fundamental sciences, mathematics included, have done so much by creating the theory of electromagnetism and thus completely transforming our life, that any future support to mathematics is more than well deserved. But what about the contribution of mathematics to our modern civilization ? Is it comparable to the contributions of the nineteenth century ?

These questions are important for common people, especially for the high school and college students. Christiane, together with her colleague Yvan Saint-Aubin, launched a lecture series at the Université de Montréal open to both undergraduate mathematics students and future high school teachers. Based on this course, they wrote a book entitled *Mathematics and Technology,* a kind of Bible on the modern applications of mathematics.

For this book, the authors were awarded the *2009 Adrien Pouliot Prize from Association mathématique du Québec*. The topics cover a wide range of domains : positioning on Earth and symmetry in arts, robotic motion and gamma-ray surgery, error-correcting codes and cryptography, mathematical foundations of Google and image compression, DNA computers, and many others. Note that all are peaceful applications of mathematics. None of the military kind, even those that are accessible to the public, are discussed in the book.

Christiane was President of the Canadian Mathematical Society between 2002 and 2004 and a Vice-President of the International Union of Mathematicians, 2010–2014. Her work at these positions is well characterized by her own words :

> In my career, I have managed to combine my teaching, research and training activities with other activities on the side of promotion of mathematics : popularization of mathematics with the public, involvement in preservice teacher education, and activities in the schools with kids and/or school teachers. In Canada, I have been actively involved in bringing the Canadian community together (learned societies, institutes, MITACS) for the organization of joint activities : joint meetings including Canada-France congresses, Canada mathematics Education Fora, etc, Canadian bids for ICM 2010 and 2014. I am now coordinating a North-American thematic year on Mathematics of Planet Earth in 2013.
>
> I have always enjoyed bringing communities to work together. As far as promotion of mathematics is concerned, there are a lot of benefits working together at the international level, and I look forward to working with IMU on this aspect. I hope to promote a great collaboration with ICME on mathematics education matters.

Let us mention only a few of Christiane's related activities. As a President of the CMS, she helped launch a Russian study-abroad program called "Math in Moscow." This provided Canadian students a chance to do "mathematics in the Russian style" and experience firsthand Russian culture. The "Nanum" program was also initiated during Christiane's tenure as a Vice-President of the IMU ; it is a program offering financial support for young mathematicians from developing countries, including graduate students. In particular, the program allowed for hundreds of young people all over the world to participate to the ICM in Seoul.

Christiane organized and co-organized more than a dozen of scientific conferences and schools. One of them, "Normal forms, bifurcations and finiteness problems in differential equations" organized in 2002, resulted in a book with the same title coedited by C. Rousseau, G. Sabidussi, and Yu. Ilyashenko. The school was a remarkable event that brought together leading experts in the field, as well as a lot of young participants including undergraduate and graduate students.

Christiane is one of the creators and a member of the Editorial Board of the popular magazine *Accromαth*, dedicated mainly to "Mathematics of the Planet Earth." This publication gathers articles written by mathematicians that bring to a wide audience the power and beauty of our modern science, in an elementary and clear way.

There is a famous cross-country skiing tradition in our mathematical community. For instance, Kolmogorov skied regularly 40 km before beginning his work days. Christiane keeps this tradition at a very high level. With her husband Serge Robert, she participates yearly to a two-day marathon in Québec and has won several gold medals. The distance is 80 km the first day, the night is slept in sleeping bags in

the snow, all that followed by 80 km the next day. Those who survive get the gold medal. Christiane loves the countryside. Together with her husband Serge, she built their permanent house with their own hands. They also built their country house. It is a three-hour ride from Montréal, located in the wilderness and surrounded by trees. They refused both gas and electricity, to have a life as close as possible to nature. Their cooking stove is made of stone and the house is heated by wood burning in a brick stove. It has a lot of small bedrooms for guests, one extra sign of Christiane's everlasting generosity.

This book is a small sign of gratitude of the mathematical and scientific community to Christiane Rousseau. Many happy returns of the day, Christiane !

Yulij Ilyashenko

Preface

The multidisciplinary STEAM-H series (Science, Technology, Engineering, Agriculture, Mathematics, and Health) brings together leading researchers to present their own work in the perspective to advance their specific fields and in a way to generate a genuine interdisciplinary interaction transcending disciplinary boundaries. All chapters therein were carefully edited and peer-reviewed ; they are reasonably self-contained and pedagogically exposed for a multidisciplinary readership.

Contributions are by invitation only and reflect the most recent advances delivered in a high standard, self-contained way. The goals of the series are :

(1) To foster student interest in science, technology, engineering, agriculture, mathematics and health.
(2) To enhance multidisciplinary understanding between the disciplines by showing how some new advances in a particular discipline can be of interest to the other discipline, or how different disciplines contribute to a better understanding of a relevant issue at the interface of mathematics and the sciences.
(3) To promote the spirit of inquiry so characteristic of mathematics for the advances of the natural, physical, and behavioral sciences by featuring leading experts and outstanding presenters.
(4) To encourage diversity in the readers' background and expertise, while at the same time structurally fostering genuine interdisciplinary interactions and networking.

Current disciplinary boundaries do not encourage effective interactions between scientists ; researchers from different fields usually occupy different buildings on university campuses, publish in journals specific to their field, and attend different scientific meetings. Existing scientific meetings usually fall into either small gatherings specializing on specific questions, targeting specific and small group of scientists already aware of each other's work and potentially collaborating, or large meetings covering a wide field and targeting a diverse group of scientists but usually not allowing specific interactions to develop due to their large size and a crowded program. Traditional departmental seminars are becoming so technical as to be largely inaccessible to anyone who did not coauthor the research being presented.

Here, contributors focus on how to make their work intelligible and accessible to a diverse audience, which in the process enforces mastery of their own field of expertise.

In honor of Professor Rousseau, a pioneer of mathematical approaches to human earthly challenges, this volume strongly advocates multidisciplinarity with the goal to generate new interdisciplinary approaches, instruments, and models including new knowledge, transcending scientific boundaries to adopt a more holistic approach. For instance, it should be acknowledged, following Nobel Laureate and President of the UK's Royal Society of Chemistry, Professor Sir Harry Kroto, "that the traditional chemistry, physics, biology departmentalised university infrastructures–which are now clearly out-of-date and a serious hindrance to progress–must be replaced by new ones which actively foster the synergy inherent in multidisciplinarity." The National Institutes of Health and the Howard Hughes Medical Institute have strongly recommended that undergraduate biology education should incorporate mathematics, physics, chemistry, computer science, and engineering until "interdisciplinary thinking and work become second nature." Young physicists and chemists are encouraged to think about the opportunities waiting for them at the interface with the life sciences. Mathematics is playing an ever more important role in the physical and life sciences, engineering, and technology, blurring the boundaries between scientific disciplines.

The series is to be a reference of choice for established interdisciplinary scientists and mathematicians and a source of inspiration for a broad spectrum of researchers and research students, graduate, and postdoctoral fellows ; the shared emphasis of these carefully selected and refereed contributed chapters is on important methods, research directions, and applications of analysis including within and beyond mathematics. As such, the volume promotes mathematical sciences, physical and life sciences, engineering, and technology education, as well as interdisciplinary, industrial, and academic genuine cooperation.

Toward such goals, the following chapters are featured in the current volume.

Chapter 1 by Faina Berezovskaya, studies a FitzHugh model modified to include a cross-diffusion connection between the potential and recovery variables, investigating successful propagation of an excitable neuron but also propagation failures, which are extremely important for many applications. The model demonstrates two types of behaviors, so-called "slow" and "fast" traveling waves.

Chapter 2 by Terence Blows, outlines a simple but imperfect approach to the study of degenerate foci and uses the method to give an example of a cubic system with four local limit cycles about a degenerate focus.

In Chap. 3 by Pietro-Luciano Buono and Raluca Eftimie, the authors establish the applicability of the Lyapunov–Schmidt reduction and the Centre Manifold Theorem for a class of hyperbolic partial differential equation models with nonlocal interaction terms describing the aggregation dynamics of animals/cells in a one-dimensional domain with periodic boundary conditions.

Chapter 4 by Magdalena Caubergh and Robert Roussarie, deals with relaxation oscillations from a generic balanced canard cycle subjected to three breaking

parameters of Hopf or jump type and proves that at most five relaxation oscillations bifurcate in a rescaled layer of the cycle.

In Chap. 5 by Colin Christopher, Wuria Muhammad Ameen, and Zhaoxia Wang, the authors first present cases where all integrability conditions are either uncovered by Darboux method or by a monodromic argument, and then investigate the integrability of the critical points which do not lie at the origin.

Chapter 6 by Morgan Craig, Mario González-Sales, Jun Li, and Fahima Nekka is a study to substantiate and situate the use of physiological modeling in pharmacometrics and provide incentives to continue to improve understanding of the underlying physiological mechanisms of a given system. It is also a testimony to the necessity of building bridges between diverse actors from different backgrounds (pharmaceutical scientists, clinicians, biomathematician, statisticians, engineers, etc.) in the pharmaceutical community to best serve patients and their needs.

Chapter 7 by Bui Xuan Dieu, Luu Hoang Duc, Stefan Siegmund, and Nguyen van Minh, is concerned with the strong stability of solutions of a class of non-autonomous equations with an unbounded operator in a Banach space and almost periodically time-dependent. A general condition on strong stability is given in terms of Perron conditions on the solvability of the associated inhomogeneous equation.

Chapter 8 by Mohamed El Morsalani and Abderaouf Mourtada, presents a new approach to the problem of limit cycles, which appear near hyperbolic polycycles of vector fields, upon a small deformation. Namely, the authors show a "preparation theorem" for quasi-regular functions, which appear as return maps associated to deformations of hyperbolic polycycles.

Chapter 9 by Raymond Fletcher, studies cubic curves that invert onto themselves, stemming for an investigation of group circle systems.

Chapter 10 by Lili Guadarrama, presents an emerging technique for noninvasive imaging with broad application in several disciplines including biomedical imaging and nondestructive testing. It summarizes different approaches for the imaging technique of elastography : quasi-static, harmonic, and transient elastography, along with establishing models for viscoelasticity.

Chapter 11 by Gerard Kientega is a chapter that studies affine completeness of algebras using a generalized metric to prove an extension theorem leading to new results such as an answer to a question of Karrli and Pixley.

In Chap. 12 by Bernd Krauskopf and Hinke M. Osinga, the authors review how a conjectural codimension-four unfolding of the full family of cubic Liénard equations helps to identify the central singularity as an excellent candidate for the organizing center that unifies different types of spiking action potentials of excitable cells. This point of view and the subsequent numerical investigation of the respective bifurcation diagrams led, in turn, to new insight on how this codimension-four unfolding manifests itself as a sequence of bifurcation diagrams on the surface of a sphere.

Chapter 13 by Yu Ilyashenko considers the long and glorious history of the theory of planar bifurcations which one could initially split into two parts : one part on local bifurcations such as the Poincaré–Andronov–Hopf bifurcation and another part on

nonlocal bifurcations dealing with the bifurcations of separatrix polygons and the polycycles such as separatrix loops of hyperbolic saddles and homoclinic curves of saddle-nodes. The chapter shows that there is a third part, not yet developed, that may be called "global bifurcations," the main features being termed "sparking bifurcations." They were discovered by Malta-Palis in the 1980s.

Chapter 14 by Chengzhi Li first introduces some basic concepts about slow-fast dynamics and its application to a biological model, a predator–prey system with response functions of Holling type, and a medical model, a SIS epidemic model with nonlinear incidence.

Chapter 15 by Pavao Mardesic, Dominique Sugny, and Leo van Damme, exposes the key role played by Abelian integrals in the infinitesimal version of the Hilbert 16th problem, as well as in the study of Hamiltonian monodromy of fully integrable systems. The authors treat in particular the simplest example presenting nontrivial Hamiltonian monodromy : the spherical pendulum.

Chapter 16 by Thanh Nguyen, Debarun Kar, Matthew Brown, Arunesh Sinha, Albert XinJiang, and Milind Tambe, the authors present how security is a critical concern around the world and computational game theory can help design security schedules in many domains from counterterrorism to sustainability where limited security resources prevent full security coverage at all times ; casting the problem as a Bayesian Stackelberg game, the authors developed new algorithms that are now deployed over multiple years in multiple applications for security scheduling. These applications are leading to real-world use-inspired research in the emerging research area of "security games" ; specifically, the research challenges posed by these applications include scaling up security games to large-scale problems, handling significant adversarial uncertainty, dealing with bounded rationality of human adversaries, and other interdisciplinary challenges.

Chapter 17 by Michael Pohrivchak, John Adam, and Umaporn Nuntaplook starts off with a nice review discussing some of the seminal advances of the last few centuries and their relation to electromagnetic radiation scattering off spheres of varied sizes, and then continues with an investigation of the backscattering of inhomogeneous spheres with different refractive index profiles, which affect the reflection, refraction, and diffraction properties of the spheres, following the approach of Uslenghi and Weston by making use of a modified Watson transformation.

Chapter 18 by Martha Alvarez Ramirez and Rodríguez José Antonio García, reviews the relations between Hamiltonian systems and symplectic geometry and uses these relations to reduce the system degrees of freedom, leading, in particular, to the solutions of the 2-body problem.

Chapter 19 by Anthony Ruffa, Michael Jandron, and Bourama Toni, presents an approach that can support a parallelized solution of banded linear systems without communication between processors using a scheme based on adjoining as many unknowns as the number of superdiagonals. The chapter also introduces p-adic computation, a step toward the development and implementation of a full parallel p-adic linear solver.

In Chap. 20 by Laban Rutto, Vitalis Temu, and Myong-Sook Ansari, the authors address the question of genetic loss and erosion of indigenous food cultures as

a preamble to making a case for investment in research on underutilized and alternative crops. Here is an area of research that could greatly benefit from an interdisciplinary approach to include mathematical and statistical models.

Chapter 21 by Mahlet Tadesse, Frederic Mortier, and Stefano Monni, reviews frequentist and Bayesian methods proposed to address in a unified manner the problems of cluster identification and cluster-specific variable selection in the context of mixture of regression models and also in the context of high-dimensional data analysis. Illustrations of these method performances are taken from ecology for modeling species-rich ecosystems and from genomic for integrating data from different genomic sources.

Chapter 22 by Loïc Teyssier addresses more precisely germs of parametric families of vector fields in the complex plane, with a saddle-node bifurcation and corresponding to first-order non linear differential equations.

Chapter 23 by Henryk Zoladek presents a new and corrected proof of the existence of 11 small amplitude limit cycles in a perturbation of some special cubic plane vector field with center.

The book as a whole certainly enhances the overall objective of the series, that is, to foster student interest and enthusiasm in the STEAM-H disciplines (Science, Technology, Engineering, Agriculture, Mathematics, and Health), stimulate graduate and undergraduate research, and generate collaboration among researchers on a genuine interdisciplinary basis.

The STEAM-H series is hosted at Virginia State University, Petersburg, Virginia, USA, an area that is socially, economically, intellectually very dynamic, and home to some of the most important research centers in the USA, including NASA Langley Research Center, manufacturing companies (Rolls-Royce, Canon, Chromalloy, Sandvik, Siemens, Sulzer Metco, NN Shipbuilding, Aerojet) and their academic consortium (CCAM), University of Virginia, Virginia Tech, the Virginia Logistics Research Center (CCAL), Virginia Nanotechnology Center, The Aerospace Corporation, C3I Research and Development Center, Defense Advanced Research Projects Agency, Naval Surface Warfare Center, Thomas Jefferson National Accelerator Facility, and the Homeland Security Institute.

The STEAM-H series, by now well established with a high impact through its intensive seminars and books published by Springer a world-renown publisher, is expected to become a national and international reference in interdisciplinary education and research.

Petersburg, VA, USA Bourama Toni

In Memoriam: Dr. Abderaouf Mourtada

Before the completion of this volume, we learned with great sadness and heavy heart the passing of Dr. Abderaouf Mourtada on April 13, 2015. He is survived by his wife Fatima, daughter Rhita, and son Ismail.

Co-author of Chap. 8, he was on the Faculty of the Université de Bourgogne, Dijon, France.

Dr. Mourtada has many important contributions to mathematics, in particular in the area of dynamical systems in relation to Hilbert 16th problem. His focus has been on the hyperbolic polycycles in an effort to develop a final proof for their finite cyclicity, achieving his goal during the last years. That is, Dr. Mourtada has been able to find a new strategy to solve the finite cyclicity for hyperbolic polycycles. This has been done by investigating the action of irreducible derivations on some Hilbert's quasi-regular algebras QRH of germs at 0, of local real analytic functions. He showed that these algebras are finite or locally finite. These new ideas represent a breakthrough in the field and will certainly provide a most needed new perspective on the cyclicity problems.

A highly innovative mathematician, Dr. Mourtada was also a humble and kind person. He will be deeply missed by all who came to know him and will remain forever in their hearts and minds !

May this book contribute to perpetuate his memory and his passion for mathematics !

Our heartfelt condolences to the family !

Acknowledgements

We would like to express our sincere appreciation to all the contributors and to all the anonymous referees for their professionalism. They all made this volume a reality for the greater benefice of the community of Science, Technology, Engineering, Agriculture, Mathematics, and Health.

Special thanks to Petro-Luciano Buono, Colin Christopher, and Yvan Saint-Aubin for their kind and professional assistance in formatting, correcting, and completing this volume. Special thanks indeed to Professor Yulij Ilyashenko for accepting, in addition to his chapter, to write such a wonderful Foreword and for his insightful guidance during the entire process.

It has been a great pleasure and a privilege to edit such a book !

Contents

Contributors

John A. Adam Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA, USA

Martha Alvarez-Ramirez Departamento de Matemáticas, UAM-Iztapalapa, México, D.F., Mexico

Myong-Sook Ansari Agricultural Research Station, Virginia State University, Petersburg, VA, USA

Faina Berezovskaya Department of Mathematics, Howard University, Washington, DC, USA

Terence R. Blows Department of Mathematics and Statistics, Northern Arizona University, Flagstaff, AZ, USA

Matthew Brown University of Southern California, Los Angeles, CA, USA

Pietro-Luciano Buono Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON, Canada

Magdalena Caubergh Departament de Matemàtiques, Edifici C, Facultat de Ciències, Universitat Autònoma de Barcelona, Barcelona, Spain

Colin Christopher School of Computing, Electronics and Mathematics, Faculty of Science and Engineering, Plymouth University, Plymouth, Devon, UK

Morgan Craig Faculté de Pharmacie, Université de Montréal, Montréal, QC, Canada

Léo van Damme Laboratoire Interdisciplinaire Carnot de Bourgogne, Université de Bourgogne, Dijon Cedex, France

Institute for Advanced Study, Technische Universität München, Garching, Germany

Bui Xuan Dieu Department of Mathematics, Dresden University of Technology, Dresden, Germany

School of Applied Mathematics and Informatics, Hanoi University of Science and Technology, Ha Noi, Vietnam

Luu Hoang Duc Department of Mathematics, Dresden University of Technology, Dresden, Germany

Institute of Mathematics, Vietnam Academy of Science and Technology, Ha Noi, Vietnam

Raluca Eftimie Division of Mathematics, University of Dundee, Dundee, UK

Mohamed El Morsalani Landesbank Baden-Württemberg, Stuttgart, Germany

Raymond R. Fletcher Department of Mathematics and Economics, Virginia State University, Petersburg, VA, USA

Rodríguez José Antonio García Departamento de Matemáticas, UAM-Iztapalapa, México, D.F., Mexico

Mario Gonzalez-Sales Faculté de Pharmacie, Université de Montréal, Montréal, QC, Canada

InVentiv Health Clinical, Montréal, QC, Canada

Lili Guadarrama CONACyT/CIMAT Aguascalientes, Aguascalientes AGS, Mexico

Wuria M.A. Hussein Department of Mathematics, College of Science, University of Salahaddin, Kurdistan Region, Erbil, Iraq

Yulij Ilyashenko National Research University Higher School of Economics, Moscow, Russia

Department of Mathematics, College of Arts and Sciences, Cornell University, Ithaca, NY, USA

Independent University of Moscow

Michael Jandron Naval Undersea Warfare Center, Newport, RI, USA

Albert Xin Jiang Trinity University, San Antonio, TX, USA

Debarun Kar University of Southern California, Los Angeles, CA, USA

Gérard Kientega UFR des Sciences Exactes et Appliquées, Université de Ouagadougou, Ouagadougou, Burkina Faso

Bernd Krauskopf Department of Mathematics, The University of Auckland, Auckland, New Zealand

Chengzhi Li School of Mathematical Sciences, Peking University, Beijing, China

Jun Li Faculté de Pharmacie, Université de Montréal, Montréal, QC, Canada

Pavao Mardešić, Institut de Mathématiques de Bourgogne, UMR CNRS, Dijon Cedex, France

Nguyen van Minh Department of Mathematics, Columbus State University, Columbus, GA, USA

Department of Mathematics & Statistics, University of Arkansas at Little Rock, Little Rock, AR, USA

Stefano Monni Department of Mathematics, Faculty of Arts and Sciences, American University of Beirut, Beirut, Lebanon

Frédéric Mortier UPR Biens et Services des Ecosystémes Forestiers Tropicaux (B&SEF), Campus International de Baillarguet, Montpellier Cedex, France

Abderaouf Mourtada[†] I.M.B., Université de Bourgogne, Dijon, France

Fahima Nekka Faculté de Pharmacie, Université de Montréal, Montréal, QC, Canada

Thanh Hong Nguyen University of Southern California, Los Angeles, CA, USA

Umaporn Nuntaplook Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand

Hinke M. Osinga Department of Mathematics, The University of Auckland, Auckland, New Zealand

Michael A. Pohrivchak The Naval Research Laboratory, Washington, DC, USA

Robert Roussarie I.M.B., Université de Bourgogne, Dijon, France

Anthony A. Ruffa Naval Undersea Warfare Center, Newport, RI, USA

Laban K. Rutto Agricultural Research Station, Virginia State University, Petersburg, VA, USA

Stefan Siegmund Department of Mathematics, Dresden University of Technology, Dresden, Germany

Arunesh Sinha University of Southern California, Los Angeles, CA, USA

Dominique Sugny Laboratoire Interdisciplinaire Carnot de Bourgogne, Université de Bourgogne, Dijon Cedex, France

Institute for Advanced Study, Technische Universität München, Garching, Germany

Mahlet G. Tadesse Department of Mathematics and Statistics, Georgetown University, Washington, DC, USA

[†] Author was deceased at the time of publication.

Milind Tambe University of Southern California, Los Angeles, CA, USA

Loïc Teyssier Université de Strasbourg, U.F.R. de Mathématiques et d'Informatique and Institut de Recherche Mathématique Avancée, Strasbourg Cedex, France

Vitalis W. Temu Agricultural Research Station, Virginia State University, Petersburg, VA, USA

Bourama Toni Department of Mathematics and Economics, Virginia State University, Petersburg, VA, USA

Zhaoxia Wang School of Mathematical Sciences, University of Electronic Science and Technology, Chengdu, Sichuan, P.R. China

Henryk Żołądek Institute of Mathematics, University of Warsaw, Warsaw, Poland

Chapter 1
Traveling Waves Impulses of FitzHugh Model with Diffusion and Cross-Diffusion

Faina Berezovskaya

Abstract The FitzHugh equations have been used as a caricature of the Hodgkin–Huxley equations of neuron firing and to capture, qualitatively, the general properties of an excitable membrane. The spatial propagation of neuron firing due to diffusion of the current potential was described by the FitzHugh–Nagumo model. Assuming that the spatial propagation of neuron firing is caused by not diffusion but cross-diffusion connection between the potential and recovery variables the cross-diffusion version of the FitzHugh model gives rise to the typical fast traveling wave solutions characteristic to the FitzHugh model, and additionally gives rise to the slow traveling wave solutions exhibited in the diffusion FitzHugh–Nagumo equations (Berezovskaya et al., Math Biosci Eng 5:239–260, 2008).

In this paper the FitzHugh model with both diffusion and cross-diffusion terms is studied; it is shown that this new version of spatial FitzHugh model gives rise to fast and slow traveling impulses.

Keywords FithHugh model • Slow and fast traveling wave solutions • Diffusion and cross-diffusion

1.1 Introduction

Hodgkin, Huxley, and Katz in the 1940s explored experimentally and mathematically the nature of nerve impulses. Their work revealed that the electrical pulses across the membrane arise from the uneven distribution between the intracellular fluid and the extracellular fluid of potassium (K+), sodium (Na+), and protein anions (see [1] for details). This entire process of rapid change in potential from threshold to peak reversal and then back to the resting potential level is called an *action potential*, impulse, or spike (see schematic diagram in Fig. 1.1).

The process was mathematically investigated by Hodgkin and Huxley in 1952 with a four-variable model [2]. FitzHugh in 1961 proposed a simplified two-variable

F. Berezovskaya (✉)
Department of Mathematics, Howard University, Washington, DC 20059, USA
e-mail: fberezovskaya@howard.edu

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_1

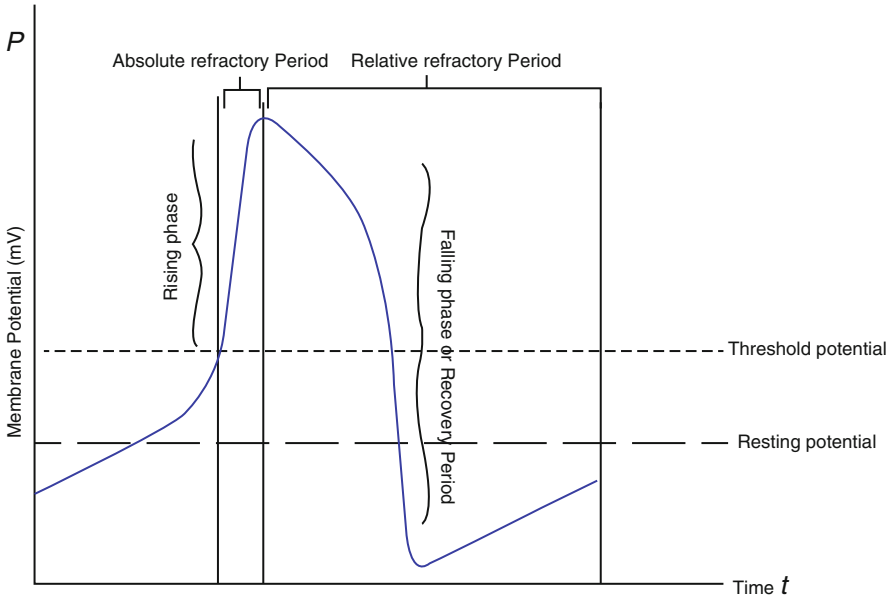


Fig. 1.1 Neuron spike in the (time-potential) plane obtained from experiments of neuron firing (see [1, 3] for details)

model of an excitable membrane, which made it possible to illustrate the various physiological states involved in an action potential (such as resting, active, refractory, enhanced, and depressed, see Fig. 1.1) in the phase plane (see [3, 4]). The FitzHugh system captures much of the same dynamical behavior and, in particular, demonstrates the spike-like behavior (see Sect. 1.2 and Fig. 1.2).

A more realistic model is the one that depends on both space and time since electric currents cross the membrane of the cell move along its axon lengthwise inside and outside. This mechanism makes it possible for electrical signals to be transmitted over long distance and thus propagate throughout the membrane without ever weakening or decreasing their initial strength. A mathematical model of the diffusion of current potential was first proposed and studied by FitzHugh in 1961 and 1969 (see [3]), Nagumo et al. in 1962 (see [5]), and many others. This model and its various modifications became one of the sources of many new methods of analysis of traveling waves, their stability, shapes and velocity of a propagation ([6–14], etc.).

Note that the initial FitzHugh and FitzHugh–Nagumo models "phenomenologically" (in the simplest way) described recovery process in the spike and spike-like wave propagation.

Recently models have been proposed, where the spatial solutions are conditioned by the effects of cross-diffusion "control" or "interactions" between components of the system ([6, 8, 10, 16–24], etc.). From a mathematical point of view

traveling wave solutions of cross-diffusion systems possess certain properties that essentially distinguish them from diffusion. Specifically, for some velocities of wave propagation they "repeat" structures of solutions of the model local system [14, 17, 18, 21]. This property of cross-diffusion systems may explain certain similarity of dynamics of local FitzHugh model and spatially distributed FitzHugh–Nagumo model.

Motivated by these works we modified the FitzHugh model to include a cross-diffusion connection between the potential and recovery variables. We *hypothesize* that the cross-diffusion regulation plays an important (perhaps, crucial) role in the spatial spreading of potential ([17, 25]). This version of the model provides an avenue both for investigating successful propagation of an excitable neuron and propagation failures, which are extremely important for many applications (see, for example, [26], and Sect. 1.4 below). In [17] we studied the "pure" cross-diffusion modification of FitzHugh model and investigated the characteristics of the spatial propagation of nerve impulses brought on by changes in the velocity of propagation and intensity of the cross-diffusion regulation. We showed that this model demonstrates two types of behaviors, the so-called slow and fast traveling waves. It was shown recently [8] that some biologically motivated modifications of the FitzHugh model demonstrate traveling waves even with three different velocities of their propagation.

The main goal of this paper is to show that at least two types of traveling waves of FHN-model are presented in generalization of FH-model by cross-diffusion.

The paper is organized as follows. Section 1.2 contains a brief description of local neuron dynamics within the framework of the FitzHugh model, as well as the bifurcation portrait of the model. Spatial dynamics of the FitzHugh model using its wave system, which provides an explanation of spatial regimes, e.g., "traveling waves," is given in Sect. 1.3.

Traveling wave solutions of the cross-diffusion modification of the FitzHugh model are described in Sect. 1.4. We show that for any fixed values of the cross-diffusion coefficient and other parameters of the model there exists the critical velocity C^* of wave propagation. The system behavior for $0<C<C^*$(slow waves) dramatically differs from the case when $C>C^*$(fast waves). In particular, the traveling spike with "large amplitude" appeared only for $C>C^*$. We present bifurcation diagrams for both slow and fast waves.

In Sect. 1.5 we consider traveling wave solutions of the FitzHugh model supplemented by diffusion and cross-diffusion terms. Using the techniques of the Tikhonov theorem ([27, 28], etc.) we show that some slow traveling wave solutions of the FitzHugh cross-diffusion model are preserved in the FitzHugh–Nagumo model. We also describe the computer experiments, which show that the fast traveling wave solutions of the FitzHugh cross-diffusion model look similar to those of the FitzHugh–Nagumo model.

Some details of analysis are given in the Appendix.

1.2 FitzHugh Model

The original FitzHugh model (1.1) describes the time dynamics of the neuron excitable membrane potential $P(t)$, which is responsible for the rising phase of neuron firing (see Fig. 1.1), and recovery membrane potential $Q(t)$, which is responsible for the falling phase of the action potential.

In slightly modified form [14] the FitzHugh model is presented as

$$\begin{aligned} eP_t &= -P^3 + P - Q \equiv F_1(P, Q), \\ Q_t &= k_1 P - Q - k_2 \equiv F_2(P, Q) \end{aligned} \tag{1.1}$$

where e, k_1, k_2 are parameters, reflecting intrinsic characteristics of the modeling system. The system has from one (a non-saddle, i.e., a node or a spiral or center) up to three (two non-saddles and a saddle) positive equilibria, $(P*, Q*)$ where $P*, Q*$ are common roots of $F_1(P, Q)$ and $F_2(P, Q)$. Thus, the model has domains of monostability and bistability. FitzHugh's computer analysis [3, 4] revealed that the model can also have limit cycles, namely, "small" cycles (containing a unique equilibrium inside) and "large" one (containing three equilibria inside). FitzHugh hypothesized that the "large separatrix loop" could be realized in the phase plane of system (1.1) for certain parameter values; the trajectory corresponding to this loop was considered as a model of a firing neuron.

The complete analysis of the FitzHugh model was done significantly later ([12, 14], etc.). It was proven in [11] that the principal dynamics of model (1.1) is described by the bifurcation diagram of the bifurcation "3-multiple neutral singular point with the degeneration, focus case," schematically presented in Fig. 1.2 (exact presentation of the bifurcations of high co-dimensions similar to those of our interest in this work was given in [13, 29–31]). In our model this bifurcation is realized in a vicinity of the parameter point $M(k_1 = 0, k_2 = 1, e = 1)$. The description of the bifurcation diagram is given by the following statement [11, 17].

Theorem 1 (i) *The space of parameters (k_1, k_2, e) can be subdivided into 21 domains of topologically different phase portraits of system (1.1). The cut of the complete parameter portrait to the plane (k_1, k_2) is topologically equivalent to the diagram presented in Fig. 1.2a (left) for arbitrary fixed $0 < e < 1$ and to the diagram presented in Fig. 1.2a (right) for arbitrary fixed $e > 1$.*

The parameter boundary surfaces correspond to the following bifurcations in system (1.1):

SN_1, SN_2: appearance/disappearance of a pair of equilibria on the phase plane;

H_1^-, H_2^- / H_1^+, H_2^+: change of stability of each of the non-saddle singular points in the Andronov–Hopf supercritical/subcritical bifurcation, respectively;

C: saddle-node bifurcation of a pair of limit cycles;

L_1, L_2: appearance/disappearance of a small limit cycle in one of two homoclinics of the saddle; and

R^+, R^-: appearance/disappearance of a large limit cycle in one of two homoclinics of the saddle.

Domains in Fig. 1.2 are numerated by integer numbers. Parametric portrait of system (1.1) possesses certain symmetry. The domains, which have respective symmetric properties, were numbered by integer with index $\boldsymbol{a}$, whereas their

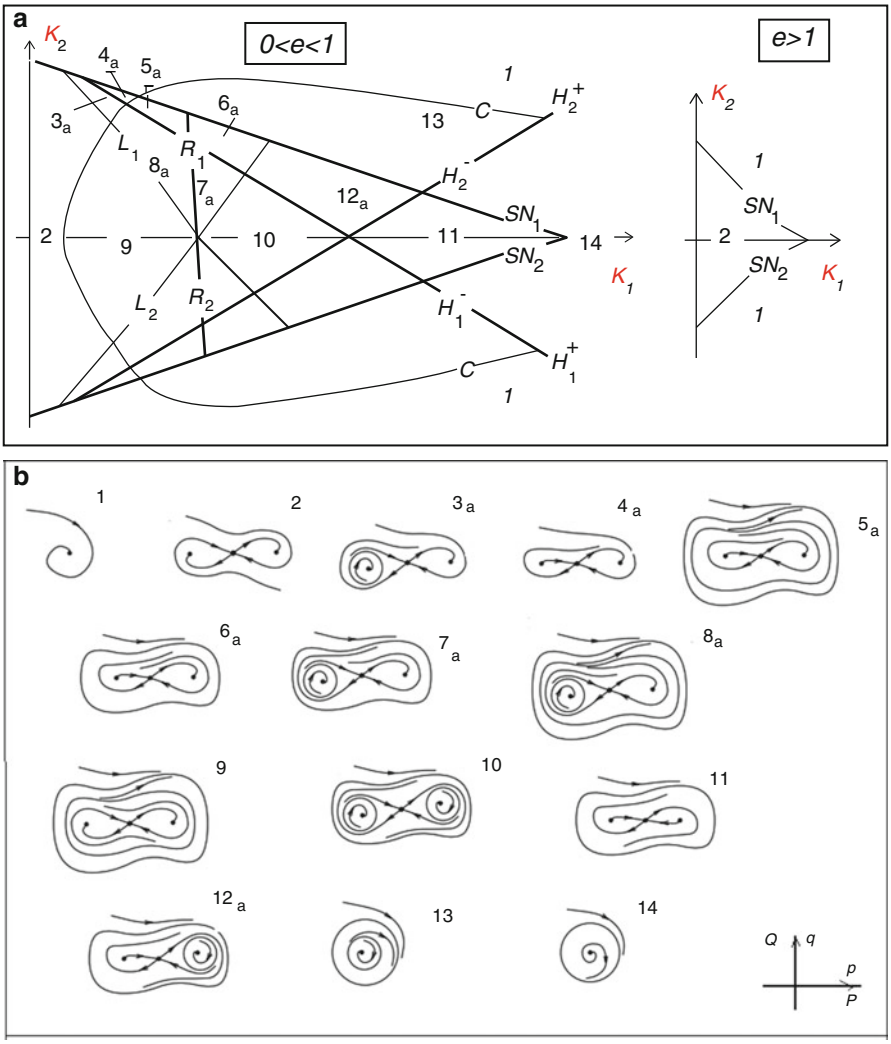


Fig. 1.2 Schematically presented (**a**) (k_1, k_2) − cuts of the bifurcation diagram of FitzHugh model (1.1) for fixed $0 < e < 1$ (*left*) and for $e > 1$ (*right*); (**b**) phase portraits. For any positive e the model has one stable topological node in domain 1 and three equilibria, two non-saddles, and a saddle, inside domain bounded by SN_1, SN_2; boundaries H_1^-, H_2^-/ H_1^+, H_2^+ correspond to the change of stability of each of the non-saddles in the Andronov–Hopf supercritical/subcritical bifurcation, respectively; each of these cycles disappears at homoclinics when the parameter values cross the boundaries L_1, L_2; two limit cycles appear in the phase plane when the parameter values cross the boundary C; the model has the large loop of the saddle separatrixes (a "large homoclinics") for parameter values on the boundaries R_1, R_2

symmetric counterparts have no index in the parameter portraits (Fig. 1.2a) and corresponding phase behaviors are not presented in Fig. 1.2b.

Let us emphasize that the spike-regime (see Fig. 1.1) can be P-component of the trajectory $\{P(t), Q(t)\}$ of the FitzHugh model; this trajectory corresponds to the phase curve of system (1.1), which is *the large separatrix loop* containing two equilibria inside (see the lower left panel in Fig. 1.3c, where coordinate ξ corresponds to t). The loop is realized with parameter values $(k_1 < 1, k_2, e < 1)$ belonging to the boundaries $\boldsymbol{R}_1$, $\boldsymbol{R}_2$ in the parameter portrait of the model in Fig. 1.2a. When parameter values crossing this boundary the limit cycle appears/disappears in the phase plane, its shape is "close" to the shape of the corresponding loop. Remark that the phase "8-shape" is realized at the parameter "point of intersection" of boundaries $\boldsymbol{L}_1, \boldsymbol{L}_2, \boldsymbol{R}_1$, and $\boldsymbol{R}_2$.

1.3 The Wave System for FitzHugh Model

1.3.1 FitzHugh Model with Diffusion and Cross-Diffusion

Spatial generalizations of FitzHugh model take into consideration diffusion processes and provide "spread" solutions in a space. Many works were devoted to the study of FHN dynamics, and in particular, to the investigation of "traveling wave" solutions ([4, 5, 7–9, 13, 15, 20, 23, 24], etc.). One of the most recent publications [8] (see also the references therein) describes different time-scale solutions of FHN-model and developed methods of computations that are related to the singular perturbation theoretical approach.

The generalized FitzHugh model, which takes into the consideration diffusion and cross-diffusion, is of the form

$$\begin{aligned} eP_t &= -P^3 + P - Q + D_P P_{xx} + D_Q Q_{xx} \equiv F_1(P,\ Q) + D_P P_{xx} + D_Q Q_{xx}, \\ Q_t &= k_1 P - Q - k_2 \equiv F_2(P,\ Q) \end{aligned} \tag{1.2}$$

where t is time, x is a one-dimensional space variable, and non-negative constants D_P, D_Q are the diffusion and cross-diffusion coefficients, respectively. For $D_P > 0, D_Q = 0$ we get FHN-model, and for $D_P = 0, D_Q \neq 0$ we get the cross-diffusion spatial modification of FH-model (1.1)

$$\begin{aligned} eP_t &= -P^3 + P - Q + D_Q Q_{xx}, \\ Q_t &= k_1 P - Q - k_2 \end{aligned} \tag{1.3}$$

which was investigated in [17].

1.3.2 Wave System of the Model

In what follows, we explore "traveling wave" solutions of system (1.2):

$$P(x,t) = P(x+Ct) \equiv p(\xi), \quad Q(x,t) = Q(x+Ct) \equiv q(\xi),$$

where $\xi = x + Ct$ and positive C is the velocity of the wave propagation. We get the ODE system:

$$\begin{aligned} eCp_\xi &= -p^3 + p - q + D_P p_{\xi\xi} + D_Q q_{\xi\xi} \equiv F_1(p,q) + D_P p_{\xi\xi} + D_Q q_{\xi\xi}, \\ Cq_\xi &= k_1 p - q - k_2 \equiv F_2(p,q) \end{aligned}$$

Differentiating the second equation by ξ, expressing $q_{\xi\xi}$ as

$$q_{\xi\xi} = k_1 p_\xi / C - q_\xi / C^2 = k_1 p_\xi / C - F_2(p,q)/C^2$$

and substituting it to the first equation we get finally that $(p(\xi),\ q(\xi))$ satisfy the "*wave system*":

$$\begin{aligned} p_\xi &= r \\ D_P r_\xi &= r\left(eC^2 - D_Q k_1\right)/C - F_1(p,q) + D_Q F_2(p,q)/C^2 \\ q_\xi &= \left(k_1 p - q - k_2\right)/C \equiv F_2(p,q)/C \end{aligned} \tag{1.4}$$

It is easy to verify that for $D_P = 0$ the wave system is two-dimensional:

$$\begin{aligned} p_\xi\left(eC^2 - D_Q k_1\right)/C &= F_1(p,q) - D_Q F_2(p,q)/C^2 \\ q_\xi &= F_2(p,q)/C \end{aligned} \tag{1.5}$$

Thus, the problem of describing all traveling wave solutions of system (1.2) and their rearrangements is reduced to the analysis of phase curves and bifurcations of solutions of three-dimensional wave system (1.4), which has the additional parameter C. Note that for $D_P = 0$ the wave system (1.5) is two-dimensional; this circumstance essentially simplifies the problem.

Remark 1 Mathematically, cross-diffusion equations possess some special properties, which facilitate their investigation [32, 33]; the most important one is that addition of the cross-diffusion term does not increase the dimensionality of corresponding wave system [6, 17–19, 21, 22].

1.3.3 *Traveling Waves of Reaction–Diffusion Model in the Frame of Wave System*

Between the bounded traveling wave solutions $(p(\xi), q(\xi))$ of the spatial model (1.2) and the phase curves of the wave system (1.4) there exists a known (see, for instance, [13, 15]) correspondence (Fig. 1.3a–c), which we formulate in the most important cases for p-component $p(\xi)$. The same statements are clearly valid for any component of the model.

Proposition 1 (Definitions)

(i) *A wave front $p(\xi)$ of model (1.2) corresponds to the heteroclinic orbit of wave system (1.5) such that for $\xi \to \pm\infty$ it tends to different in p singular points (Fig. 1.3a).*

(ii) *A wave train $p(\xi)$ of model (1.2) corresponds to the limit cycle of (1.5) (Fig. 1.3b).*

(iii) *A wave impulse $p(\xi)$ of model (1.2) corresponds to the homoclinic orbit of singular point of (1.5) (Fig. 1.3c).*

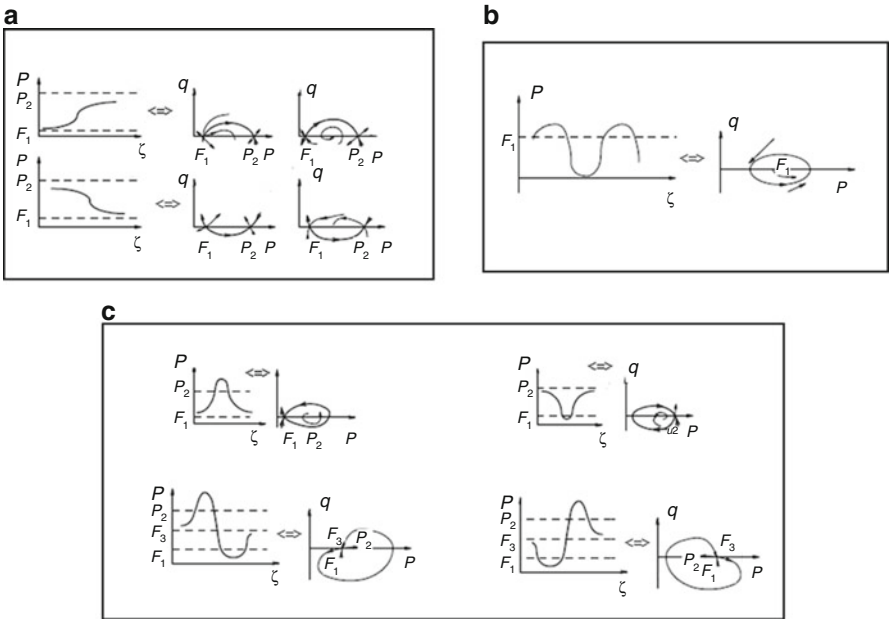


Fig. 1.3 Correspondence between the bounded "traveling wave" solutions of system (1.2) and the phase curves of its wave system (1.5). (**a**) The wave fronts correspond to the heteroclinic phase curves, a separatrix from a saddle to a node or to another saddle; (**b**) the wave train corresponds to the limit cycle; and (**c**) the wave impulses correspond to the homoclinic phase curves, *small* (*upper panel*) or *large* (*lower panel*) separatrix loops

By virtue of this statement, the description of all possible wave solutions of Eq. (1.2), as well as of their changing with variation of parameters of the reaction functions $F_1(p,q), F_2(p,q)$, is reduced to the analysis of phase curves and bifurcations in the wave system (1.5) depending on an additional parameter that is the propagation velocity C of waves. We will consider the behavior of system (1.4), (1.5) depending on variation of the parameters.

1.4 Traveling Waves of Cross-Diffusion Model

1.4.1 Behaviors of Wave System

If $eC^2 \neq D_Q k_1$, then (1.5) can be presented in the form

$$\begin{aligned} p_\xi &= \alpha \left(F_1(p,q) - D_Q F_2(p,q)/C^2\right) \\ q_\xi &= F_2(p,q)/C, \end{aligned} \tag{1.6$\pm$}$$

where $\alpha = C/\left(eC^2 - D_Q k_1\right)$; sign "+" in the denotation corresponds to the case $\alpha > 0$ and the system is denoted as (1.6+), sign "−" corresponds to the case $\alpha < 0$ and the system is denoted as (1.6−).

It was shown in [17] that the wave system exhibits different behaviors depending on sign of α.

Theorem 2

(i) *Let $C^2 e > D k_1$ (i.e., $\alpha > 0$). There exists a neighborhood of the parameter point $M\,(e = 1, k_1 = 0, k_2 = 1)$, in which the vector field defined by system (1.6+) has a bifurcation diagram, whose cut to the plane (k_1, k_2) is topologically equivalent to the one presented in Fig. 1.2. The boundaries in (e, k_1, k_2)—parameter space (lines at e—cuts at Fig. 1.2) correspond to the same bifurcations that have been mentioned in the Theorem 1.*

(ii) *Let $eC^2 < D k_1$ (i.e., $\alpha < 0$). There exists a neighborhood of the parameter point $M\,(e = 1, k_1 = 0, k_2 = 1)$, in which the vector field defined by system (1.6−) has a bifurcation diagram, whose cut to the plane (k_1, k_2) is topologically equivalent to the one presented in Fig. 1.4a for arbitrary fixed positive $0 < e < 1$ (left) and for arbitrary fixed $e > 1$ (right). The boundary surfaces in the parameter space correspond to the following bifurcations:*

SN_1, SN_2: appearance/disappearance of a pair of equilibria on the phase plane;

H: change of stability of the non-saddle equilibrium in Andronov–Hopf subcritical bifurcation;

L_1, L_2: appearance/disappearance of a small limit cycle in homoclinic bifurcations of the saddle;

SC_1, SC_2: upper and lower (respectively) heteroclinics of saddles.

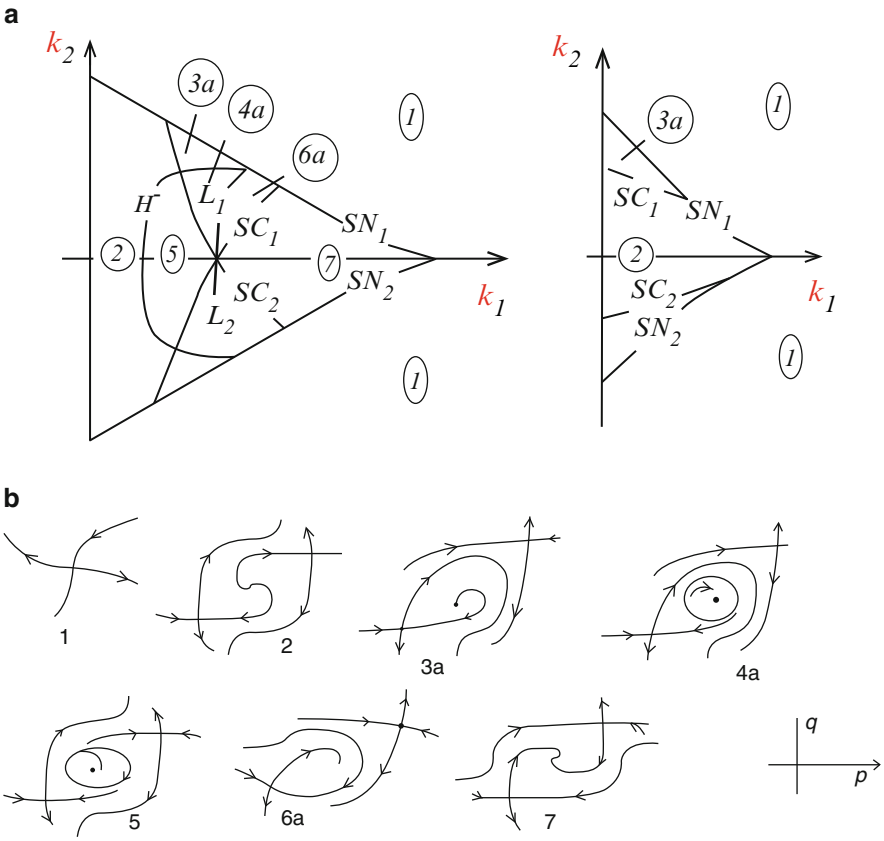


Fig. 1.4 Schematically presented (**a**) (k_1, k_2) − cuts of the bifurcation diagram of wave system (1.6−) of FitzHugh model (1.2) for fixed $0 < e < 1$ (*left*) and for $e > 1$ (*right*); (**b**) phase portraits of the system. Inside the domain bounded by SN_1, SN_2 the system has three equilibria, two saddles, and non-saddle; the boundaries SC_1, SC_2 correspond to the *right* and *left* heteroclinics of saddles; the boundary H corresponds to changing of stability of the non-saddle in Andronov–Hopf subcritical bifurcation; each of these cycles disappears at homoclinics (the boundaries L_1, L_2)

Remark 2 The bifurcation presented in Fig. 1.4 is known as "3-multiple neutral equilibrium with the degeneration, saddle case." In the wave systems (1.6−) of FitzHugh cross-diffusion model (1.3) the bifurcations are realized close to the parameter point M $(e = 1, k_1 = 0, k_2 = 1)$ (see also [31]).

1.4.2 Fast and Slow Wave Solutions of FitzHugh Model with Cross-Diffusion

According to Theorem 2 system (1.6) exhibits different behaviors depending on sign of $eC^2 - D_Q k_1 \neq 0$

Traveling wave solution of model (1.3) is called the s*low* wave if its velocity $0 < C < \sqrt{D_Q k_1/e}$ (i.e., $\alpha < 0$) and the *fast* wave if $C > \sqrt{D_Q k_1/e}$ (i.e., $\alpha > 0$).

Collecting together the statements of Theorems 1, 2, and Proposition 1 and taking into consideration that only positive values of model parameters k_1, k_2, e have biophysical meaning, we arrive at the following description of all possible wave solutions.

Theorem 3

(1) *Model (1.3) has the fast traveling wave solutions of the following types (see Fig. 1.2* and *Fig. 1.3):*

the fronts in every Domain of the portraits of Fig. 1.2 except the Domain 1, 13, and 14;

the single train in Domains 3a, 6a, 11, and 14; two trains, differing in their "amplitudes" in Domains 5a, 7a, 9, 12a, and 13; three different trains in Domains 8a and 10; and

the impulses on the boundaries L_1, L_2, and R_1.

(2) *Model (1.3) has the slow traveling wave solutions of the following types (see Fig. 1.4* and *Fig. 1.3):*

the fronts in every Domain of the portraits in Fig. 1.4a except the Domain 1; the monotonous fronts with the maximal "amplitude" on the boundary SC_1, SC_2;

the trains in the domains 4a, 5; and

the impulses with small amplitudes on the boundaries L_1, L_2.

The existence and the shapes of wave impulses, which may be different for slow and fast wave systems, is the problem of our main interest. Figures 1.5 and 1.6 demonstrate some typical phase portraits and solutions for slow wave system (1.6−) and fast wave system (1.6+) for different parameter values.

1.4.3 Possible Role of Cross-Diffusion Mechanism in Forming of Traveling Waves

According to Proposition 1, model (1.3) possesses a traveling impulse (spike) if and only if its wave system has a separatrix loop; the impulse has a large amplitude (see Figs. 1.1 and 1.3c) if the separatrix loop contains two points inside itself, and a small amplitude if the separatrix loop contains one point inside itself. Our analysis revealed that only fast wave system exhibits large separatrix loop, whereas slow wave system exhibits small separatrix loops only.

In the work [17] we utilized a modified version of the FitzHugh equations to model the spatial propagation of neuron firing; we assumed that this propagation is essentially caused by the cross-diffusion connection between the potential and recovery variables. This modification, which includes the implicit (although hypothetical) cross-diffusion mechanism, helped explore the effect of a generic drug in the neuron firing process, and explain other biophysical questions still arising [21, 25, 26].

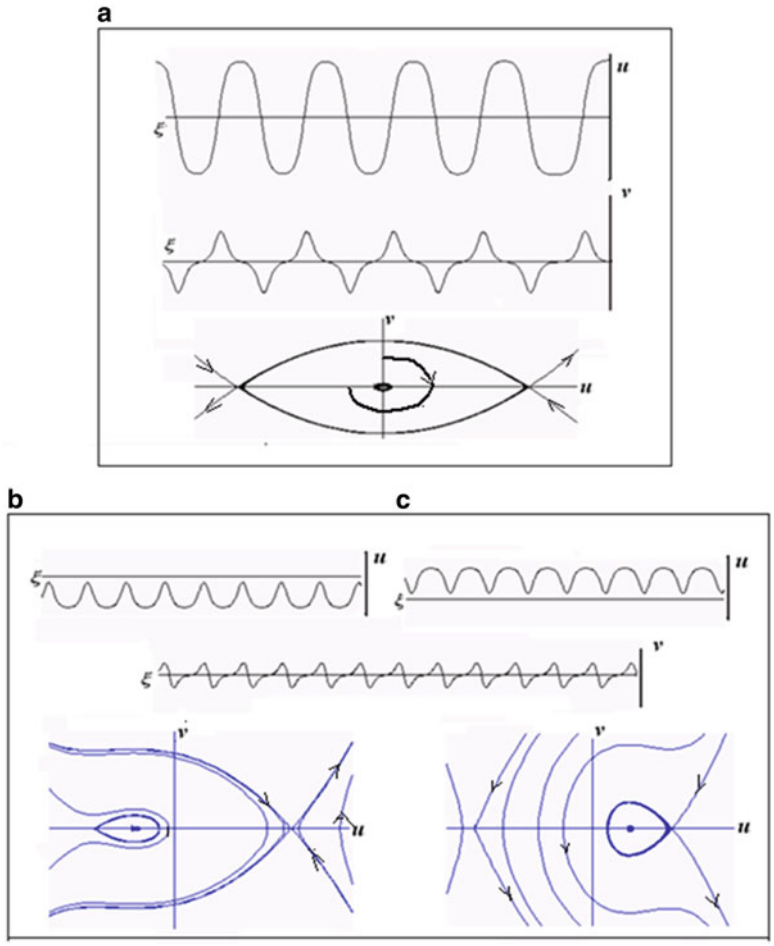


Fig. 1.5 $(u,\xi),(v,\xi)-$ solutions and $(u,v)-$ phase portraits of slow wave system (1.6−) for $e=.942, k_1=.9, D_Q=2, C=.1$ Here $u=q+k_2, v=k_1p-q-k_2$, where $p=p(\xi)$, $q=q(\xi)$ are components of system (1.6−). The system has three equilibria, the central equilibrium is a spiral, and the side equilibria are saddles. (**a**) $k_2=0$, stable equilibrium O is placed inside the unstable limit cycle that appeared from saddle heteroclinics; (**b**) $k_2=.01$, and (**c**) $k_2=-.01$. The system has "*left*" (**b**) and "*right*" (**c**) unstable limit cycle containing the stable equilibrium inside; the cycle appears from the saddle homoclinics loop, see the "*left*" (**b**) and "*right*" (**c**) panels

The mathematical problem of interest was the appearance and transformations of the traveling wave solutions, which depended on the model parameters, as those (e,k_1,k_2) which are "intrinsic" to the local system, the cross-diffusion coefficient D_Q, and the propagation speed C, that characterize the axons' abilities for the firing propagation. We studied the wave system of the cross-diffusion version of the model and explored its bifurcation diagram.

We have shown that the cross-diffusion model possesses a large set of traveling wave solutions; besides giving rise to the typical "fast" traveling wave solutions

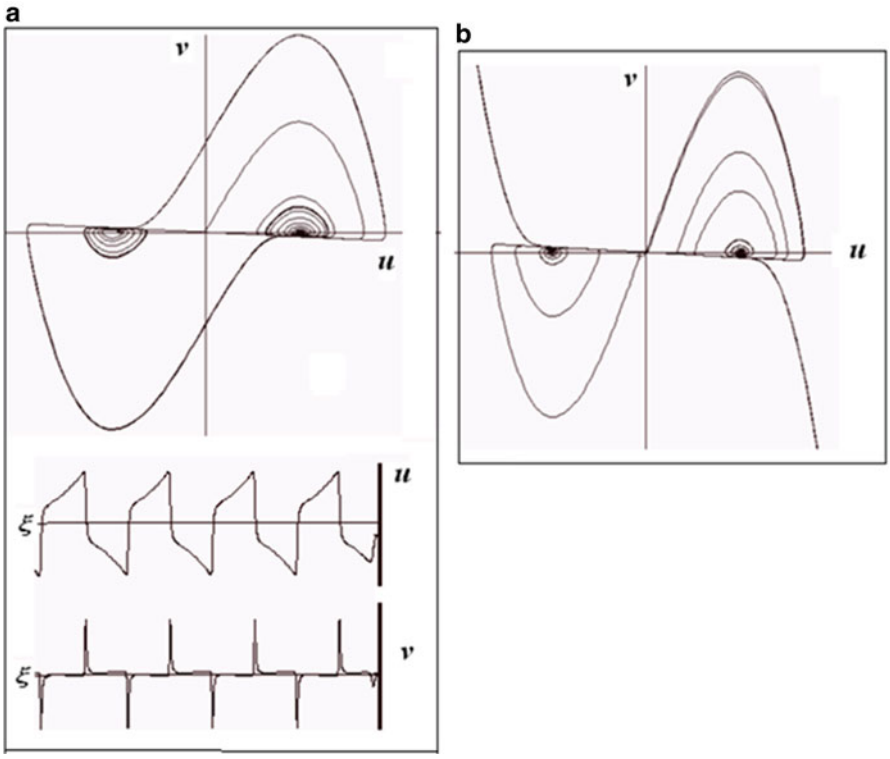


Fig. 1.6 (u, ξ), (v, ξ) − solutions and (u, v) − phase portraits of fast wave system (1.6+) for $e = .09, k_1 = .689, k_2 = 0, D_Q = .3$ Here $u = q + k_2, v = k_1 p - q - k_2$, where $p = p(\xi)$, $q = q(\xi)$ are variables of (1.6+). The system has three equilibria, a saddle, and two spirals. (**a**) $C = 4.1$. The system has a limit cycle which appears from the homoclinics of the saddle separatrices, the cycle contains inside two spiral equilibria; (**b**) $C = 3.1$, limit cycle is destroyed

exhibited in the original "diffusion" FitzHugh–Nagumo equations, it also gives rise to "slow" traveling wave solutions. This more sophisticated approach indicates that instead of a "one-parametric" set of waves ordered by the propagation speed C, one should consider a two-parametric set of traveling wave solutions with parameters (C, D_Q). We then proved that in the parametric space (C, D_Q) (under fixed parameter values e, k_1, k_2) there exists a parabolic boundary, $D_Q = KC^2$, where constant $K = e/k_1$ separates the domains of existence of the fast and slow waves. The system behavior qualitatively changes with the intersection of this boundary. Let us emphasize that the "traveling spike" that we consider as the "normal" propagation of a nerve impulse is a "fast" traveling wave. Hence, the parabola $D_Q = KC^2$ bounds the area where the "normal" spike propagation is possible. After the intersection of this boundary, due to a very large cross-diffusion coefficient or too small speed of impulse propagation, a "normal" propagation of the nerve impulse is impossible and some violations are inevitable: nerve impulses propagate with decreasing amplitude

or as damping oscillations. Introducing cross-diffusion regulations in the FitzHugh model allowed us to observe the propagation of spikes and spike-like oscillations but restricted their velocities from below or, equivalently, maintained the upper boundary for the cross-diffusion coefficient. It means that if, for any reason (e.g., as a result of the effect of a generic drug), the speed of transmission of a signal along the axon is reduced, then the "normal" neuron firing propagation in the form of a traveling spike is impossible. The increase of the cross-diffusion coefficient beyond the "normal" value implies the same result.

1.5 Traveling Wave Solution of FHN-Model

1.5.1 Slow Waves

Consider more precisely wave system (1.4):

$$\begin{aligned} p_\xi &= r \\ D_P r_\xi &= r\left(eC^2 - D_Q k_1\right)/C - F_1(p,q) + D_Q F_2(p,q)/C^2 \\ q_\xi &= \left(k_1 p - q - k_2\right)/C \equiv F_2(p,q)/C \end{aligned} \tag{1.7}$$

where $F_1(p,q) = \left(-p^3 + p - q\right)$, $F_2(p,q) = k_1 p - q - k_2$, $e > 0$, $k_1 > 0, k_2 \geq 0$.

Suppose that diffusion coefficient $D_P \to 0$. For the limiting system

$$\begin{aligned} p_\xi &= C\left(F_1(p,q) - D_Q F_2(p,q)/C^2\right)/\left(eC^2 - D_Q k_1\right) \\ q_\xi &= \left(k_1 p - q - k_2\right)/C \equiv F_2(p,q)/C \end{aligned} \tag{1.8}$$

is equivalent to the wave system (1.5) of cross-diffusion model (1.3).

Following the idea of the Tikhonov theorem [28] and its numerous generalizations (see, e.g., [27] and references therein) we prove numerically the following statement.

Statement 1 With parameter values in neighborhood of point $M\left(D_Q e = 1, k_1 = 1, k_2 = 0\right)$ and C, D_Q such that condition $\alpha \equiv C/\left(eC^2 - D_Q k_1\right) < 0$ holds there exist area $\Omega(p,q)$, $(0,0) \in \Omega$ in the phase plane (p,q) where the wave profiles $(p(\xi), q(\xi))$ of system (1.5) approximate two components $(p(\xi), q(\xi))$ of the wave profiles $(p(\xi), p_\xi(\xi), q(\xi))$ of system (1.4) for $(p_0 = p(\xi_0), q_0 = q(\xi_0)) \in \Omega$, $\xi \in (const, \infty)$, including homoclinics of equilibria.

According to this Statement, the slow wave solutions $(p(\xi), q(\xi))$ of model (1.1) with diffusion and cross-diffusion can be qualitatively approximated by the solutions of the model with only cross-diffusion term (compare Figs. 1.5 and 1.7).

Thus, the solution of the modification of the FitzHugh model, which accounts for the cross-diffusion and small diffusion terms, demonstrates the *slow traveling waves* (having relatively small velocity of propagation) similarly to the model with only cross-diffusion term under certain values of the model parameters.

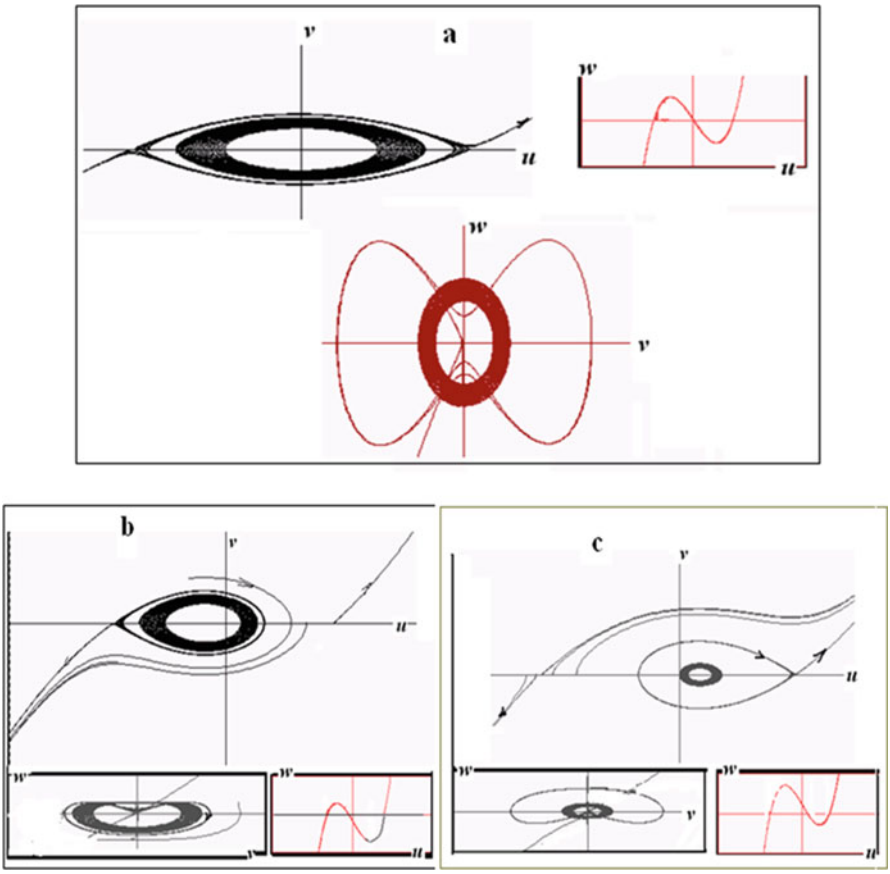


Fig. 1.7 $(u,v)-$, $(u,w)-,(w,v)-$ cuts of phase portraits of wave system (1.4). Here $u = q + k_2, v = k_1 p - q - k_2$, $w = v_\xi$, where $p = p(\xi)$, $q = q(\xi)$ are variables of system (1.4). The parameter values are $k_1 = .9, e = .9335, C = .1, D_P = .5, D_Q = 1.5$. (**a**) "Symmetric" case, k_2=0. The system demonstrates an unstable limit cycle that arose from "heteroclinics cycle" composed by separatrixes of the saddle points, the stable equilibrium O is placed inside this "funnel"(see (w,v)—cut); (**b**, **c**) $k_2 = \pm.005$. The system has "homoclinics" cycle containing stable point O; $k_2 = .005$ at the *left* and $k_2 = -.005$ at the *right panels*, correspondingly

1.5.2 Fast Waves

Numerous studies showed that FHN-model possesses spike type "fast" wave solutions (see, for example, [8] and reference therein).

Our computer experiments revealed that the fast solutions observed in the fast FH-cross-diffusion wave system (1.5+) have counterparts in the wave system (1.4) where $\overline{\alpha} \equiv eC^2 - D_Q k_1 > 0$ (compare Fig. 1.6 and Fig. 1.8a–c. The latter Figure demonstrates phase curves and trajectories of the model with different values of the

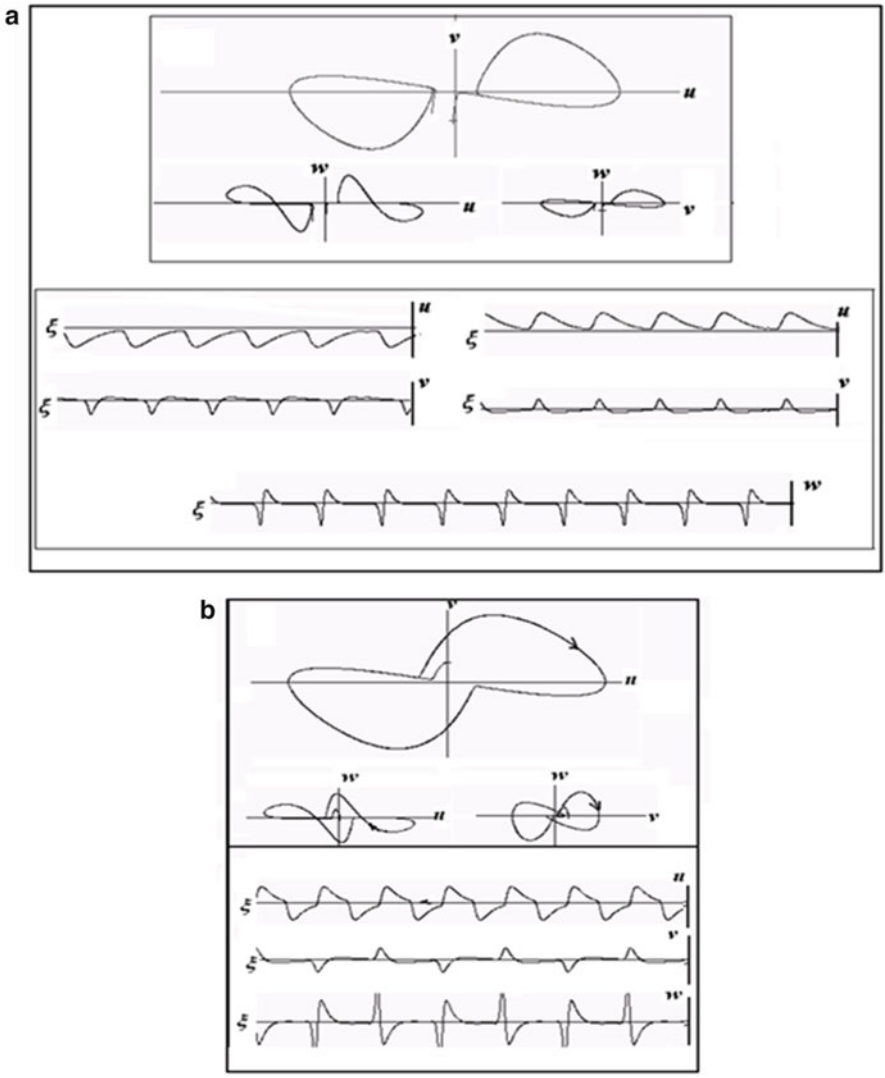


Fig. 1.8 $(u, v) -$, $(u, w) -$, $(w, v) -$ cuts of phase portraits and solutions $u(\xi)$, $v(\xi)$, $w(\xi)$ of wave system (1.4) in the case when $\overline{\alpha} = \left(eC^2 - D_Q k_1\right) > 0$. Here $u = q + k_2$, $v = k_1 p - q - k_2$, $w = v_\xi$, where $p = p(\xi)$, $q = q(\xi)$, $r = r(\xi)$ are variables of system (1.4). Parameters are $k_1 = .689, k_2 = 0, e = .15, D_P = .7, D_Q = .5$. (**a**) $C = 7$. The system has two limit cycles, whose shapes are close to the small homoclinic loop (see Fig. 1.3c). (**b**) $C = 4.8$. The system has a limit cycle, whose shape is close to the large homoclinic loop (this case is similar to the case that was presented in Fig. 1.6a); (**c**) $C = 2.5$. The limit cycle is destroyed

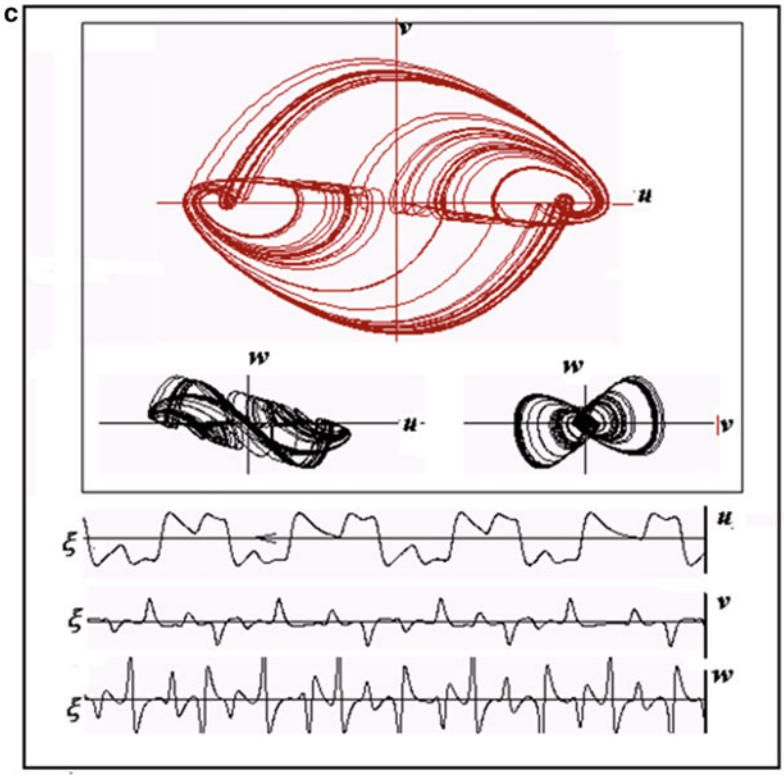


Fig. 1.8 (Continued)

speed propagation C but identical values of all other parameters). Notice that the qualitative behaviors of the model presented in these figures actually do not depend on cross-diffusion coefficient D_Q, in particular, for $D_Q = 0$.

Overall, our investigation of the diffusion—cross-diffusion modification of the FitzHugh model (the full model for brevity)—reveals an interesting phenomenon: if we consider the "fast" wave solutions, then the qualitative behavior of the full system coincides with that of the FH-model accounting for only diffusion term, but if we consider the "slow" wave solution, then the qualitative behavior of the full model coincides with that of FH-model accounting only for the cross-diffusion term. Both types of solutions qualitatively coincide with corresponding solutions of cross-diffusion wave system.

A.1 Appendix

A.1.1 Lienard Form of the FitzHugh Model and Its Wave System

Through the change of variables

$$(P,Q)\to(U,V):U=Q+k_2,\ V=F_2(P,Q)\equiv k_1P-Q-k_2,$$
$$(P=(U+V)/k_1,\ Q=U-k_2,\ k_1\neq 0) \tag{1.9}$$

the local model (1.1) is transformed to the generalized Lienard form:

$$U_t=V, eV_t=(U+V)/k_1-(U+V)^3/{k_1}^3$$
$$+k_2-U\equiv f(U)+V(g_1(U)+VG(U))\equiv\Phi(U,V), \tag{1.10}$$

Where

$$\begin{aligned} f(u)&=-u^3/k_1^2+u(1-k_1)+k_1k_2,\\ g_1(u)&=(1-e)-3u^2/k_1^2,\\ G(u,v)&=-(3u+v)/k_1^2 \end{aligned} \tag{1.11}$$

Model (1.2) after transformation (1.10) reads

$$\begin{aligned} &U_t=V,\\ &eV_t=\Phi(U,V)+D_PV_{xx}+\left(D_P+D_Qk_1\right)U_{xx} \end{aligned} \tag{1.12}$$

Model (1.3) after transformation (1.10) reads

$$\begin{aligned} &U_t=V,\\ &eV_t=\Phi(U,V)+D_Qk_1U_{xx} \end{aligned} \tag{1.13}$$

A *traveling wave solution* of systems (1.12) and (1.13) is defined as a pair of bounded functions

$$U(x,t)=U(x+Ct)\equiv u(\xi),\quad V(x,t)=V(x+Ct)\equiv v(\xi),$$

where $C>0$ is a velocity of propagation.

Let's now replace the capital letters in (1.9) with small letters, reduce p and q via $p=(u+v)/k_1,\ q=u-k_2,\ k_1\neq 0$.

Take into the consideration that $u_t=Cu_\xi,\ u_x=u_\xi;\ v_t=Cv_\xi,\ v_x=v_\xi;\ u_{xx}=u_{\xi\xi}=v_\xi/C$ and put $w=v_\xi, w_\xi=v_{\xi\xi}$ we get the wave system of system (1.12) in the form

$$\begin{aligned} u_\xi &= v/C \\ v_\xi &= w \\ D_P w_\xi &= (eC - (D_P + D_Q k_1)/C)w - \Phi(u, v) \end{aligned} \tag{1.14}$$

where $\Phi(U, V) = f(U) + V(g_1(U) + VG(U, V))$ and functions $f(u), g_1(u), G(u, v)$ are given by (1.11).

The wave system of (1.13) takes the form

$$\begin{aligned} &u_\xi = v/C, \\ &\left(eC - D_Q k_1\right)/C\Big)v_\xi = \Phi(u, v) \end{aligned} \tag{1.15}$$

System (1.15) contains the factor $1/\alpha \equiv \left(eC - D_Q k_1\right)/C\Big)$ which we assumed to be non-zero.

Behaviors of system (1.15) depend on the sign of parameter α. For $\alpha > 0$ there exists a parameter domain containing point $M(e = 1, k_1 = 0, k_2 = 1)$, where the vector field defined by system (1.15) is topologically orbitally equivalent to those defined by local system (1.1). It realizes the bifurcation of co-dimension 4 with symmetry ("spiral case") [11]. For $\alpha < 0$ parameter point $M(e = 1, k_1 = 0, k_2 = 1)$ is also the point that corresponds to the bifurcation of co-dimension four with symmetry but as a "saddle case." So, behaviors of system (1.15) for $\alpha > 0$ are different from those for $\alpha < 0$ (see details in [17]).

References

1. Sherwood, L.: Human Physiology: From Cells to Systems, 4th edn. Brooks and Cole, Belmont, CA (2001)
2. Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. **117**, 500–544 (1952)
3. FitzHugh, R.: Impulses and physiological states in theoretical models of nerve membrane. Biophys. J. **1**, 445–466 (1961)
4. FitzHugh, R.: Mathematical models of excitation and propagation in nerve. In: Schwan, H.P. (ed.) Biological Engineering. McGraw-Hill, New York (1969)
5. Nagumo, J., Arimoto, S., Yoshizawa, S.: An active pulse transmission line simulating nerve axon. Proc. IRE **50**, 2061–2070 (1962)
6. Berezovskaya, F., Karev, G.: Bifurcation approach to analysis of travelling waves in some Taxis–Cross-Diffusion models. Math. Model. Nath. Phenom. **8**(3), 133–153 (2013)
7. Evans, J., Fenichel, N., Feroe, J.: Double impulse solutions in nerve axon equations. SIAM J. Appl. Math. **42**, 219–234 (1982)
8. Guckenheimer, J., Krauskopf, B., Osinga, H.M., Sanstede, B.: Invariant manifolds and global bifurcations. CHAOS **25**, 097604 (2015)
9. Hastings, S.P.: On the existence of homoclinic and periodic orbits for the FitzHugh-Nagumo equations. Quart. J. Math. (Oxford) **27**, 123–134 (1976)
10. Ieda, M., Mimira, M., Ninomia, H.: Diffusion, cross-diffusion and competitive interaction. J. Math. Biol. **53**, 617–641 (2006)

11. Khibnik, A., Krauskopf, B., Rousseau, C.: Global study of a family of cubic Lienard equations. Nonlinearity **11**, 1505–1519 (1998)
12. Kostova, T., Ravindran, R., Schonbek, M.: FitzHugh-Nagumo revisited: types of bifurcations, periodical forcing and stability regions by a Lyapunov functional. Int. J. Bifurcation Chaos **14**, 913–925 (2004)
13. Kuznetsov, Y.: Elements of Applied Bifurcation Theory. Applied Mathematics Sciences. Springer, Berlin (1995)
14. Volokitin, E.P., Treskov, S.A.: Parameter portrait of FitzHugh system. Math. Model. **6**(12), 65–78 (1994) (in Russian)
15. Volpert, A.I., Volpert, V.A., Volpert, V.A.: Travelling Wave Solutions of Parabolic Systems. AMS, Providence, RI (1994)
16. Almirantis, Y., Papageorgiou, S.: Cross-diffusion effects on chemical and biological pattern formation. J. Theor. Biol. **151**, 289–311 (1991)
17. Berezovskaya, F., Camacho, E., Wirkus, S., Karev, G.: Traveling wave solutions of FitzHugh model with cross-diffusion. Math. Biosci. Eng. **5**(2), 239–260 (2008)
18. Berezovskaya, F.S., Karev, G.P.: Bifurcations of traveling waves in population models with taxis. Physics-Uspekhi **169**(9), 1011–1024 (1999)
19. Berezovskaya F., Novozhilov A., Karev G.: Pure cross-diffusion models: implications for traveling wave solutions. Dynam. Contin. Discret. Impulsive Syst. (Ser. A) **16**(S1), 141–146 (2009)
20. Biktashev, V., Tsyganov, M.: Solitary waves in excitable systems with cross-diffusion. Proc. R. Soc. A **461**, 3711–3730 (2005)
21. Kuznetsov, Y., Antonovsky, M., Biktashev, V., Aponina, E.: A cross-diffusion model of forest boundary dynamics. J. Math. Biol. **32**, 219–232 (1994)
22. Sherratt, J.: Travelling wave solutions of a mathematical model for tumor encapsulation. SIAM J. Appl. Math. **60**(2), 392–407 (2000)
23. Tsyganov, M.A., Biktashev, V.N., Brindley, J., Holden, A.V., Ivanitskii, G.R.: Waves in systems with cross-diffusion as a new class of nonlinear waves. Physics-Uspekhi **177**(3), 275–300 (2007)
24. Zemskov, E.P., Epstein, I.R., Muntean, A.: Oscillatory pulses in FitzHugh-Nagumo type systems with cross-diffusion. Math. Med. Biol. **28**(2), 217–226 (2011)
25. Verzi, W., Rheuben, M.B., Baer, S.M.: Impact of time-dependent changes in spine density and spine change on the input–output properties of a dendritic branch: a computational study. J. Neurophysiol. **93**, 2073–2089 (2005)
26. Highfield R.: The brains teasing chemical cocktail that gets us drunk. UK News, Electronic Telegraph, Issue 582 (1996)
27. Hoppensteadt, F.: Singular perturbations on the infinite interval. Trans. Amer. Math. Soc. **123**, 521–535 (1966)
28. Tikhonov, N.: Systems of differential equations containing small parameters multiplying the derivatives. Matematicheskii sbornik **31**, 575–586 (1952)
29. Dangelmayr, G., Guckenheimer, J.: On a four parameter family of planar vector fields. Arch. Ration. Mech. Anal. **97**, 321–352 (1987)
30. Dumortier, F., Rossarie, R., Sotomayor, J.: Bifurcations of planar vector fields. Lect. Notes Math. **1480**, 1–164 (1991)
31. Wiggins, S.: Introduction to Applied Non-Linear Dynamical Systems and Chaos. Springer, New York/Berlin/Heidelberg (1990)
32. Ni, W.-M.: Diffusion, cross-diffusion and their spike-layer steady states. Not. Am. Math. Soc. **45**(1), 9–18 (1998)
33. Murray, J.D.: Mathematical Biology. Springer, New York (1993)

Chapter 2
Local Limit Cycles of Degenerate Foci in Cubic Systems

Terence R. Blows

Abstract The problem of determining the stability of a weak focus in a quadratic or cubic system has been the focus of much research. Here we outline a simple but imperfect approach to the study of degenerate foci and use the method to give an example of a cubic system with four local limit cycles about a degenerate focus.

Keywords Cubic system • Limit cycles degenerate focus

2.1 Introduction

From his famous list of problems the second part of Hilbert's Sixteenth Problem was the topic of much interest in the 1980s and 1990s. The papers of Shi [1] and of Chen and Wang [2] which gave examples of quadratic systems with four limit cycles were a catalyst for this, but a major contributor to the increased work in this area was the rise of computer algebra systems that allowed lengthy algebraic manipulations to be carried out by a machine. With advances in bifurcation theory happening at the same time, see Rousseau [3], this was a rich period for research in planar polynomial systems.

A fixed point of a planar system of differential equations is called a *center* if a neighborhood of the fixed point is filled with closed orbits. Centers can occur in two ways; as well as the much studied case when the critical point is a *weak focus* (purely imaginary eigenvalues) there is the case where the critical point is a *degenerate focus*. The simplest type of the latter occurs under certain conditions when the linearization about the critical point is nilpotent but non-zero. These conditions are described with proof in Andronov et al. [4] and summarized in Perko [5].

Andronov's condition for monodromicity does not specify whether the fixed point is a center or focus, and in this sense the situation is similar to that of a weak focus. The problem of determining the stability of a weak focus has been well-studied and dates back to Poincaré. One approach is to construct a Liapunov

T.R. Blows (✉)
Department of Mathematics and Statistics, Northern Arizona University, Box 5717, Flagstaff, AZ 86011, USA
e-mail: Terence.Blows@nau.edu

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_2

function, and this can be done using an algorithm which is easily implemented using symbolic computing. See, for example, Blows and Lloyd [6]. Here we use a similar approach and use a Liapunov function to determine the stability of a degenerate focus. The method described is imperfect—it does not determine the stability for every degenerate focus—but we are able to use the method with some success. In particular we describe and apply the method to degenerate foci of cubic systems and extend a result of Andreev et al. [7].

It should be noted that another possibility is that the localization about the critical point has no linear terms. An example of this was studied in Blows and Rousseau [8] where the localization was about the point at infinity of a cubic system of a certain type.

2.2 Method

We consider cubic systems that have a degenerate focus at the origin. These may be written such that the linear part has a canonical form corresponding to a Jordan block with double zero eigenvalue:

$$\begin{cases} x' = y + P_2(x,y) + P_3(x,y) \\ y' = Q_2(x,y) + Q_3(x,y) \end{cases} \tag{2.1}$$

Also for monodromicity it is necessary that ([4, 5]) $Q_2(x, 0) = 0$ and $Q_3(x, 0) < 0$.

To study the stability of the origin we seek to construct a Liapunov function of the form

$$V(x,y) = V_2(x,y) + V_3(x,y) + V_4(x,y) + \cdots + V_n(x,y) + \cdots$$

where $V_k(x, y)$ is homogeneous of degree k. This gives

$$V' = \frac{\partial V_2}{\partial x} y + \cdots$$

and to be one-signed we therefore need $V_2(x, 0) = 0$. For V itself to be positive in a neighborhood of the origin it is therefore necessary that $V_2(x, y) = cy^2$ for some $c > 0$ and we make the arbitrary and convenient choice $c = ½$ to get

$$V_2 = \frac{1}{2} y^2 \tag{2.2}$$

We have

$$V' = \left(\frac{\partial V_2}{\partial x} + \frac{\partial V_3}{\partial x} + \frac{\partial V_4}{\partial x} + \cdots \right) (y + P_2(x,y) + P_3(x,y)) + \left(\frac{\partial V_2}{\partial y} + \frac{\partial V_3}{\partial y} + \frac{\partial V_4}{\partial y} + \cdots \right) (Q_2(x,y) + Q_3(x,y)),$$

and gathering like terms gives

$$\begin{aligned} V' = & \left(\tfrac{\partial V_2}{\partial x}\right) y \\ & + \left(\tfrac{\partial V_3}{\partial x}\right) y + \left(\tfrac{\partial V_2}{\partial x}\right) P_2(x,y) + \left(\tfrac{\partial V_2}{\partial y}\right) Q_2(x,y) \\ & + \left(\tfrac{\partial V_4}{\partial x}\right) y + \left(\tfrac{\partial V_3}{\partial x}\right) P_2(x,y) + \left(\tfrac{\partial V_3}{\partial y}\right) Q_2(x,y) + \left(\tfrac{\partial V_2}{\partial x}\right) P_3(x,y) \\ & + \left(\tfrac{\partial V_2}{\partial y}\right) Q_3(x,y) \\ & + \left(\tfrac{\partial V_5}{\partial x}\right) y + \left(\tfrac{\partial V_4}{\partial x}\right) P_2(x,y) + \left(\tfrac{\partial V_4}{\partial y}\right) Q_2(x,y) + \left(\tfrac{\partial V_3}{\partial x}\right) P_3(x,y) \\ & + \left(\tfrac{\partial V_3}{\partial y}\right) Q_3(x,y) + \cdots \end{aligned}$$

In order to guarantee that the quadratic and cubic terms of V' are both zero, the choice (2.2) then implies that

$$V_3(x,y) = -\int Q_2(x,y)\,dx$$

Indeed we have $V' \equiv \mathbf{0}$ if we can recursively choose V_k such that

$$\begin{aligned} \left(\tfrac{\partial V_k}{\partial x}\right) y = & -\left(\tfrac{\partial V_{k-1}}{\partial x}\right) P_2(x,y) - \left(\tfrac{\partial V_{k-1}}{\partial y}\right) Q_2(x,y) - \left(\tfrac{\partial V_{k-2}}{\partial x}\right) P_3(x,y) \\ & - \left(\tfrac{\partial V_{k-2}}{\partial y}\right) Q_3(x,y) \end{aligned}$$

for all integers $k \geq 4$. However the term on the right-hand side may contain terms of the form x^{k+1} and so the best we can do when choosing the V_k is to have

$$V' = \eta_5 x^5 + \eta_6 x^6 + \eta_7 x^7 + \cdots + \eta_k x^k + \cdots$$

If the leading non-zero η_k is such that k is even, then V' is one-signed in a neighborhood of the origin, and the stability of the origin is determined by the sign of η_k. If all η_k terms are zero, then the origin is a center. However if the leading non-zero η_k is such that k is odd, then the construction fails to give a Liapunov function. Such cases will require a different method. See, for example, Sadovskii [9].

The center problem parallels the case of a weak focus. Although there are an infinite number of η_k, this set has a finite basis which we denote $<L(1), L(2), L(3) \ldots L(N)>$ where the Liapunov numbers $L(k)$ are numbered in order as they arise from the η_k. Calculating the η_k and reducing them to a finite set of Liapunov numbers is a difficult problem, and it is likely, as with the case of a weak focus, that the full solution to the problem may lie out of reach even with fast computers and Gröbner basis methods.

Another connection with weak foci lies in the generation of small amplitude limit cycles by perturbation methods. This is described below in the proof of Theorem 2.

2.3 Results

Using a judicious linear coordinate change, we may assume without loss of generality that $\boldsymbol{P_3}(\boldsymbol{x},\ \boldsymbol{0}) = \boldsymbol{0}$ and $\boldsymbol{Q_3}(\boldsymbol{x},\ \boldsymbol{0}) = -\boldsymbol{1}$ in (2.1). We therefore consider systems of the form

$$\begin{cases} \boldsymbol{x'} = \boldsymbol{y} + \boldsymbol{Cx^2} + \boldsymbol{Dxy} + \boldsymbol{Fy^2} + \boldsymbol{Nx^2y} + \boldsymbol{Qxy^2} + \boldsymbol{Ry^3} \\ \boldsymbol{y'} = \boldsymbol{Axy} + \boldsymbol{By^2} - \boldsymbol{x^3} + \boldsymbol{Kx^2y} + \boldsymbol{Lxy^2} + \boldsymbol{My^3} \end{cases}$$

Applying the algorithm we find that

$$\begin{aligned} \eta_5 &= 1/2(A + 2C)(AC + 1) \\ \eta_6 &= -\left(5\mathrm{AB} + 14\mathrm{BC} + 5\mathrm{A^2BC} + 17\mathrm{ABC^2} + 6\mathrm{BC^3} + 2\mathrm{AD} - 6\mathrm{CD} + 5\mathrm{A^2CD}\right. \\ &\quad \left. +12\mathrm{AC^2D} + 2\mathrm{K} - \mathrm{ACK} + 6\mathrm{C^2K}\right)/6 \end{aligned}$$

However the solution to $\eta_5 = \eta_6 = 0$ is far from simple, and we are already faced with computational difficulties that we do not wish to get into here. Instead we make the convenient choice $A = C = 0$ to easily get $\eta_5 = 0$. It is easy to see that then $\eta_6 = -K/3$. Under the conditions $A = C = K = 0$ we find using *Mathematica* 8 that

$$\begin{aligned} \eta_7 &= \mathrm{F}/4 \\ \eta_8 &= (-19\mathrm{BF} + 8\mathrm{DF} - 12\mathrm{M} - 4\mathrm{Q})/20 \\ \eta_9 &= \left(209\mathrm{B^2F} - 60\mathrm{D^2F} + 55\mathrm{FL} + 90\mathrm{DM} + 40\mathrm{FN} + 207\mathrm{BDF}\right. \\ &\quad \left. +42\mathrm{BM} - 66\mathrm{BQ} + 40\mathrm{DQ}\right)/120 \\ \eta_{10} &= \left(-1509\mathrm{B^3F} + 480\mathrm{D^3F} - 818\mathrm{DFL} - 720\mathrm{DFN} - 660\mathrm{D^2M} - 360\mathrm{D^2Q}\right. \\ &\quad - 3261\mathrm{B^2DF} + 18\mathrm{B^2M} + 726\mathrm{B^2Q} - 1982\mathrm{BD^2F} + 439\mathrm{BFL} - 22\mathrm{BDM} \\ &\quad \left. + 1090\mathrm{BFN} + 956\mathrm{BDQ} + 552\mathrm{LM} + 480\mathrm{MN} + 264\mathrm{LQ} + 240\mathrm{NQ}\right)/840 \end{aligned}$$

In terms of Liapunov quantities, where $\mathrm{L}(1) = \eta_5$ and $\mathrm{L}(2) = -\mathrm{K}/3$ have been set to zero, we have

$$\begin{aligned} \mathrm{L}(3) &= \mathrm{F}/4 \\ \mathrm{L}(4) &= (3\mathrm{M} + \mathrm{Q})/5 \\ \mathrm{L}(5) &= (\mathrm{D} + 2\mathrm{B})\,\mathrm{M}/4 \end{aligned}$$

We have a choice from L(5): Either $D+2B = 0$ or $M = 0$. However, as we show in the proof of Theorem 1, the latter gives a center. So we assume M is non-zero. We make the choices $F = 0$, $Q = -3M$, and $D = -2B$ to get

$$\eta_{10} = 2\mathrm{M}\,(\mathrm{L} + \mathrm{N})\,/7$$
$$\eta_{11} = 3\mathrm{BM}\left(2300\mathrm{B}^2 + 216\mathrm{L} + 181\mathrm{N}\right)/112$$

Substituting $\mathrm{N} = -\mathrm{L}$ from η_{10} gives

$$\mathrm{L}(7) = 0$$

If B or M is equal to zero, then, as we show in the proof of Theorem 1, we have a center. Otherwise

$$\eta_{12} = -\mathrm{M}\big(4436580\mathrm{B}^4 + 387976\mathrm{B}^2\mathrm{L} - 1504\mathrm{L}^2 + 263331\mathrm{B}^2\mathrm{N} - 3968\mathrm{LN} - 2464\mathrm{N}^2\big)/5040$$

And subbing $L = -14B^2$ and $N = 4B^2$ gives

$$\mathrm{L}(8) = (108161/560)\ \mathrm{B}^4\mathrm{M}$$

So $\eta_5 = \eta_6 = \eta_7 = \eta_8 = \eta_9 = \eta_{10} = \eta_{11} = \eta_{12} = 0$ implies a center.

Theorem 1 *The origin of the system*

$$\begin{cases} \boldsymbol{x' = y + Dxy + Fy^2 + Nx^2y + Qxy^2 + Ry^3} \\ \boldsymbol{y' = By^2 - x^3 + Kx^2y + Lxy^2 + My^3} \end{cases}$$

is a center if and only if one of the following two conditions holds:

$$1)\ K = F = M = Q = 0$$
$$2)\ K = F = B = D = 0; \quad Q = -3M, N = -L$$

Proof Necessity has already been shown. For the sufficiency of 1) note that in this case the origin is a center due to the symmetry $(x, y, t) \to (x, -y, -t)$. Condition 2) gives a Hamiltonian system.

In the following theorem we start with the weakest possible degenerate focus, namely when $\eta_5 = \eta_6 = \eta_7 = \eta_8 = \eta_9 = \eta_{10} = \eta_{11} = 0$ but $\eta_{12} \neq 0$, we may perturb η_{10}, η_8, and η_6 away from zero in turn to get three local limit cycles in the same manner as using multiple Hopf bifurcation from a weak focus. We then perturb λ non-zero to get a fourth in a manner that is new. It is possible that other perturbations will produce more than one local limit cycle; this would require a complete analysis of the unfolding of the degenerate critical point in a manner similar to that of Rousseau and Zhu [10] for an elliptic nilpotent singularity in quadratic systems.

Theorem 2 *The system*

$$\begin{cases} \boldsymbol{x' = y - 2Bxy + Nx^2y + (\delta - 3M)\, xy^2 + Ry^3} \\ \boldsymbol{y' = \lambda\, (-x + sgn\, (M)\, y) + By^2 - x^3 + \mu x^2y + Lxy^2 + My^3} \end{cases}$$

where $M \neq 0, B \neq\ 0,\ \delta M < 0, \mu M > 0,\ O << |\lambda| << |\mu| << |\delta| << |\varepsilon| << 1$ *has at least three local limit cycles in a neighborhood of the origin.*

Proof With $\lambda = \mu = \delta = 0$, the origin is a degenerate focus whose stability is given by the sign of M. The perturbations of δ, and μ away from zero in turn each cause a change in stability and produce local limit cycles. At this point, the origin has stability given by the sign opposite to M. Finally perturbing $\lambda\ \neq\ 0$ produces a strong focus at the origin whose stability is given by the sign of M to produce one final local limit cycle.

Appendix: Mathematica 8

Mathematica was used interactively to produce the results in Sect. 2.3. Firstly the base functions are put in place:

$$\mathrm{P2} = cx\hat{}2 + Dxy + F\hat{}2$$

$$\mathrm{Q2} = Axy + By\hat{}2$$

$$\mathrm{P3} = Nx\hat{}2y + Qxy\hat{}2 + Ry\hat{}3$$

$$\mathrm{Q3} = -x\hat{}3 + Kx\hat{}2y + Lxy\hat{}2 + My\hat{}3$$

$$\mathrm{V2} = 1/2y\hat{}2$$

$$\mathrm{V3} = -\mathrm{Integrate}\,[\mathrm{Q2}, x]$$

After this each iteration of the algorithm has a sequence of similar steps. The first set is as follows:

$$T4 = -D\,[\mathrm{V3}, x]\,\mathrm{P2} - D\,[\mathrm{V3}, y]\,\mathrm{Q2} - D\,[\mathrm{V2}, x]\,\mathrm{P3} - D\,[\mathrm{V2}, y]\,\mathrm{Q3}$$

$$\mathrm{Collect}\,[\%, \{x, y\}]$$

$$\mathrm{X4} = \mathrm{Coefficient}\,[\%, x\hat{}4]$$

$$\mathrm{V5} = \mathrm{Simplify}\,[(\mathrm{T4} - \mathrm{X4}\, x\hat{}4)\,/y]$$

Each X terms give us a focal value η, and the V terms give us the homogeneous pieces of the Liapunov function that we are constructing. We continue through as many of these steps as is necessary.

References

1. Shi, S.-l.: A concrete example of the existence of four limit cycles for plane quadratic systems. Sci. Sin. 153–158 (1980)
2. Chen, L.S., Wang, M.S.: The relative position and number of limit cycles of a quadratic differential system. Acta Math. Sin. **22**, 751–758 (1979)
3. Rousseau, C.: Bifurcation methods in polynomial systems. In: Proceedings of the Nato Advanced Study Institute (Séminaire de Mathématiques Supérieures), "Bifurcations and periodic orbits of vector fields", pp. 383–428. Kluwer Academic Publishers (1992)
4. Andronov, A.A., Leontovich, E.A., Gordon, I.I., Meier, A.G.: Qualitative Theory of Second-order Dynamical Systems. Wiley, New York (1973)
5. Perko, L.M.: Differential Equations and Dynamical Systems. Springer, New York (1991)
6. Blows, T.R., Lloyd, N.G.: The number of limit cycles of certain polynomial differential equations. Proc. R. Soc. Edinburgh Sect. A **98**, 215–239 (1984)
7. Andreev, A.F., Sadovskii, A.P., Tsikalyuk, V.A.: The center-focus problem for a system with homogeneous nonlinearities in the case of zero eigenvalues of the linear part. Differ. Equ. **39**, 155–164 (2003)
8. Blows, T.R., Rousseau, C.: Bifurcations at infinity in polynomial vector fields. J. Differ. Equ. **104**, 215–242 (1993)
9. Sadovskii, A.P.: Solution of the center-focus problem for some systems of nonlinear oscillations. Diff. Uravn. **14**, 268–269 (1978)
10. Rousseau, C., Zhu, H.: PP-graphics with a nilpotent elliptic singularity in quadratic systems and Hilbert's 16th problem. J. Differ. Equ. **196**, 169–208 (2004)

Chapter 3
Lyapunov–Schmidt and Centre Manifold Reduction Methods for Nonlocal PDEs Modelling Animal Aggregations

Pietro-Luciano Buono and R. Eftimie

Abstract The goal of this paper is to establish the applicability of the Lyapunov–Schmidt reduction and the Centre Manifold Theorem (CMT) for a class of hyperbolic partial differential equation models with nonlocal interaction terms describing the aggregation dynamics of animals/cells in a one-dimensional domain with periodic boundary conditions. We show the Fredholm property for the linear operator obtained at a steady-state and from this establish the validity of Lyapunov–Schmidt reduction for steady-state bifurcations, Hopf bifurcations and mode interactions of steady-state and Hopf. Next, we show that the hypotheses of the CMT of Vanderbauwhede and Iooss (Center manifold theory in infinite dimensions. In: Jones, C., Kirchgraber, U., Walther, H.O. (eds.) Dynamics Reported, vol. 1, pp. 125–163. Springer, Berlin, 1992) hold for any type of local bifurcation near steady-state solutions with $\mathbf{SO}(2)$ and $\mathbf{O(2)}$ symmetry. To put our results in context, we review applications of hyperbolic partial differential equation models in physics and in biology. Moreover, we also survey recent results on Fredholm properties and Centre Manifold reduction for hyperbolic partial differential equations and equations with nonlocal terms.

Keywords Hyperbolic PDE • Animal aggregation • Centre Manifold reduction • Lyapunov–Schmidt reduction • Symmetry Fredholm property

3.1 Introduction

Collective self-organised behaviour is a phenomenon observed in a variety of organisms. Familiar examples include schooling fish, flocking birds, swarming insects, aggregating bacteria, etc. The way such aggregations are formed, maintained and

P.-L. Buono (✉)
Faculty of Science, University of Ontario Institute of Technology, Oshawa, ON, Canada L1H 7K4
e-mail: Luciano.buono@uoit.ca

R. Eftimie
Division of Mathematics, University of Dundee, Dundee DD1 4HN, UK

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_3

the transitions between different patterns is a fascinating subject which has been increasingly studied in the last 25 years. The elegance and beauty of motion in aggregations are remarked in early writings, going back to antiquity, in Pliny the Elder's book *The Natural History* [59], where collective movement in flocks of birds is described. Apart from the appeal this problem has for curiosity-driven research, understanding of collective self-organised motion also has applications to environmental and societal problems such as the formation and motion of swarms of locusts [65], which affect rural communities in several locations worldwide, as well as applications to the small scale phenomenon of cell–cell interactions in developmental biology or cancer research.

Mathematical modelling has been an important aspect of the study of collective motion and aggregation using both particle-based and density-based models. Those models have been used to probe the possible biological mechanisms leading to the formation and persistence of aggregations, and also as a means to investigate transient aggregations. Several modelling approaches assume local interactions with conspecifics [9]. However, in many cases it is preferable for the models to assume that animals/cells can interact with conspecifics positioned further away [29, 57, 58, 64, 66]. For instance, in migratory flocks of birds, radar-tracking observations have shown that individuals 200–300 m apart can fly at the same speed and in the same direction [49]. Nonlocal interactions are also seen in developmental biology where collective cell movement results from cell–cell adhesion forces with an interaction range proportional to cell size [1].

In this study, we focus on density-based models and consider a class of 1D hyperbolic first-order partial differential equations with nonlocal terms:

$$\frac{\partial u^+}{\partial t} + \frac{\partial g^+[u^+,u^-]u^+}{\partial x} = f^+[u^+,u^-], \tag{3.1a}$$

$$\frac{\partial u^-}{\partial t} + \frac{\partial g^-[u^+,u^-]u^-}{\partial x} = f^-[u^+,u^-]. \tag{3.1b}$$

Such equations could be used, for instance, to model the dynamics of animal aggregations in 1D (i.e., on domains much longer than wide) [18], and in this case $u^\pm$ describe the densities of left-moving $(-)$ and right-moving $(+)$ animals, $g^\pm$ are the (possibly nonlocal) density-dependent speeds, and $f^\pm$ incorporate (possibly nonlocal) turning behaviour and population dynamics. Moreover, such equations are known to exhibit a large variety of spatio-temporal patterns, ranging from stationary aggregations that can be time-variant or time-invariant, to different types of moving aggregations; see the patterns in [6, 7, 16, 17]. Many of these patterns have complex dynamical features, which are still not fully understood in terms of invariant sets of phase space.

One way to determine the phase space origin of many of the patterns exhibited by these nonlocal hyperbolic models is to determine whether they emerge from a sequence of bifurcations starting with a homogeneous steady-state solution, as the main parameters of the system are varied. However, in order to study the unfolding

of bifurcations near a steady-state solution, it is necessary to find out if Lyapunov–Schmidt and Centre Manifold reduction methods can be applied to this class of nonlocal hyperbolic equations.

Our main focus in this paper is to show the applicability of Lyapunov–Schmidt (LS) and Centre-Manifold (CM) reduction methods near steady-state bifurcation points of equations such as (3.1). Reduction methods are the first step to investigate the bifurcation and formation of patterns in mathematical models. While these methods have been commonly applied to ODEs and parabolic PDEs, their use to hyperbolic PDEs is still scarce. Moreover, the few analytical results existent in the literature for these hyperbolic PDEs are mainly applied to models describing phenomena in physics [8, 11, 24, 52].

Previous studies of nonlocal hyperbolic models for animal aggregations investigated local bifurcations near codimension-two Steady-state/Hopf [7] and Hopf/Hopf [6] bifurcation points with $\mathbf{O}(\mathbf{2})$-symmetry using *weakly-nonlinear analysis* (WNA) techniques, also known as the *method of multiple scales*. The formal equivalence of the reduced equations near bifurcation obtained with WNA and either LS and CM reduction has not been studied in a systematic way for these nonlocal models. Nevertheless, comparison between the results of WNA and CM reductions has been performed only for some specific cases of nonlocal hyperbolic models of type (3.1); see [7]. Note that such comparisons have been performed quite often for local fluid models [8, 24]. Moreover, the equivalence between LS and CM reductions has been investigated by Chossat and Golubitsky [11] in the context of Hopf bifurcation with symmetry.

In order to establish the validity of LS and CM reductions for nonlocal hyperbolic models, we first need to understand the linear operators associated with these models. The investigation of properties of linear operators coming from *local* hyperbolic first-order partial differential equations has attracted the attention of several authors [36, 45, 47]. We review some of these contributions in Sect. 3.3.2. Moreover, we present some details of the results for Fredholm operators inspired by integro-differential equations [21] and from functional differential equations (FDEs), e.g. differential equations on lattices, FDEs of mixed-type [38, 56]. We also review some Centre Manifold reduction results obtained for hyperbolic first-order partial differential equations and for general PDE systems, as well as mentioning recent results from FDE theory. Then, in the context of nonlocal models (3.1), we show that for the Lyapunov–Schmidt reduction the linear operator at a steady-state solution is a Fredholm operator of index zero. Moreover, for the Centre Manifold reduction, we verify that the nonlinear hyperbolic system (3.1) satisfies the conditions of an infinite-dimensional version of the Centre Manifold Theorem (CMT) of Vanderbauwhede and Iooss [67] (see also Haragus and Iooss [32]). Because the nonlocal hyperbolic models (3.1) for animal aggregations and movement are symmetric with respect to a group isomorphic to $\mathbf{O}(\mathbf{2})$, the reduction methods also respect the symmetry group so that the equations obtained in the reduced space have the required symmetry properties. In this paper, we do not perform explicit computations for particular cases of bifurcations arising in the context of nonlocal

aggregation models such as (3.1). However, our CMT puts on a rigorous footing the use of WNA computations done in [6, 7].

The content of the paper is organised as follows. We start in Sect. 3.2 with a short literature review of applications of hyperbolic PDEs to physics and biology. In particular, in Sect. 3.2.3, we focus on a hyperbolic first-order partial differential equations model with nonlocal terms for animal aggregation. Then, in Sect. 3.3 we discuss some general results on the Lyapunov–Schmidt and Central Manifold reductions and state our main results. We return to our nonlocal hyperbolic model in Sect. 3.4, where we show the main properties of the linear operator of the animal aggregation model, prove the main results concerning the Fredholm property and the use of Lyapunov–Schmidt reduction, and the applicability of the CMT. We conclude with Sect. 3.5, where we discuss some interesting future research directions.

3.2 1D Hyperbolic Models

Before discussing the various reduction methods, we first review briefly some 1D hyperbolic mathematical models derived to describe phenomena in physics and biology. This allows us to emphasise the importance of these models, and the lack of analytical studies to investigate the patterns exhibited by them. Moreover, by presenting some physics models, it allows us to review the existent analytical results developed for these models. Since generalisations of 1D hyperbolic models to 2D are not only more realistic but also more complex, their analytic investigation is more difficult. For this reason, we ignore them in this study.

3.2.1 Hyperbolic Models in Physics: Laser Models

To understand the complex dynamics of distributed feedback multi-section semiconductor lasers, [52] have investigated the following hyperbolic system describing the forward and backward propagating complex amplitudes of the light (u_1, u_2), coupled to an equation for the carrier density (v):

$$\frac{\partial u}{\partial t} = \left(-\frac{\partial u_1}{\partial x}, \frac{\partial u_2}{\partial x}\right) + G(x, u(x,t), v(x,t)), \tag{3.2a}$$

$$\frac{\partial v}{\partial t} = I(x,t) + H(x, u(x,t), v(x,t)) + \sum_{k=1}^{m} b_k \chi_{S_k}(x)\left(\frac{1}{x_k - x_{k-1}} \int_{S_k} v(y,t)dy - v(x,t)\right), \tag{3.2b}$$

The model describes the longitudinal dynamics of edge emitting lasers [52], and thus the 1D domain $[0, L] = U_{k=1}^{m} \bar{S}_k$, which is formed of m sub-sectional intervals $S_k := (x_{k-1}, x_k)$, $k = 1, \ldots, m$. The nonlinear operators $G : (0, L) \times \mathbb{C}^2 \times \mathbb{R} \to \mathbb{C}^2$

and $H : (0, L) \times \mathbb{C}^2 \times \mathbb{R} \to \mathbb{R}^2$ are differentiable with respect to (u, v), and measurable and bounded with respect to $x \in [0, L]$. Because these two operators have very complex descriptions, we will not show them here. However, for more examples of such nonlinear operators, see [52, 63]. Model (3.2) was completed with reflective boundary conditions for $u = (u_1, u_2)$

$$u_1(0, t) = r_0 u_2(0, t) + \alpha(t), \quad u_2(L, t) = r_L u_1(L, t),$$

and initial conditions for both u and v:

$$u(x, 0) = u_0(x), \quad v(x, 0) = v_0(x).$$

This class of models have been shown to exhibit very rich dynamics: from bifurcations to self-pulsations, hysteresis, excitability, frequency synchronisation [52]. While there are many studies that focus on the numerical description of these behaviours, only a few studies focus on their analytical investigation; see [52, 62, 63]. The complexity of these equations makes it difficult to show, for example, the existence and persistence of smooth invariant manifolds for general cases required for bifurcation results. Fortunately, a CMT and Fredholm properties for equations similar to (3.2) have been established in [44, 51, 52]. We will return to these results in Sect. 3.3.2. For more details on modelling of multi-section lasers, see the short introduction with many references given in Lichtner [51].

3.2.2 *Hyperbolic Models in Biology: Predator–Prey Models, Chemotaxis Models, Aggregation Models, Age-Dependent Models*

The last 20 years have seen an increase in the use of hyperbolic models to describe various biological phenomena: from self-organised biological aggregations (i.e., aggregations in the absence of a leader or external stimuli; [17]), to chemotactic aggregations (i.e., aggregations in the presence of a chemotactic signal produced by the members of the group; [23, 35, 37]), predator–prey dynamics [2, 3] or age-structured models [39, 42, 69].

One of the simplest models of type (3.1) with constant speed and constant turning rates was introduced and discussed extensively in [34, 36, 37]. The general form of this model for particles/animals aggregations, which includes a turning behaviour ($\lambda^{\pm}$) as well as a birth/death processes ($h^{\pm}(u^+, u^-)$), is given by

$$\frac{\partial u^+}{\partial t} + \gamma \frac{\partial u^+}{\partial x} = f^+[u^+, u^-] = -\lambda^+ u^+ + \lambda^- u^- + \frac{1}{2} h^+(u^+, u^-), \tag{3.3a}$$

$$\frac{\partial u^-}{\partial t} - \gamma \frac{\partial u^-}{\partial x} = f^-[u^+, u^-] = \lambda^+ u^+ - \lambda^- u^- + \frac{1}{2} h^-(u^+, u^-). \tag{3.3b}$$

Since the change in particles/animals movement directions is not always constant, but usually depends on (local or nonlocal) interactions with other particles, Eftimie et al. [18] considered a model of type (3.1) with constant speed $g^{\pm}[u^+, u^-] = \gamma$ (constant), nonlocal density-dependent turning rates $f^+[u^+, u^-] = -\lambda^+[u^+, u^-]u^+ + \lambda^-[u^+, u^-]u^-$, $f^- = -f^+$, and no birth/death dynamics ($h^{\pm} = 0$). This model was introduced to describe the formation and movement of self-organised biological aggregations in response to nonlocal social interactions among group members. Because of the complex spatial and spatio-temporal dynamics exhibited by this model we will review it in more detail in Sect. 3.2.3.

The chemotactic movement of animal/cell aggregations can also be described by (3.1), which is now coupled with an equation for the dynamics of the chemical $c(x, t)$ [35]:

$$\frac{\partial c(x,t)}{\partial t} = p(c, u^+, u^-) + D\frac{\partial^2 c(x,t)}{\partial x^2}, \tag{3.4}$$

where $p(c, u^+, u^-)$ describes the production/degradation of this chemical, and D is its diffusion rate. The chemical can influence the speed of animals/cells (i.e., $g^+[u^+, u^-, c]$ in (3.1)), their turning behaviour and even the birth–death dynamics of the population (i.e., $f^{\pm}[u^+, u^-, c]$ in (3.1)).

The 1D hyperbolic predator–prey models do not usually consider turning behaviour, i.e., $\lambda^{\pm} = 0$ (however, the 2D kinetic models can incorporate changes in movement direction in response to prey/predator behaviour; see [22]). In this case, the functions $f^{\pm}[u^+, u^-]$ incorporate only the predator–prey dynamics between the two populations. Usually, this dynamics is described by Lotka–Volterra-type terms [13], but other terms such as Holling-type functional responses can also be used [2]. Moreover, the interactions between the prey and predator populations can affect the speed of either prey or predator [13], as the animals speed up to avoid or to catch up with the other population. Note here that not all 1D predator–prey models are of the type (3.1). For example, Barbera et al. [2] derived a hyperbolic model where the hyperbolic equations for the two populations are coupled with transport equations for the dissipative fluxes.

A final type of hyperbolic model that we would like to mention briefly describes age-structured populations. The hyperbolic age-structured models (of the McKendrick–von Foerster type) have the general form [42]

$$\frac{\partial u(a,t)}{\partial t} + \frac{\partial u(a,t)}{\partial a} = -\lambda(a)u(a,t), \tag{3.5}$$

with $u(t, a)$ representing the density of the population of age a at time t, and $\lambda(a)$ describing the mortality rate. The description of the model is completed with conditions for the initial population $u(0, a) = Q(a)$, $a \geq 0$, and conditions for the newborn population:

$$u(0,t) = \int_{\alpha}^{\beta} u(x,t)m(x)dx, \tag{3.6}$$

with m the maternity function.

While all these models can exhibit a large variety of spatial and spatio-temporal patterns ranging from stationary and moving aggregations of animals/cells (e.g., stationary pulses, travelling pulses, breathers, ripples, zigzags; see [17]) to networks of cells [23], thorough investigations of these patterns are still not the common approach in mathematical biology. For a more in-depth review of pattern formation in hyperbolic models in biology, and the analytical and numerical techniques available to investigate them, see [15, 68]. Existence of reduction methods (e.g., Centre Manifold reduction) for local bifurcations of various types of equations described in this section has been established for parabolic equations [32] and for hyperbolic age-structured models [54].

Next, we focus on a particular class of nonlocal mathematical models for self-organised biological aggregations, for which there are a few preliminary studies on the local bifurcation of patterns near codimension-1 and codimension-2 bifurcation points [6, 7].

3.2.3 Self-organised Animal Aggregation Models

Here, we present in more detail a class of 1D nonlocal hyperbolic models derived to describe the formation and movement of various animal, cell and bacterial aggregations as a result of inter-individual communication [17, 18]. The evolution of densities of right-moving (u^+) and left-moving (u^-) individuals, which travel with constant velocity γ and change their movement direction from right to left (with rate λ^+) and from left to right (with rate λ^-) [17] is given by

$$\partial_t u^+(x,t) + \partial_x(\gamma u^+(x,t)) = -\lambda^+[u^+,u^-]u^+(x,t) + \lambda^-[u^+,u^-]u^-(x,t), \tag{3.7a}$$

$$\partial_t u^-(x,t) - \partial_x(\gamma u^-(x,t)) = \lambda^+[u^+,u^-]u^+(x,t) - \lambda^-[u^+,u^-]u^-(x,t), \tag{3.7b}$$

$$u^{\pm}(x,0) = u_0^{\pm}(x). \tag{3.7c}$$

The turning rates are defined as

$$\begin{aligned}\lambda^{\pm}[u^+,u^-] &= \lambda_1 + \lambda_2 f(y_r^{\pm}[u^+,u^-] - y_a^{\pm}[u^+,u^-] + y_{al}^{\pm}[u^+,u^-]), \\ &= \Big(\lambda_1 + \lambda_2 f(0)\Big) + \lambda_2\Big(f(y_r^{\pm} - y_a^{\pm} + y_{al}^{\pm}) - f(0)\Big).\end{aligned} \tag{3.8}$$

The terms $\lambda_1 + \lambda_2 f(0)$ and $\lambda_2\big(f(y^{\pm}) - f(0)\big)$ describe the baseline random turning rate and the bias turning rate, respectively. The function f is a positive function saturating for large values of its argument (to describe the biologically realistic situation

Table 3.1 Nonlocal social interaction terms ($y_j^{\pm}, j \in \{a, al, r\}$) introduced in [17]

Mechanisms	Communic. models	Interaction terms: attraction ($y_a^{\pm}$), repulsion ($y_r^{\pm}$), alignment ($y_{al}^{\pm}$)
Omnidirectional perception,	M2	$y_{a,r}^{\pm} = q_{r,a} \int_0^{\infty} K_{a,r}(s)\big(u(x \pm s) - u(x \mp s)\big)ds$
Omnidirectional emission		$y_{al}^{\pm} = q_{al} \int_0^{\infty} K_{al}(s)\big(u^{\mp}(x \mp s) + u^{\mp}(x \pm s) - u^{\pm}(x \mp s) - u^{\pm}(x \pm s)\big)ds$
Unidirectional perception,	M3	$y_{r,a}^{\pm} = q_{r,a} \int_0^{\infty} K_{r,a}(s)u(x \pm s)ds$
Omnidirectional emission		$y_{al}^{\pm} = q_{al} \int_0^{\infty} K_{al}(s)\left(u^{\mp}(x \pm s) - u^{\pm}(x \pm s)\right)ds$
Omnidirectional perception,	M4	$y_{r,a}^{\pm} = q_{r,a} \int_0^{\infty} K_{r,a}(s)\left(u^{\mp}(x \pm s) - u^{\pm}(x \mp s)\right)ds$
Unidirectional emission		$y_{al}^{\pm} = q_{al} \int_0^{\infty} K_{al}(s)\left(u^{\mp}(x \pm s) - u^{\pm}(x \mp s)\right)ds$
Unidirectional perception,	M5	$y_{a,r}^{\pm} = q_{r,a} \int_0^{\infty} K_{a,r}(s)u^{\mp}(x \pm s)ds$
Unidirectional emission		$y_{al}^{\pm} = q_{al} \int_0^{\infty} K_{al}(s)u^{\mp}(x \pm s)ds$

Constants q_a, q_{al}, q_r describe the magnitudes of the attractive, alignment and repulsive interactions, respectively. Kernels $K_{a,al,r}(s)$ describe the spatial ranges for each of these social interactions. Note that $u = u^+ + u^-$

of bounded turning rates). An example of such function is $f(y) = 0.5 + 0.5\tanh(y)$; see [6, 7, 17, 18]. These turning rates are influenced by the social interactions among individuals: attraction towards far-away neighbours ($y_a^{\pm}$), alignment with neighbours at intermediate distances ($y_{al}^{\pm}$) and repulsion from individuals at very close distances ($y_r^{\pm}$). Moreover, these social interactions depend on the perception of neighbours, which communicate via different mechanisms involving visual, sound, tactile or chemical signals. Table 3.1 shows the social interaction terms $y_{r,al,a}^{\pm}$ corresponding to four examples of communication mechanisms introduced in [17]. Note that in [17] the authors considered also a fifth mechanism (denoted M1), which combined attraction/repulsion forces as described by M2 and alignment forces as described by M4. Since this mechanism did not bring any new results in terms of pattern formation or model symmetry, it was ignored in more recent studies [6, 7] and thus we ignore it throughout this study too. The parameters $q_{r,a,al}$ are the magnitudes of the repulsive (r), attractive (a) and alignment (al) interactions. The kernels $K_{r,a,al}$ that model long-distance social interactions are given by Gaussian functions

$$K_j(s) = \frac{1}{2\pi m_j^2} e^{-(s-s_j)^2/(2m_j^2)}, \text{ with } j = r, a, al, \text{ and } m_j = s_j/8, \tag{3.9}$$

with $s_j, j = r, a, al$ being the width of the interaction ranges.

The integrals in Table 3.1 can be re-written by defining the operator $\mathcal{I}^{\pm}_{i,\ell}(u^+(x), u^-(x), s)$, with $\ell = a, r, al$, to describe the integrand for model Mi, $i = 2, 3, 4, 5$. The superscript $^{\pm}$ in $\mathcal{I}^{\pm}$ corresponds to the superscript in $y^{\pm}$. Thus, the social interaction terms become

$$y^{\pm}_{i,\ell}(u(x)) := \int_0^{\infty} K_\ell(s)\mathcal{I}^{\pm}_{i,\ell}(u^+(x), u^-(x), s)\, ds. \tag{3.10}$$

Note that $\mathcal{I}^{\pm}$ satisfies the following relation:

$$\mathcal{I}^{\pm}_{i,\ell}(v_1^+(x)+v_2^+(x), v_1^-(x)+v_2^-(x), s) = \mathcal{I}^{\pm}_{i,\ell}(v_1^+(x), v_1^-(x), s)+\mathcal{I}^{\pm}_{i,\ell}(v_2^+(x), v_2^-(x), s),$$

for all $i = 2, 3, 4, 5$ and $\ell = a, r, al$.

3.2.3.1 Periodic Boundary Conditions

Because numerical simulations of system (3.7) are performed on a finite domain $[0, L]$, we complete the description of the model by imposing boundary conditions. For a detailed discussion of biologically realistic boundary conditions for hyperbolic systems, see [30, 36]. Here, we consider periodic boundary conditions, which approximate the dynamics on infinite domains:

$$u^{\pm}(0, t) = u^{\pm}(L, t). \tag{3.11}$$

Hillen [36] showed the existence of solutions for *local hyperbolic systems* that satisfy periodic, homogeneous Dirichlet and homogeneous Neumann boundary conditions. Since model (3.7) is nonlocal, next we confirm that the integrals (3.10) are well defined for $u^{\pm}$ satisfying conditions (3.11). First define the space

$$L^2_{per} = \{u \in L^2(\mathbb{R}) \mid u(x) = u(x + L) \text{ for all x} \in [0, \text{L})\}.$$

We now show that for the interaction kernels $K(s)$ as in (3.9) and for $v \in L^2_{per}$ and

$$\tilde{K}^{\pm}v(x) := \int_0^{\infty} K(s)v(x \pm s)\, ds, \tag{3.12}$$

we have $\tilde{K}^{\pm}v(x) \in L^2_{per}$. To this end, we write $v(x) = \sum_{n=-\infty}^{\infty} c_n e^{ik_n x}$, where $k_n = 2\pi n/L$. Then,

$$\begin{aligned}\tilde{K}^{+}v(x) &= \int_0^{\infty} K(s) \sum_{n=-\infty}^{\infty} c_n e^{ik_n(x+s)}\, ds \\ &= \sum_{n=-\infty}^{\infty} c_n e^{ik_n x} \int_0^{\infty} K(s) e^{ik_n s}\, ds \\ &= \sum_{n=-\infty}^{\infty} c_n \hat{K}(n) e^{ik_n x}.\end{aligned}$$

Here, $\hat{K}(n)$ is the Fourier transform of $K(s)$, and $\hat{K}(\eta) \to 0$ as $|\eta| \to \infty$ exponentially fast (since $K(s)$ is Gaussian). Next we know that $|c_n|^2 < 1$ if $|n| > N$ for some $N \in \mathbb{N}$:

$$\sum_{n=0}^{\infty} |c_n \hat{K}(n)|^2 \leq \sum_{n=-N}^{N} |c_n|^2 |\hat{K}(n)|^2 + \sum_{n=-\infty}^{-(N+1)} |\hat{K}(n)|^2 + \sum_{n=N+1}^{\infty} |\hat{K}(n)|^2 < \infty.$$

Thus, $\tilde{K}^{+}v(x) \in L^2_{per}$ and the same holds for $\tilde{K}^{-}v(x)$.

Remark 1. Note that if we choose to work with functions in C^0_{per} with the sup-norm $||v||_\infty = \sup\{|v(x)| \mid x \in [0, L]\}$, it is a straightforward exercise to show that $\tilde{K}^{\pm}$ is a bounded operator from C^0_{per} to itself.

3.2.3.2 Reflective Boundary Conditions

Another type of boundary condition that is commonly used for systems of hyperbolic models (both in biology and physics; see, for example, [43, 47]) is the homogeneous Neumann condition. On the domain $[0, L/2]$, this condition reads

$$u^{+}(0,t) = u^{-}(0,t), \quad u^{+}(L/2,t) = u^{-}(L/2,t), \quad t \geq 0. \tag{3.13}$$

These Neumann (reflective) conditions describe the case where cells/animals cannot leave the domain and turn around at the boundary [30, 36, 53]. In regard to the equivalence between periodic and reflective boundary conditions for local hyperbolic systems, Lutscher [53] and Hillen [36] showed that for solutions that satisfy the mirror symmetry condition

$$u^{+}(x) = u^{-}(L - x), \quad x \in [0, L], \tag{3.14}$$

if one considers $w_0^{\pm}$ the initial data on $[0, L/2]$ that satisfies the no-flux boundary conditions (3.13), then it can be shown that

$$u_0^{\pm}(x) = \begin{cases} w_0^{\pm}(x) & \text{for } x \in [0, L/2], \\ w_0^{\mp}(L - x) & \text{for } x \in [L/2, L], \end{cases}$$

defines initial data on $[0, L]$ that satisfies periodic boundary conditions. Moreover, considering solutions $u^{\pm}$ of a local version of (3.7) with periodic boundary conditions (3.11), one can construct restrictions $w^{\pm}(x,t) = u^{\pm}(x,t)$, for $x \in [0, L/2]$, which are solutions of the same system with no-flux boundary conditions (3.13).

The steady-state solutions of nonlocal system (3.7) described below do satisfy the mirror symmetry condition (3.14). Therefore, the results of the next sections obtained for periodic (or zero-flux) boundary conditions can be easily generalised to zero-flux (or periodic) conditions.

3.2.3.3 Symmetry of Hyperbolic Models for Self-organised Biological Aggregations

Consider functions $u(x,t) = (u^+(x,t), u^-(x,t))$ satisfying the boundary condition $u(0,t) = u(L,t)$. We introduce the translation operator T_θ with $\theta \in [0, L)$, and the involution κ acting on $u(x,t)$ by

$$T_\theta.u(x,t) := u(x-\theta,t) \quad \text{and} \quad \kappa.(u^+(x,t), u^-(x,t)) := (u^-(L-x,t), u^+(L-x,t)). \tag{3.15}$$

The elements T_θ generate a group isomorphic to $\mathbf{SO}(2)$ because of the periodic boundary condition. One can check that $T_\theta \circ \kappa = \kappa \circ T_\theta^{-1}$, and so T_θ and κ generate a group isomorphic to $\mathbf{O(2)}$. Moreover, it is shown in [7] that system (3.7) is $\mathbf{O(2)}$-equivariant for any of the models M2, M3, M4, M5 described in Table 3.1; that is, for any solution $u(x,t)$ of (3.7), then $\kappa.u(x,t)$ and $T_\theta.u(x,t)$ are also solutions of (3.7) for any $\theta \in [0, L)$.

Consider now the action of a group Γ on a vector space V. The *isotropy subgroup* of the point $v \in V$ is

$$\Gamma_v := \{\rho \in \Gamma \mid \rho.v = v\}.$$

The symmetry of solutions of (3.7) is encoded in the isotropy subgroup.

3.2.3.4 Steady-State Solutions

Steady-state solutions of (3.7) are found by setting $\partial_t u^{\pm} = 0$ and solving the remaining integro-differential system. As shown in [7], by adding the two equations in (3.7), one notices that steady-state solutions $(u_*^+(x), u_*^-(x))$ satisfy $u_*^+(x) = u_*^-(x) + C$, where C is a constant.

For homogeneous steady-state solutions, let us first define the total conserved population density

$$A = \frac{1}{L}\int_0^L (u_*^+(x,t) + u_*^-(x,t))\, dx.$$

Then, the homogeneous steady-state solutions are of the form $(u_*^+(x), u_*^-(x)) = (A/2, A/2)$ and $(u_*^+(x), u_*^-(x)) = (A^*, A^{**})$, where $A^* \neq A^{**}$ and $A^* + A^{**} = A$. These solutions have isotropy subgroups $\mathbf{O(2)}$ and $\mathbf{SO}(2)$, respectively.

It is also possible to find non-homogeneous symmetric steady-state solutions with isotropy subgroup $\mathbf{D}_n$. Such solutions for $n = 1$ and $n = 3$ are observed in [7, 48]. It is shown in [7] that if $(u_*^+(x), u_*^-(x))$ has isotropy subgroup $\Sigma \supset \kappa$, then $u_*^+(x) = u_*^-(x)$. Therefore, steady-state solutions with isotropy subgroups $\mathbf{O(2)}$ and $\mathbf{D}_n$ have this property, but not steady-state solutions with isotropy subgroup $\mathbf{SO}(2)$.

3.3 Lyapunov–Schmidt and Centre Manifold Reductions

Before discussing the application of the Lyapunov–Schmidt and Centre Manifold reduction methods to model (3.7), we first present in Sects. 3.3.1 and 3.3.2 some general results on these methods. Then, in Sect. 3.4 we verify that the reduction methods can be applied to the nonlocal hyperbolic models (3.7).

3.3.1 General Theory

Let X be a Banach space and $\mathcal{L}$ a closed linear operator on X with dense domain $D(\mathcal{L})$. Consider a differential equation

$$\frac{d}{dt}u = \mathcal{L}(u, \mu) + F(u, \mu) := G(u, \mu), \tag{3.16}$$

where $F : \overline{D(\mathcal{L})} \times \mathbb{R}^\ell \to X$ is the nonlinear part of the operator which satisfies a Lipschitz condition, and μ is a bifurcation parameter. Suppose that $G(u_0, \mu_0) = 0$, and that the point spectrum of $\mathcal{L}(u_0, \mu_0)$ has values on the imaginary axis. That is, (u_0, μ_0) is a bifurcation point of the μ family. Without loss of generality, we can assume that $(u_0, \mu_0) = (0, 0)$.

To unfold this bifurcation using the Lyapunov–Schmidt (LS) reduction, the linear operator

$$\mathcal{T} = \frac{d}{dt} - \mathcal{L}(\cdot, \mu)$$

has to be Fredholm over a suitably chosen function space X'. Recall that an operator $\mathcal{T} : X' \to X'$ is *Fredholm* if the range of $\mathcal{T}$ is closed, and $\ker \mathcal{T}$ and $\operatorname{coker} \mathcal{T}$ are finite. The *index* of $\mathcal{T}$ is $\dim \ker \mathcal{T} - \dim \operatorname{coker} \mathcal{T}$. Notice that if $\mathcal{L}(u_0, \mu_0)$ has only zero eigenvalues on the imaginary axis (in its point spectrum), then the eigenfunctions do not depend on time, and the time derivative vanishes. Thus, it

is sufficient to verify the Fredholm property for $\mathcal{L}$ only. Typically, X' is a function space of 2π-periodic functions (if (3.16) has been suitably rescaled).

The Fredholm property of $\mathcal{L}$ enables a splitting $X' = \ker \mathcal{L} + M = N + \operatorname{coker} \mathcal{L}$, where M and N are, respectively, complementary subspaces to $\ker \mathcal{L}$ and $\operatorname{coker} \mathcal{L}$. Projection operators exist for each of these subspaces. Then, one can split G into the operators $G_1 : \ker \mathcal{L} \times M \times \mathbb{R}^k \to N$ and $G_2 : \ker \mathcal{L} \times M \times \mathbb{R}^k \to \operatorname{coker} \mathcal{L}$. Given coordinates (x_1, x_2, μ) for $\ker \mathcal{L} \times M \times \mathbb{R}^k$, one can solve $G_1 = 0$ using a properly chosen implicit function theorem (e.g. see Chicone [10]) and obtain $x_2 = \phi(x_1, \mu)$ near $(u_0, \mu_0) = (x_{10}, x_{20}, \mu_0)$. Because $\mathcal{L}$ has finite index, this leads to a finite-dimensional mapping $\tilde{G}_2 : \ker \mathcal{L} \times M \times \mathbb{R}^k \to \operatorname{coker} \mathcal{L}$ which contains the information about particular types of bifurcating solutions, depending on the choice of function space X'. A detailed description of the LS reduction can be found in [25] and in Chossat and Lauterbach [12].

We now briefly discuss the CMT for (equivariant) infinite-dimensional systems of Vanderbauwhede and Iooss [67]. See also [12] for the CMT in the context of equivariant systems. Suppose that the linear operator $\mathcal{L}$ at $(0, 0)$ satisfies the following assumptions:

(A0) The operator $\mathcal{L} : D(\mathcal{L}) \to X$ is bounded (in the graph norm).

(A1) For some $k \geq 2$, there exists a neighborhood $\mathcal{V} \subset D(\mathcal{L}) \times \mathbb{R}^\ell$ of $(0, 0)$ and Y a Banach space ($Y \subset X$) such that the nonlinear operator F is $C^k(\mathcal{V}, Y)$ and $F(0, 0) = 0$ and $DF(0, 0) = 0$.

(A2) The spectrum σ of $\mathcal{L}$ can be decomposed as $\sigma = \sigma_+ \cup \sigma_0 \cup \sigma_-$ where σ_+, σ_- contain, respectively, all λ such that $\operatorname{Re}(\lambda) > 0$ and $\operatorname{Re}(\lambda) < 0$ while σ_0 has all eigenvalues λ with $\operatorname{Re}(\lambda) = 0$. There exists $\delta > 0$ such that $\inf_{\lambda \in \sigma_+} \operatorname{Re}(\lambda) > \delta$ and $\sup_{\lambda \in \sigma_-} \operatorname{Re}(\lambda) < -\delta$. Moreover, σ_0 consists of a finite number of eigenvalues with finite algebraic multiplicity.

(A3) Let P_0 be the projection onto the generalised eigenspaces of σ_0 and $\mathcal{L}_h = I - P_0$, where the linear operator $\mathcal{L}_h$ is defined as $\mathcal{L}$ restricted to $D(\mathcal{L})_h = P_h D(\mathcal{L})$. Then for any $\eta \in [0, \delta]$ and any $f \in C_\eta(\mathbb{R}, Y_h)$ the linear problem

$$\frac{du_h}{dt} = \mathcal{L}_h u_h + f(t)$$

has a unique solution $u_h = K_h f$, where K_h is a bounded operator from $C_\eta(\mathbb{R}, Y_h)$ to $C_\eta(\mathbb{R}, D(\mathcal{L})_h)$ and $C_\eta(\mathbb{R}, \mathcal{X})$ is the space of exponentially growing functions with the norm

$$||u(t)||_{C_\eta} = \sup_{t \in \mathbb{R}} e^{-\eta|t|} ||u(t)||_X.$$

The norm of K_h is bounded by a continuous function of $\eta \in [0, \delta]$.

Then, there exists a parameter-dependent finite-dimensional manifold

$$M_0(\mu) = \{u_0 + \Psi(u_0, \mu) \mid u_0 \in E_0\},$$

where $E_0 = \text{Ran}\, P_0$, and such that $M_0(\mu)$ is locally invariant and contains the set of all bounded solutions. Letting $\mathcal{L}_0$ be the restriction of $\mathcal{L}$ to E_0, the reduced system of equations on the centre manifold has the form

$$\frac{du_0}{dt} = \mathcal{L}_0 u_0 + P_0 F(u_0 + \Psi(u_0, \mu), \mu) := g(u_0, \mu).$$

Moreover, if G is Γ-equivariant, then Γ acting on the vector space E_0 and M_0 can be chosen to be Γ-invariant. Therefore, g satisfies a Γ-equivariant condition: $g(\gamma u_0, \mu) = \gamma g(u_0, \mu)$.

Remark 2. The verification of assumption (A3) is often done by checking an inequality estimate on the resolvent operator $(\lambda I - \mathcal{L}_h)^{-1}$. This is illustrated in several examples in [32]. However, there are cases where the resolvent estimate does not hold, but (A3) does, see [40]. In our case, we do not attempt to prove the resolvent estimate due to the complexity of the resolvent operator coming from the nonlocal nature of the linear operator $\mathcal{L}_h$. Instead, we use the symmetries to decompose the problem into a family of finite-dimensional systems for which (A3) is easily satisfied.

We now introduce the function spaces for which we show our results. Recall that for some $\Omega \subset \mathbb{R}^n$, $W^{k,p}(\Omega, \mathbb{R}) \subset L^p(\Omega, \mathbb{R})$ is the Banach space of functions for which the first k weak derivatives are in $L^p(\Omega, \mathbb{R})$, and note that $W^{1,2}$ is a Hilbert subspace of L^2. We let $Y = W^{1,2}([0, L], \mathbb{R}^2)$ and $X = L^2([0, L], \mathbb{R}^2)$ and so $D(\mathcal{L}) = \{(u^+, u^-) \in Y \mid u^{\pm}(0) = u^{\pm}(L)\}$. We also define

$$Y_{per} = \{(u^+, u^-) \in W^{1,2}(\mathbb{R}, \mathbb{R}^2) \mid u^{\pm}(x) = u^{\pm}(x + L),\ x \in [0, L)\}$$

and

$$X_{per} = \{(u^+, u^-) \in L^2(\mathbb{R}, \mathbb{R}^2) \mid u^{\pm}(x) = u^{\pm}(x + L),\ x \in [0, L)\}.$$

For time-periodic solutions, we introduce

$$X_{2\pi} = \{u \in L^2([0, L] \times \mathbb{R}, \mathbb{R}^2) \mid u(x, t + 2\pi) = u(x, t)\},$$

$$Y_{2\pi} = X_{2\pi} \cap W^{1,2}([0, L] \times \mathbb{R}, \mathbb{R}^2)$$

and $D(\mathcal{T}) = \{u = (u^+, u^-) \in Y_{2\pi} \mid u^{\pm}(0, t) = u^{\pm}(L, t)\}$. Note that Hilbert spaces are chosen in order to exploit the orthogonal projection properties when showing that assumption (A3) of the CMT is satisfied. If one chooses to perform the analysis using Banach spaces $Y = C^1([0, L], \mathbb{R}^2)$ and $X = C^0([0, L], \mathbb{R}^2)$ along with their periodic counterparts, then assumption (A3) of the CM theorem of [67] becomes much more cumbersome to satisfy as we need to consider explicitly the resolvent operator $(\lambda I - \mathcal{L})^{-1}$ in order to define projections using the Dunford integral formula [19].

We are now ready to state the main results of this paper. Our first main result is the following.

Proposition 3 (Fredholm Operators). *Let $u_*(x)$ be a steady-state solution of (3.7) (for any of the models M2,...,M5 described in Table 3.1) and let $\mathcal{L}$ be the linearised operator at $u_*(x)$. Then, $\mathcal{L} : D(\mathcal{L}) \to X_{2\pi}$ and*

$$\mathcal{T} = \frac{d}{dt} - \mathcal{L}(\cdot, \mu) : D(\mathcal{T}) \to X_{2\pi}$$

are Fredholm operators of index zero.

A consequence of Proposition 3 is that the Lyapunov–Schmidt procedure can be performed on the operator $\mathcal{L}$ in the context of zero eigenvalues. If $\mathcal{L}$ has purely imaginary eigenvalues (after rescaling) $\pm i, \pm ik_1, \ldots, \pm ik_m$ where $k_1, \ldots, k_m \in \mathbb{Z}$ or a mixture of zero eigenvalues and purely imaginary eigenvalues as above, then the Lyapunov–Schmidt reduction is performed on $\mathcal{T}$ with a function space of 2π-periodic functions. The case of nonresonant purely imaginary eigenvalues is not easily handled using the Lyapunov–Schmidt reduction because the choice of Banach space of periodic functions cannot be chosen to simultaneously obtain all solutions with the two frequencies. Therefore, steady-state, Hopf (including resonances) and steady-state/Hopf (including resonances) bifurcation problems can be unfolded with the reduced equations obtained via Lyapunov–Schmidt reduction. Moreover, if $u_*(x)$ has isotropy subgroup $\Sigma \subset \mathbf{O}(2)$, then the reduced equation is Σ-equivariant for steady-state bifurcations and is $\Sigma \times \mathbf{S}^1$-equivariant for Hopf and steady-state/Hopf bifurcation problems, see [25].

The second result of this paper concerns the spectral properties at the linearisation near a steady-state.

Proposition 4 (Spectral Properties). *Let $u_*(x)$ be a steady-state solution of (3.7) (for any of the models M2,...,M5) and let $\mathcal{L}$ be the linearised operator at $u_*(x)$. Then, the spectrum of $\mathcal{L}$ is made up of isolated eigenvalues with finite multiplicity and with no accumulation point in $\mathbb{C}$. In particular, $\mathcal{L}$ can only have a finite number of eigenvalues with finite multiplicity on the imaginary axis.*

Therefore, property (A2) of the CMT of [67] is satisfied. This leads to the following result.

Proposition 5 (Centre Manifold Theorem). *Let $u_*(x)$ be a steady-state solution of (3.7) (for any of the models M2,...,M5) with isotropy subgroup $\mathbf{SO}(2)$ or $\mathbf{O}(2)$ and suppose that $\mathcal{L}$ has a finite number of eigenvalues on the imaginary axis. Then, assumptions (A0)–(A3) of the CMT of [67] are satisfied by $\mathcal{L}$ and F.*

The proof of these results is found in Sect. 3.4. In the next section, we describe recent results on Fredholm properties for linear operators originating from integro-differential equations and hyperbolic partial differential equations with local and nonlocal linear terms. We do the same for recent results about the applicability of Centre Manifold reduction in similar contexts.

3.3.2 Recent Results for Hyperbolic PDEs and FDEs

We summarise some recent results for FDEs and hyperbolic systems that use the theory for Lyapunov–Schmidt (LS) and Central Manifold (CM) reductions.

In regard to the LS reductions, it is well known that for ODEs and parabolic PDEs, the linear operator $\mathcal{L}$ is Fredholm [25]. In fact if $\mathcal{L}$ is a strongly continuous linear operator, then it is automatically Fredholm [4, 19]. This includes the case of FDEs (retarded and neutral) [31] and evolution semigroups [14]. If Eq. (3.16) has additional properties such as Hamiltonian structure, symmetry (including reversibility), those are preserved by the LS reduction [25, 28].

There are also several results available for linear operators $\mathcal{L}$ that are not strongly continuous. Mallet-Paret [56] establishes that for mixed-type FDEs

$$\dot{x} = \mathcal{L}(\xi)x_\xi = \sum_{j=1}^{N} A_j(\xi)x(\xi + r_j),$$

with $\mathcal{L}(\xi)$ asymptotically hyperbolic, the linear operator $(\Lambda_L x)(\xi) = x'(\xi) - \mathcal{L}(\xi)x_\xi$ is Fredholm. Moreover, in the case where the linear operator $\mathcal{L}(\xi)$ admits constant coefficient asymptotic operators $L_\pm$ as $\xi \to \pm\infty$, the index depends only on $L_\pm$ and a formula is given by the spectral flow, which counts the net number of eigenvalue crossings in the family $L(\xi)$ from L_- to L_+.

This work stimulated several other advances, especially in linking the Fredholm property to exponential dichotomy of the linear operator [33, 50]. Hupkes and Verduyn-Lunel [38] provided a direct generalisation of [56] to nonhyperbolic autonomous linear operators for mixed-type equations and they use their result to show a CMT for mixed-type FDEs. A more recent development is found in Faye and Scheel [21]. Here, the authors studied the following class of mixed-type equations with nonlocal terms, which is commonly used in neural models [20]:

$$\frac{d}{d\xi}U(\xi) = \int_{\mathbb{R}} K(\xi - \xi'; \xi)U(\xi')\,d\xi' + \sum_{j\in\mathcal{J}} A_j(\xi)U(\xi - \xi_j) + H(\xi),$$

where $U(\xi), H(\xi) \in \mathbb{C}^n$ and $K(\zeta;\xi)$, $A_j(\xi)$ are $n \times n$ complex matrices and $\mathcal{J}$ is countable with the shifts satisfying $\xi_1 = 0$, $\xi_k \neq \xi_k$ for $j \neq k \in \mathcal{J}$. As in [56], the authors showed that under some assumptions on the asymptotic operators defined for $\xi \to \pm\infty$, the operator defined by the right-hand side is Fredholm and the index can also be computed via its spectral flow.

In a different direction, Kmit and Recke [43, 44, 46] established the Fredholm property for linear operators associated with a class of first-order local hyperbolic systems of equations with reflective and Dirichlet boundary conditions. For example, in [44], they looked at the system of equations

$$\partial_t u + \gamma\partial_x u + a(x)u + b(x)v = f(x,t), \tag{3.17a}$$

$$\partial_t v - \gamma\partial_x v + c(x)u + d(x)v = g(x,t), \tag{3.17b}$$

where $x \in [0,1], f, g$ are 2π-periodic with respect to t, and u, v are 2π-periodic and satisfy the reflection boundary conditions

$$u(0,t) = r_0 v(0,t), \qquad v(1,t) = r_1 u(1,t).$$

By letting $W^\gamma = H^{0,\gamma} \times H^{0,\gamma}$ and $V^\gamma(r_0, r_1) = \{(u,v) \in W^\gamma \mid (\partial_t u + \partial_x u, \partial_v - \partial_x v) \in W^\gamma\}$, they showed that the linear operator on the left-hand side of (3.17) is Fredholm of index zero from V^γ to W^γ, for $a, d \in L^\infty(0,1)$ and $b, c \in BV(0,1)$. This Fredholm result was generalised in [45] to an n-dimensional system analogous to system (3.26), along with corresponding boundary conditions where the coefficient functions satisfy weak conditions. The same system was investigated in [46] but this time with C^1 coefficients satisfying some optimal non-resonance conditions. There, the authors showed that the linear operator is also Fredholm of index zero, but this time from $C_n \to C_n$, where C_n is the space of continuous mappings $u : [0,1] \times \mathbb{R} \to \mathbb{R}^n$ with a sup-norm. The result in [46] was then used in [47] to show a Hopf bifurcation theorem for semilinear hyperbolic systems

$$\omega \partial_t u_j + a_j(x,\lambda)\partial_x u_j + b_j(x,\lambda,u) = 0, \quad x \in (0,1),\ j = 1,\dots,n,$$

with smooth coefficients a_j, b_j, where the a_j were satisfying some extra nondegeneracy conditions. Moreover, the authors assumed that the solutions satisfied $u_j(x, t+2\pi) = u_j(x,t)$ for $x \in [0,1], j = 1,\dots,n$, and the reflection boundary conditions were chosen to be

$$\begin{aligned} u_j(0,t) &= \textstyle\sum_{k=m+1}^{n} r_{jk} u_k(0,t), \qquad j = 1,\dots,m \\ u_j(1,t) &= \textstyle\sum_{k=1}^{m} r_{jk} u_k(1,t), \qquad j = m+1,\dots,n. \end{aligned}$$

The statement of the Hopf theorem in [47] also depended on the coefficients r_{jk}.

In regard to the Centre Manifold reduction, Renardy [60] proved a version of the CMT for quasilinear hyperbolic equations, and then applied it to a Bénard problem describing viscoelastic fluid. On the other hand, Lichtner et al. [52] proved a version of the CMT for the class of local semilinear hyperbolic systems (3.2) describing laser dynamics. Here, the authors showed that the spectrum of the infinitesimal generator of the operator consists only of eigenvalues of finite algebraic multiplicity. Then, using a spectral gap property, they constructed an exponentially attracting invariant manifold on which can be defined the flow of the reduced system

$$\begin{aligned} \frac{\partial u_c}{\partial t} &= U_c(v)u_c + \epsilon G_c(t, v, u_c, \gamma(t, u_c, v, \epsilon)), \\ \frac{\partial v}{\partial t} &= \epsilon F(t, v, B(v)u_c + C(v)\gamma(t, u_c, v, \epsilon) + g(t, v)), \end{aligned}$$

with $u = B(v)u_c + C(v)u_s$ (and B and C smooth bases), $B(v)$ the spectral projection for the critical eigenvalues, and $u_s = \gamma(u_c, v, t, \epsilon)$ the C^k smooth graph representation of the invariant manifold.

Moreover, Hillen [36] investigated the existence of solutions for a local, linear version of the hyperbolic system (3.7) (i.e., $\lambda^{\pm}[u^+, u^-] = \lambda^{\pm} =$const.). He showed that the linear operator $\mathcal{L}$ with Neumann or periodic boundary conditions generates a strongly continuous semigroup on $L^p[0, L] \times L^p[0, L]$. Moreover, he calculated the spectrum of the linear operator for different types of boundary conditions. In a separate study, Hillen [34] showed the existence of an invariant manifold for the class of local hyperbolic reaction random-walk systems (3.3).

Finally, we mention the work of Magal and Ruan for hyperbolic semilinear equations with non-dense domains, which model age-structured populations [54, 55]. In reformulating the equation in operator form, the nonlocal boundary condition describing the fertility of the population enters the nonlinear terms of the functional equation. The authors adapted the approach of [67] to the context of integrated semigroups to prove the existence of centre manifolds near steady-state solutions. The main difficulty to overcome with non-dense domains was to determine a spectral decomposition of the whole function space X, while for densely defined domains of the linear operator, only the decomposition of the domain is necessary.

We can conclude from here that for hyperbolic systems, these reductions have been applied mainly to *local* systems. Next, we will focus on applying such results to *nonlocal* first-order hyperbolic models.

3.4 Application of LS and CM Reductions to Nonlocal Hyperbolic Systems for Biological Aggregations

Due to the nonlocal nature of the linear operator associated with system (3.7), the results of the previous section do not apply directly to the nonlocal hyperbolic model (3.7). In this section, we discuss the particularities of the Lyapunov–Schmidt and Centre Manifold reductions for this model. First, we focus on the linear operator associated with system (3.7), and investigate its compactness. Then, we prove the Fredholm property for this nonlocal operator. Finally, we discuss the spectrum of $\mathcal{L}$ and the application of the Central Manifold Theorem to model (3.7), therefore providing proofs of Propositions 3–5.

3.4.1 The Linear Operator

In this section, we extract the linear operator of (3.7) at an equilibrium solution $u_*(x) = (u_*^+(x), u_*^-(x))$. We rewrite Eq. (3.7) as

$$\pm\partial_t u^{\pm} = -\gamma\partial_x u^{\pm} - \lambda^+(x)u^+ + \lambda^-(x)u^- \tag{3.19}$$

Consider a small perturbations $u^{\pm}(x) = u_*^{\pm}(x) + u_1^{\pm}(x)$ near $u_*(x)$, where $|u_1^{\pm}(x)| << 1$, which we substitute in (3.7). Taking a Taylor expansion of $\lambda^{\pm} = \lambda_1 + \lambda_2 f(y_r^{\pm} - y_a^{\pm} + y_{al}^{\pm})$, we obtain after truncating beyond order one

$$\lambda^{\pm} \approx (\lambda_1 + \lambda_2 f(0)) + \lambda_2 f'(0)(y_r^{\pm} - y_a^{\pm} + y_{al}^{\pm}).$$

Let $L_1 = \lambda_1 + \lambda_2 f(0)$, $R_1 = \lambda_2 f'(0)$. Performing a Taylor expansion of (3.19) near (u_*^+, u_*^-), and keeping only the linear terms we obtain

$$\begin{aligned} -\lambda^+ u^+ + \lambda^- u^- = &-L_1(u_1^+(x) - u_1^-(x)) - R_1 u_*^+(x)\mathcal{K}_i^+(u_1(x)) - R_1 u_1^+(x)\mathcal{K}_i^+(u_*(x)) \\ &+R_1 u_*^-(x)\mathcal{K}_i^-(u_1(x)) + R_1 u_1^-(x)\mathcal{K}_i^-(u_*(x)). \end{aligned}$$

Here, we define $\mathcal{K}_i^{\pm}(u(x)) := y_{i,a}^{\pm}[u(x)] - y_{i,r}^{\pm}[u(x)] + y_{i,al}^{\pm}[u(x)]$. Then, the linear system is given by

$$\begin{aligned} \pm\partial_t u_1^{\pm} = &-\gamma\partial_x u_1^{\pm} - L_1(u_1^+ - u_1^-) - R_1 u_*^+ \mathcal{K}_i^+(u_1) \\ &-R_1 u_1^+ \mathcal{K}_i^+(u_*) + R_1 u_*^- \mathcal{K}_i^-(u_1) + R_1 u_1^- \mathcal{K}_i^-(u_*). \end{aligned}$$

The linear operator on the right-hand side is denoted by $\mathcal{L}(v, \mu)$, where μ is a vector of parameters—the main ones being q_a, q_r and q_{al}. We write $\mathcal{L} = \mathcal{L}_u + \mathcal{L}_c$, where

$$\mathcal{L}_u(v^+, v^-)^T = \begin{pmatrix} -\gamma\partial_x v^+ \\ \gamma\partial_x v^- \end{pmatrix} + \begin{pmatrix} -L_1 v^+ \\ -L_1 v^- \end{pmatrix}$$

and

$$\mathcal{L}_c(v^+, v^-, \mu)^T = \begin{pmatrix} L_1 v^- \\ L_1 v^+ \end{pmatrix} - R_1 \begin{pmatrix} \mathcal{K}_i^+(u_*)v^+ - \mathcal{K}_i^-(u_*)v^- + u_*^+\mathcal{K}_i^+(v) - u_*^-\mathcal{K}_i^-(v) \\ -\mathcal{K}_i^+(u_*)v^+ + \mathcal{K}_i^-(u_*)v^- - u_*^+\mathcal{K}_i^+(v) + u_*^-\mathcal{K}_i^-(v) \end{pmatrix}.$$

3.4.2 Compactness of $\mathcal{L}_c$

We now show that $\mathcal{L}_c$ is a compact operator. It is sufficient to show that $\tilde{K}^{\pm}u$ is compact as an operator from L^2_{per} to itself. We proceed in a standard way (see [61] for instance) by defining a family of operators with finite range (and therefore compact) which converges in the L^2 norm to the operators $\tilde{K}^{\pm}$, and thus the limit is also compact. To this end, we use a windowed orthonormal basis of $L^2(\mathbb{R})$ over intervals of length L defined by

$$g_{n,j}(x) = \frac{1}{\sqrt{L}} e^{ik_n x} \chi_{[jL,(j+1)L)}(x), \tag{3.20}$$

where $k_n = 2\pi n i/L$, χ is the characteristic function of the interval, and $n, j \in \mathbb{Z}$. Then,

$$K(s) = \sum_{n,j\in\mathbb{Z}} \alpha_{n,j} g_{n,j}(s), \quad \text{where} \quad \sum_{n,j\in\mathbb{Z}} |\alpha_{n,j}|^2 < \infty. \tag{3.21}$$

Consider the approximations of $K(s)$ on $(0,\infty)$ given by

$$K_{N,M}(s) = \sum_{j=0}^{M} \sum_{n=-N}^{N} \alpha_{n,j} g_{n,j}(s)$$

and define the operators $K^{\pm}_{N,M} : L^2_{per} \to L^2_{per}$ by

$$\tilde{K}^{\pm}_{N,M} u := \int_0^{\infty} K_{N,M}(s) u(x \pm s)\, ds.$$

Letting $u(x) = \sum_{\ell=-\infty}^{\infty} c_\ell e^{ik_\ell x} \in L^2_{per}$ and substituting (3.20) into $\tilde{K}^{+}_{N,M} u$ gives us

$$\begin{aligned}
\tilde{K}^{+}_{N,M} u &= \int_0^{\infty} K_{N,M}(s) u(x+s)\, ds \\
&= \int_0^{\infty} \sum_{j=0}^{M} \sum_{n=-N}^{N} \alpha_{n,j} g_{n,j}(s) \sum_{\ell=-\infty}^{\infty} c_\ell e^{ik_\ell x} e^{ik_\ell s}\, ds \\
&= \sum_{j=0}^{M} \int_{jL}^{(j+1)L} \sum_{n=-N}^{N} \alpha_{n,j} \frac{1}{\sqrt{L}} e^{ik_n s} \sum_{\ell=-\infty}^{\infty} c_\ell e^{ik_\ell x} e^{ik_\ell s}\, ds \\
&= \sum_{j=0}^{M} \sum_{n=-N}^{N} \sum_{\ell=-\infty}^{\infty} \alpha_{n,j} c_\ell \frac{1}{\sqrt{L}} e^{ik_\ell x} \int_{jL}^{(j+1)L} e^{ik_n s} e^{ik_\ell s}\, ds \\
&= \sum_{j=0}^{M} \sum_{n=-N}^{N} \alpha_{n,j} c_{-n} \sqrt{L}\, e^{ik_{-n} x},
\end{aligned}$$

where the last equality comes from the integral being nonzero (and equal to L) if and only if $n = -\ell$. Thus, $\tilde{K}_{N,M}$ has finite range and so is compact. Consider now

$$\begin{aligned}
&||\tilde{K} - \tilde{K}_{N,M}||^2 = \\
&\quad = \sup_{||u||_2=1} \int_0^L |(K - K_{N,M})(u)|^2\, dx \\
&\quad = \sup_{||u||_2=1} \int_0^L \left|(K(s) - K_{N,M}(s)) u(x+s)\, ds\right|^2\, dx \\
&\quad \le \sup_{||u||_2=1} \int_0^L \left(\int_0^{\infty} \left|(K(s) - K_{N,M}(s)) u(x+s)\right| ds \right)^2 dx \\
&\quad = \sup_{||u||_2=1} \frac{1}{L} \int_0^L \left(\sum_{j=M+1}^{\infty} \int_{jL}^{(j+1)L} \chi_{[jL,(j+1)L)}(s) \left| \sum_{n=N+1}^{\infty} \alpha_{n,j} e^{ik_n s} + \overline{\alpha_{n,j}} e^{-ik_n s} \right| |u(x+s)|\, ds \right)^2 dx \\
&\quad = \sup_{||u||_2=1} \frac{1}{L} \int_0^L \sum_{j=M+1}^{\infty} \left(\int_0^L \chi_{[jL,(j+1)L)}(s) \left| \sum_{n=N+1}^{\infty} \alpha_{n,j} e^{ik_n s} + \overline{\alpha_{n,j}} e^{-ik_n s} \right| |u(x+s)|\, ds \right)^2 dx
\end{aligned}$$

and using the Cauchy–Schwarz inequality we obtain

$$
\begin{aligned}
||\tilde{K} - \tilde{K}_{N,M}||^2 \leq & \\
&\leq \sup_{||u||_2=1} \frac{1}{L} \int_0^L \sum_{j=M+1}^{\infty} \int_0^L \chi_{[jL,(j+1)L)}(s) \sum_{n=N+1}^{\infty} \left| \alpha_{n,j} e^{ik_n s} + \overline{\alpha}_{n,j} e^{-ik_n s} \right|^2 ds \, ||u||_2^2 \, dx \\
&\leq \frac{1}{L} \int_0^L \sum_{j=M+1}^{\infty} \int_0^L \sum_{n=N+1}^{\infty} 2|\mathrm{Re}(\alpha_{n,j} e^{ik_n x})|^2 \, ds \, dx \\
&\leq \frac{1}{L} \int_0^L \int_0^L \sum_{j=M+1}^{\infty} \sum_{n=N+1}^{\infty} 2|\alpha_{n,j}|^2 \, ds \, dx \\
&= 2L \sum_{j=M+1}^{\infty} \sum_{n=N+1}^{\infty} |\alpha_{n,j}|^2 \to 0
\end{aligned}
$$

as $N, M \to \infty$ because the series of coefficients in (3.21) is finite. Therefore, $\tilde{K}_{N,M}$ converges to $\tilde{K}$ in the L^2-norm, which implies that $\tilde{K}$ is also a compact operator. The purely matrix portion of $\mathcal{L}_c$ is compact because it has finite range and the sum of compact operators is compact. We conclude that $\mathcal{L}_c$ is a compact operator.

3.4.3 *Fredholm Property and the Lyapunov–Schmidt Reduction*

It is shown in [41] (Chap. IV, Theorem 5.26) that if E, F are Banach spaces and $T : E \to F$ is a closed Fredholm operator and $A : E \to F$ is a compact operator, then $T + A$ is also Fredholm with the index of T and $T + A$ being equal. We use this result to show the Fredholm property for $\mathcal{L}$ and for the operator

$$\mathcal{T} = \frac{d}{dt} - \mathcal{L}$$

where $\mathcal{T} : D(\mathcal{T}) \to X$ and $D(\mathcal{T})$ is a subspace of a space of 2π time periodic solutions.

As mentioned above, for steady-state bifurcations, it is sufficient to show that $\mathcal{L}$ is Fredholm. For studying Hopf bifurcations (including resonance cases) and steady-state/Hopf mode-interactions, we need to show that $\mathcal{T}$ is a Fredholm operator.

$\mathcal{L}$ Operator: We use the splitting $\mathcal{L} = \mathcal{L}_u + \mathcal{L}_c$ above. We show that $\mathcal{L}_u$ is an isomorphism. Our goal is to solve $\mathcal{L}_u v = \tilde{h}$ with $v \in D(\mathcal{L})$ and $\tilde{h} = (h_1(s), h_2(s))^\top \in X$. Let $v(x) = (v^+(x), v^-(x))^\top$, $\mathcal{M} = \gamma^{-1} L_1 \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix} \right)$ and $h(s) = \gamma^{-1}(-h_1(s), h_2(s))^\top$. The equation is rewritten as a differential equation system $v'(x) = \mathcal{M} v(x) + h(s)$ and has solution

$$v(x) = e^{\mathcal{M}x} C + e^{\mathcal{M}x} \int_0^x e^{-\mathcal{M}s} h(s) \, ds.$$

Applying the boundary condition $C = v(0) = v(L)$ implies

$$v(0) = e^{\mathcal{M}L}v(0) + e^{\mathcal{M}L}\int_0^L e^{-\mathcal{M}s}h(s)\,ds.$$

Because $\gamma^{-1}L_1L \neq 0$, then $I - e^{\mathcal{M}L}$ is invertible and the system has a unique solution with

$$v(0) = (I - e^{\mathcal{M}L})^{-1}e^{\mathcal{M}L}\int_0^L e^{-\mathcal{M}s}h(s)\,ds.$$

Thus, $\mathcal{L}_u : D(\mathcal{L}) \to X$ is an isomorphism, and so it is a Fredholm operator of index 0. Because $\mathcal{L}_c$ is a compact operator, we conclude that $\mathcal{L}$ is also a Fredholm operator of index zero.

$\mathcal{T}$ Operator We proceed in a similar way as for $\mathcal{L}$ and use the splitting

$$\mathcal{T} = \mathcal{T}_u - \mathcal{L}_c := \frac{d}{dt} - \mathcal{L}_u - \mathcal{L}_c.$$

We show that $\mathcal{T}_u$ is an isomorphism when defined with appropriate function spaces of 2π-time-periodic functions (since we are interested in time-periodic solutions emerging from Hopf bifurcations). For $2\pi/\omega$-periodic solutions a time-rescaling gives a one-to-one correspondence with the 2π periodic solutions, see [25]. Let $h(x,t) = (h_1(x,t), h_2(x,t)) \in X_{2\pi}$ and consider the equation $\mathcal{T}_u v = h$ where $v \in D(\mathcal{T})$. This equation is transformed into the decoupled transport system

$$\partial_t v^+ + \gamma\partial_x v^+ = -L_1 v^+ + h_1(x,t) \tag{3.22a}$$

$$\partial_t v^- - \gamma\partial_x v^- = -L_1 v^- + h_2(x,t) \tag{3.22b}$$

The existence and uniqueness of solutions of (3.22) with periodic boundary conditions follows the proof in [36]. Thus $\mathcal{T}_u$ is an isomorphism and therefore it is Fredholm of index zero. Again, the compact perturbation preserves the Fredholm property.

3.4.4 Centre Manifold Theorem

We now show that we can apply the CMT as stated in [32]. We begin by investigating the spectrum of $\mathcal{L}$ in order to show that assumptions (A2) and (A3) of Sect. 3.3.1 are satisfied. Assumptions (A0) and (A1) are straightforward to verify and are discussed below.

The operator $\mathcal{L} : D(\mathcal{L}) \subset Y \to X$ is a compact perturbation of the closed differential operator $\gamma(-\partial_x, \partial_x)^T : D(\mathcal{L}) \subset Y \to X$. Therefore, they have the

same essential spectrum (Kato [41], Chap. IV, Theorem 5.35). The differential operator $\gamma(-\partial_x, \partial_x)$ (with periodic boundary conditions) has compact resolvent and its spectrum is a point spectrum given by

$$\left\{2k\frac{\gamma\pi}{L}i \mid k \in \mathbb{Z}\right\},$$

see Hillen [36]. Therefore, $\mathcal{L}$ has empty essential spectrum and so the resolvent set of $\mathcal{L}$ is non-empty. It is straightforward to conclude that $\mathcal{L}$ also has compact resolvent. This is done by noticing that for two invertible operators U, V one can verify that $U^{-1}-V^{-1} = U^{-1}(V-U)V^{-1}$. Choose λ in the resolvent set of $\gamma(-\partial_x, \partial_x)$ and letting $V = \lambda I - \gamma(-\partial_x, \partial_x)$, $U = \lambda I - \mathcal{L}$, then

$$U^{-1} = V^{-1}(I - (\mathcal{L} - \gamma(-\partial_x, \partial_x))V^{-1})^{-1}$$

is compact because V^{-1}, $\mathcal{L} - \gamma(-\partial_x, \partial_x)$ are compact, the product of a compact operator and a bounded operator is compact [41] (Chap. III, Theorem 4.8), and λ is chosen so that $(I - (\mathcal{L} - \gamma(-\partial_x, \partial_x))V^{-1})^{-1}$ exists and is bounded. The spectrum of $\mathcal{L}$ is a point spectrum consisting of isolated eigenvalues with finite multiplicity and with no accumulation points. Thus, assumption (A2) is automatically satisfied.

We now turn to assumption (A3). We consider only steady-state solutions $u_*(x)$ which have the isotropy subgroups $\Sigma = \mathbf{SO}(2)$ and $\Sigma = \mathbf{O}(2)$. For a steady-state solution $u_*(x)$ with isotropy subgroup Σ, the tangent space to X at $u_*(x)$ (which is isomorphic to X as a Hilbert space) is Σ-invariant. Since $\Sigma \subset \mathbf{O}(2)$ is a compact group acting on the separable Hilbert space X, a consequence of the Peter–Weyl theorem [5] implies that the action of Σ leads to a decomposition of the space as a direct sum of finite-dimensional irreducible representations; that is,

$$X = \bigoplus_{k=1}^{\infty} U_k$$

where U_k, $k = 1, 2, \ldots$ are irreducible representations of Σ. The isotypic decomposition of X with respect to the Σ action is obtained by grouping Σ-isomorphic representations into the so-called isotypic components $\tilde{U}_\ell$, with $\ell \in \mathcal{I}$, where $\mathcal{I}$ is the indexing set for isomorphism classes of irreducible representations of Σ; see [26, 27] for details. The cases $\Sigma = \mathbf{SO}(2)$ and $\Sigma = \mathbf{O}(2)$ have a countably infinite number of non-isomorphic irreducible representations and are studied together. The case $\mathbf{D}_n$ has a finite number of isomorphism classes which leads to a decomposition of the tangent space into infinite-dimensional isotypic blocks. We do not consider this case in this paper.

SO(2) and O(2) Symmetric Steady-States Let e_j, $j = 1, 2$, be the standard basis vectors of $\mathbb{C}^2$. The subspaces

$$V_n^j = \{ze_j e^{ik_n x} + \text{c.c.} \mid z \in \mathbb{C}\}$$

are irreducible with respect to the $\mathbf{SO}(2)$ action (3.15) and V_n^j, V_m^j are not isomorphic if $n \neq m$. Thus, we have the decomposition

$$X = \bigoplus_{n=1}^{\infty} V_n^1 \oplus V_n^2$$

where for all $n \in \mathbb{N}$, $X_n := V_n^1 \oplus V_n^2$ are the isotypic components. In the case of $\mathbf{O}(2)$, the decomposition derived in [7] has the following form:

$$X = \bigoplus_{n=1}^{\infty} X_n, \tag{3.23}$$

where the subspaces X_n are defined as follows. Let $k_n = 2\pi n/L$. For all $n \geq 1$,

$$X_n = \left\{ a e^{ik_n x} + \text{c.c.} \mid a = (a^+, a^-)^\top \in \mathbb{C}^2 \right\} \subset X,$$

are isomorphic to $\mathbb{C}^2$, $\mathbf{O}(2)$-invariant, and can be decomposed into isomorphic irreducible representations. Let $f_1 = (1,1)^\top$ and $f_2 = (1,-1)^\top$, then

$$X_n^1 = \{(v_0 e^{ik_n x} + \overline{v}_0 e^{-ik_n x}) f_1 \mid v_0 \in \mathbb{C}\} \quad \text{and} \quad X_n^2 = \{(v_1 e^{ik_n x} + \overline{v}_1 e^{-ik_n x}) f_2 \mid v_1 \in \mathbb{C}\}, \tag{3.24}$$

are real two-dimensional $\mathbf{O}(2)$-irreducible representations (written in complex notation). The basis is given by $\{e^{ik_n x} f_1, e^{ik_n x} f_2, e^{-ik_n x} f_1, e^{-ik_n x} f_2\}$. Then, $X_n = X_n^1 \oplus X_n^2$, and the subspaces X_j are called isotypic components of the $\mathbf{O}(2)$ action on X. The decomposition (3.23) is called the isotypic decomposition of X.

The inner product on X given by

$$\langle \mathbf{v}, \mathbf{w} \rangle = \int_0^L (v^+ \overline{w^+} + v^- \overline{w^-}) dx, \tag{3.25}$$

where $\mathbf{v} = (v^+, v^-)$ and $\mathbf{w} = (w^+, w^-)$, is O(2)-invariant $\mathbf{O}(2)$-invariant. Using this inner product, one can verify that the subspaces X_j, X_k are mutually orthogonal for all $j \neq k$. Let $P_{X_k} : X \to X_k$ be the orthogonal projection associated with X_k. The $\mathbf{O}(2)$-equivariance of $\mathcal{L}$ implies a block diagonalisation along the isotypic decomposition. That is, $\mathcal{L}(X_k) \subset X_k$ and we write $\mathcal{L}_k := \mathcal{L}|_{X_k}$. Therefore, $\mathcal{L}$ decomposes as a direct sum of finite-dimensional matrices.

Let $\sigma(\mathcal{L}) = \sigma_+ \cup \sigma_0 \cup \sigma_-$ where σ_0 has a (nonzero) finite number of elements. From assumption (A2), we define $\delta > 0$ such that

$$\inf_{\lambda \in \sigma_+} (\text{Re}(\lambda)) > \delta \quad \text{and} \quad (Re(\lambda)) \sup_{\lambda \in \sigma_-} < -\delta.$$

For each k, we define the projections $P_{k,+}$ and $P_{k,-}$ onto the stable and unstable subspaces of $\mathcal{L}_k$. Consider the linear system

$$\frac{du}{dt} = \mathcal{L}_h u + f(t), \tag{3.26}$$

where $f \in C_\eta(\mathbb{R}, Y_h)$ with $\eta \in [0, \delta]$. We use the isotypic decomposition of X and write $u \in C_\eta(\mathbb{R}, D(\mathcal{L})_h)$ as

$$u(t) = \sum_{k=1}^{\infty} \Phi_k(x) u_k(t),$$

where $\Phi_k(x)$ is the matrix with basis elements of X_k as columns and $u_k(t) = P_{X_k} u(t)$. This leads to the decomposition of (3.26) into an infinite family of finite-dimensional systems

$$\frac{du_k}{dt} = \mathcal{L}_{h,k} u_k + f_k(t), \tag{3.27}$$

where $f_k = P_{X_k} f$. For isotypic blocks with no eigenvalues on the imaginary axis, $\mathcal{L}_{h,k} = \mathcal{L}_k$, while for isotypic blocks with spectrum intersecting σ_0 and at least one of σ_+ or σ_-, then $\mathcal{L}_{h,k} = P_{k,+}\mathcal{L}_k + P_{k,-}\mathcal{L}_k$. We denote by $\sigma_{k,\pm}$ the hyperbolic part of the spectrum in $\mathcal{L}_{h,k}$.

The system of equations (3.27) has solution

$$u_k(t) = e^{\mathcal{L}_{h,k} t} u_0 + e^{\mathcal{L}_{h,k} t} \int_0^t e^{-\mathcal{L}_{h,k} s} f_k(s)\, ds,$$

with the constraint $||u_k(t)||_{C_\eta(\mathbb{R}, X_k)} < \infty$ forcing the unique solution

$$u_0 = -\int_0^\infty e^{-\mathcal{L}_{h,k} s} P_{k,+} f_k(s)\, ds + \int_{-\infty}^0 e^{-\mathcal{L}_{h,k} s} P_{k,-} f_k(s)\, ds,$$

where the splitting $f_k(s) = P_{k,+} f_k + P_{k,-} f_k$ guarantees that the integrals are convergent. The solution can be rewritten $u_k = K_{k,h} f_k$ where

$$(K_{k,h} f_k)(t) := e^{\mathcal{L}_{h,k} t} \left(-\int_t^\infty e^{-\mathcal{L}_{h,k} s} P_{k,+} f_k(s)\, ds + \int_{-\infty}^t e^{-\mathcal{L}_{h,k} s} P_{k,-} f_k(s)\, ds \right).$$

Because of the finite-dimensionality, it is straightforward to check that

$$K_{k,h} \in \mathscr{L}(C_\eta(\mathbb{R}, X_k), C_\eta(\mathbb{R}, X_k)) \quad \text{and} \quad ||K_{k,h}|| < C_k(\eta),$$

where C_k is a continuous function of $\eta \in [0, \delta_k]$ with δ_k chosen so that

$$\inf_{\lambda \in \sigma_{k,+}} \mathrm{Re}(\lambda) > \delta_k \quad \text{and} \quad \sup_{\lambda \in \sigma_{k,-}} \mathrm{Re}(\lambda) < -\delta_k.$$

We define

$$u = K_h f := \sum_{k=1}^{\infty} \Phi_k(x) (K_{h,k} f_k)(t),$$

and one can verify that this provides a unique solution of equation (3.26). We now show that K_h is also a bounded operator; i.e. $K_h \in \mathscr{L}(C_\eta(\mathbb{R}, Y_h), C_\eta(\mathbb{R}, D(\mathcal{L}_h))$ and that the norm of K_h is bounded by a continuous function of $\eta \in [0, \delta]$.

Let $f \in Y_h$ and we write $f = \sum_{k=1}^{\infty} f_k$, where $f_k \in X_k$. Then,

$$||f||_{Y_h} = \int_0^L |f|^2\,dx + \int_0^L \left|\frac{d}{dx}f\right|^2\,dx = \sum_{k=1}^{\infty}(1+k^2)||f_k||^2_{X_k}.$$

For any $\eta \in [0,\delta]$, we write $||K_h f||_{C_\eta(\mathbb{R},D(\mathcal{L}_h))} =$

$$= \sup_{t\in\mathbb{R}} e^{-\eta|t|}||K_h f||_{D(\mathcal{L}_h)}$$

$$= \sup_{t\in\mathbb{R}} e^{-\eta|t|}\left(\int_0^L |(K_n f)(t)|^2\,dx + \int_0^L \left|\frac{d}{dx}(K_n f)(t)\right|^2\,dx\right)$$

$$= \sup_{t\in\mathbb{R}} e^{-\eta|t|}\left(\int_0^L \left|\sum_{k=1}^{\infty}\Phi_k(x)(K_{k,h}f_k)(t)\right|^2\,dx + \int_0^L \left|\frac{d}{dx}\left(\sum_{k=1}^{\infty}\Phi_k(x)(K_{k,h}f_k)(t)\right)\right|^2\,dx\right)$$

$$= \sup_{t\in\mathbb{R}} e^{-\eta|t|}\left(\int_0^L \left(\sum_{k=1}^{\infty}\Phi_k(x)(K_{k,h}f_k)(t)\right)\cdot\overline{\left(\sum_{j=1}^{\infty}\Phi_j(x)(K_{j,h}f_j)(t)\right)}\,dx\right.$$

$$\left. + \int_0^L \left(\frac{d}{dx}\left(\sum_{k=1}^{\infty}\Phi_k(x)(K_{k,h}f_k)(t)\right)\right)\cdot\left(\overline{\frac{d}{dx}\left(\sum_{k=1}^{\infty}\Phi_k(x)(K_{k,h}f_k)(t)\right)}\right)\,dx\right)$$

where $\cdot$ is the Euclidean inner product on $\mathbb{R}^2$. Commuting with the summations and by linearity, the inner product turns into the inner product on each X_k and the last line is equal to

$$\sup_{t\in\mathbb{R}} e^{-\eta|t|}\left(\sum_{k=1}^{\infty}\sum_{j=1}^{\infty}\int_0^L \Phi_k(x)(K_{k,h}f_k)(t)\cdot\Phi_j(x)(K_{j,h}f_j)(t)\,dx\right.$$
$$\left. + \sum_{k=1}^{\infty}\sum_{j=1}^{\infty}\int_0^L k^2\Phi_k(x)(K_{k,h}f_k)(t)\cdot\Phi_j(x)(K_{j,h}f_j)(t)\,dx\right).$$

By orthogonality of X_k, X_j for $j \neq k$ this last line becomes

$$\sup_{t\in\mathbb{R}} e^{-\eta|t|}\left(\sum_{k=1}^{\infty}||(K_{h,k}f_k)(t)||^2 + k^2||(K_{h,k}f_k)(t)||^2\right)$$

$$\leq \sup_{t\in\mathbb{R}} e^{-\eta|t|}\sum_{k=1}^{\infty} C_k(\eta)^2((1+k^2)||f_k||^2_{X_k}).$$
$$= C(\eta)^2||f||_{C_\eta(\mathbb{R},Y_h)}$$

where the inequality holds for all $\eta \in [0, \delta]$ because $\delta \leq \delta_k$ for all $k = 1, 2, \ldots$, and so we can factor out C_k from the summation and drop the index k. Thus, assumption (A3) is satisfied.

Assumption (A0) is automatically satisfied because $\mathcal{L}$ is a closed operator, see [32], while assumption (A1) is satisfied because of the tanh function in the definition of (3.7). Therefore, all four assumptions of the CMT of [67] are satisfied and Proposition 5 is verified.

3.5 Discussion and Generalisation of the Results

In this study, we investigated Lyapunov–Schmidt and Centre Manifold reductions for a class of nonlocal hyperbolic systems developed to model animal aggregations. We first presented the general theory behind these reduction methods, and the application of these results to FDEs and local hyperbolic systems (describing physical or biological phenomena). This approach allowed us to summarise the results existent in the literature, and to identify the results that are still missing. Then, we applied the two reduction methods to our class of nonlocal hyperbolic models. We showed the compactness of the operator associated with the nonlocal system, and then proved that the operator is Fredholm—a condition necessary for Lyapunov–Schmidt reduction. We emphasise here that the Fredholm property for hyperbolic equations, and in particular for nonlocal hyperbolic equations, has been less studied compared to the ODE or parabolic PDE models. Hence, our study fills a gap in the literature about Fredholm operators for nonlocal hyperbolic systems.

In regard to the Central Manifold reduction, we proved that the version of the CMT described in [32] can be applied to the nonlocal model (3.7), and hence the CM reduction in [7] is valid near steady-state solutions with isotropy subgroups $\mathbf{SO}(2)$ and $\mathbf{O}(\mathbf{2})$. The extension to steady-state solutions with isotropy subgroup $\mathbf{D}_n$ would require a different approach than the one presented here.

An interesting consequence of our results in this paper is that they extend automatically to two or more population models for animal/cellular aggregations [16]. They also extend to coupled equations of animal/cellular aggregation via chemotaxis, because the chemotaxis models typically contain a Laplacian operator and the theory is well known there [32].

It is possible that the LS and CM reductions extend in a straightforward way also to 2D nonlocal kinetic models (generalisations of the 1D hyperbolic models (3.7); see [22]). However, it was not the goal of our study to investigate this aspect. Such an analysis will form the object of a future study.

Another interesting question to be addressed in the future refers to a formal comparative study of the unfolding and dynamics obtained from Lyapunov–Schmidt reduction and WNA for arbitrary bifurcation problems.

Acknowledgements PLB acknowledges the financial support from NSERC in the form of a Discovery Grant. RE acknowledges support from an Engineering and Physical Sciences Research Council (UK) First Grant number EP/K033689/1. PLB would like to thank Christiane Rousseau for her support and encouragements over the years. PLB is particularly grateful to her for proposing and championing the Mathematics of Planet Earth initiative.

References

1. Armstrong, N.J., Painter, K.J., Sherratt, J.A.: A continuum approach to modelling cell–cell adhesion. J. Theor. Biol. **243**, 98–113 (2006)
2. Barbera, E., Currò, C., Valenti, G.: Wave features of a hyperbolic prey–predator model. Math. Methods Appl. Sci. **33**(12), 1504–1515 (2010)
3. Barbera, E., Consolo, G., Valenti, G.: A two or three compartments hyperbolic reaction-diffusion model for the aquatic food chain. Math. Biosci. Eng. **12**(3), 451–472 (2015)
4. Belleni-Morante, A., McBride, A.C.: Applied Nonlinear Semigroups: An Introduction. Wiley, New York (1998)
5. Bröcker, T., tom Dieck, T.: Representations of Compact Lie Groups, vol. 98. Springer-Verlag, New-York (1985)
6. Buono, P.-L., Eftimie, R.: Analysis of Hopf/Hopf bifurcations in nonlocal hyperbolic models for self-organised aggregations. Math. Models Methods Appl. Sci. **24**(2), 327–357 (2014)
7. Buono, P.-L., Eftimie, R.: Codimension-two bifurcations in animal aggregation models with symmetry. SIAM J. Appl. Dyn. Syst. **13**(4), 1542–1582 (2014)
8. Carr, J., Muncaster, R.G.: The application of centre manifolds to amplitude expansions. II. Infinite dimensional problems. J. Differ. Equ. **50**, 260–279 (1983)
9. Chertock, A., Kurganov, A., Polizzi, A., Timofeyev, I.: Pedestrian flow models with slowdown interactions. Math. Models Methods Appl. Sci. **24**, 249–275 (2014)
10. Chicone, C.: Ordinary Differential Equations with Applications. Texts in Applied Mathematics, vol. 34 Springer-Verlag, New-York (2006)
11. Chossat, P., Golubitsky, M.: Hopf bifurcation in the presence of symmetry, center manifold and Liapunov-Schmidt reduction. In: Atkinson, F.V., Langford, W.F., Mingarelli, A.B. (eds.) Oscillation, Bifurcation and Chaos. CMS-AMS Conference Proceedings Series, vol. 8, pp. 343–352. AMS, Providence (1987)
12. Chossat, P., Lauterbach, R.: Methods in Equivariant Bifurcation and Dynamical Systems. World Scientific, Singapore River Edge, NJ (2000)
13. Colombo, R.M., Rossi, E.: Hyperbolic predators vs. parabolic prey. Commun. Math. Sci. **13**(2), 369–400 (2015)
14. Diekmann, O., van Gils, S.A., Verduyn Lunel, S.M., Walther, H.O.: Delay Equations, Functional-, Complex-, and Nonlinear Analysis. Springer-Verlag, New-York (1995)
15. Eftimie, R.: Hyperbolic and kinetic models for self-organised biological aggregations and movement: a brief review. J. Math. Biol. **65**(1), 35–75 (2012)
16. Eftimie, R.: Simultaneous use of different communication mechanisms leads to spatial sorting and unexpected collective behaviours in animal groups. J. Theor. Biol. **337**, 42–53 (2013)
17. Eftimie, R., de Vries, G., Lewis, M.A.: Complex spatial group patterns result from different animal communication mechanisms. Proc. Natl. Acad. Sci. **104**(17), 6974–6979 (2007)
18. Eftimie, R., de Vries, G., Lewis, M.A., Lutscher, F.: Modeling group formation and activity patterns in self-organizing collectives of individuals. Bull. Math. Biol. **69**(5), 1537–1566 (2007)

19. Engel, K.-J., Nagel, R.: A Short Course on Operator Semigroups. Springer, Berlin (2006)
20. Ermentrout, G.B., McLeod, J.B.: Existence and uniqueness of travelling waves for a neural network. Proc. R. Soc. Edinb. **123A**, 461–478 (1993)
21. Faye, G., Scheel, A.: Fredholm properties of nonlocal differential operators via spectral flow. Indiana Univ. Math. J. **63**(5), 1–34 (2013)
22. Fetecau, R.: Collective behaviour of biological aggregations in two dimensions: a nonlocal kinetic model. Math. Models Methods Appl. Sci. **21**, 1539–1569 (2011)
23. Filbet, F., Laurencot, P., Perthame, B.: Derivation of hyperbolic models for chemosensitive movement. J. Math. Biol. **50**(2), 189–207 (2005)
24. Fujimura, K.: Methods of centre manifold and multiple scales in the theory of weakly nonlinear stability for fluid motions. Proc. R. Soc. Lond. A **434**, 719–733 (1991)
25. Golubitsky, M., Schaeffer, D.G.: Singularities and Groups in Bifurcation Theory, vol. 1. Springer, New York (1985)
26. Golubitsky, M., Stewart, I.: The Symmetry Perspective: From Equilibrium to Chaos in Phase Space and Physical Space. Birkhäuser, Basel (2002)
27. Golubitsky, M., Stewart, I., Schaeffer, D.G.: Singularities and Groups in Bifurcation Theory, vol. 2. Springer, New York (1988)
28. Golubitsky, M., Marsden, J., Stewart, I., Dellnitz, M.: The constrained Liapunov-Schmidt procedure and periodic orbits. In: Normal Forms and Homoclinic Chaos. Fields Institute Communications, vol. 4, pp. 81–127. American Mathematical Society, Providence, RI (1995)
29. Hackett-Jones, E.J., Landman, K.A., Fellner, K.: Aggregation patterns from non-local interactions: discrete stochastic and continuum modelling. Phys. Rev. E **85**, 041912 (2012)
30. Hadeler, K.P.: Reaction transport equations in biological modeling. Math. Comput. Model. **31**(4–5), 75–81 (2000). Proceedings of the Conference on Dynamical Systems in Biology and Medicine
31. Hale, J.K., Verduyn Lunel, S.M.: Introduction to Functional Differential Equations. Springer-Verlag, London (1993)
32. Haragus, M., Iooss, G.: Local Bifurcations, Centre Manifolds, and Normal Forms in Infinite-Dimensional Systems. Springer-Verlag, London (2010)
33. Härterich, J., Sandstede, B., Scheel, A.: Exponential dichotomies for linear non-autonomous functional differential equations of mixed type. Indiana Univ. Math. J. **51**, 1081–1109 (2002)
34. Hillen, T.: Invariance principles for hyperbolic random walk systems. J. Math. Anal. Appl. **210**(1), 360–374 (1997)
35. Hillen, T.: Hyperbolic models for chemosensitive movement. Math. Models Methods Appl. Sci. **12**(07), 1007–1034 (2002)
36. Hillen, T.: Existence theory for correlated random walks on bounded domains. Canad. Appl. Math. Quart. **18**(1), 1–40 (2010)
37. Hillen, T., Hadeler, K.P.: Hyperbolic systems and transport equations in mathematical biology. In: Warnecke, G. (ed.) Analysis and Numerics for Conservation Laws, pp. 257–279. Springer, Berlin/Heidelberg (2005)
38. Hupkes, H.J., Verduyn Lunel, S.M.: Center manifold theory for functional differential equations of mixed type. J. Dynam. Differ. Equ. **19**, 497–560 (2007)
39. Inaba, H.: Threshold and stability results for an age-structured epidemic model. J. Math. Biol. **28**, 411–434 (1990)
40. Iooss, G., Kirchgässner, K.: Travelling waves in a chain of coupled nonlinear oscillators. Commun. Math. Phys. **211**, 439–464 (2000)
41. Kato, T.: Perturbation Theory for Linear Operators. Springer-Verlag, New-York (1995)

42. Keyfitz, B.L., Keyfitz, N.: The Mckendrick partial differential equation and its uses in epidemiology and population study. Math. Comput. Model. **26**(6), 1–9 (1997)

43. Kmit, I.: Fredholm solvability of a periodic Neumann problem for a linear telegraph equation. Ukrainian Math. J. **65**(3) (2013)

44. Kmit, I., Recke, L.: Fredholm alternative for periodic-Dirichlet problems for linear hyperbolic systems. J. Math. Anal. Appl. **335**(1), 355–370 (2007)

45. Kmit, I., Recke, L.: Fredholmness and smooth dependence for linear time-periodic hyperbolic systems. J. Differ. Equ. **252**(2), 1962–1986 (2012)

46. Kmit, I., Recke, L.: Periodic solutions to dissipative hyperbolic systems. I: Fredholm solvability of linear problems. 999:DFG Research Center MATHEON (2013, preprint)

47. Kmit, I., Recke, L.: Hopf bifurcation for semilinear dissipative hyperbolic systems. J. Differ. Equ. **257**, 264–309 (2014)

48. Kovacic, M.: On matrix-free pseudo-arclength continuation methods applied to a nonlocal PDE in 1+1D with pseudo-spectral time-stepping. Master's thesis, University of Ontario Institute of Technology (2013)

49. Larkin, R., Szafoni, R.: Evidence for widely dispersed birds migrating together at night. Integr. Comp. Biol. **48**(1), 40–49 (2008)

50. Latushkin, Y., Tomilov, Y.: Fredholm differential operators with unbounded coefficients. J. Differ. Equ. **208**, 388–429 (2005)

51. Lichtner, M.: Exponential Dichotomy and Smooth Invariant Center Manifolds for Semilinear Hyperbolic Systems. Ph.D. thesis, Humboldt-Universität zu Berlin, Berlin (2006)

52. Lichtner, M., Radziunas, M., Recke, L.: Well-posedness, smooth dependence and centre manifold reduction for a semilinear hyperbolic system from laser dynamics. Math. Methods Appl. Sci. **30**, 931–960 (2007)

53. Lutscher, F.: Modeling alignment and movement of animals and cells. J. Math. Biol. **45**, 234–260 (2002)

54. Magal, P., Ruan, S.: On integrated semigroups and age structured models in L^p spaces. Differ. Integr. Equ. **20**(2), 197–239 (2007)

55. Magal, P., Ruan, S.: Center Manifolds for Semilinear Equations with Non-dense Domain and Applications to Hopf Bifurcation in Age-Structured Models. American Mathematical Society, Providence (2009)

56. Mallet-Paret, J.: The Fredholm alternative for functional differential equations of mixed type. J. Dyn. Differ. Equ. **11**(1), 1–47 (1999)

57. Mogilner, A., Edelstein-Keshet, L.: A non-local model for a swarm. J. Math. Biol. **38**, 534–570 (1999)

58. Pfistner, B.: A one dimensional model for the swarming behaviour of Myxobacteria. In: Hoffmann, G., Alt, W. (eds.) Biological Motion. Lecture Notes on Biomathematics, pp. 556–563. Springer, Berlin (1990)

59. Pliny the Elder: The Natural History. Book X. Taylor and Francis, London (1855)

60. Renardy, M.: A centre manifold theorem for hyperbolic PDEs. Proc. R. Soc. Edinb. Sect. A **122**(3–4), 363–377 (1992)

61. Robinson, J.C.: Infinite-Dimensional Dynamical Systems. Cambridge University Press, Cambridge (2001)

62. Sieber, J., Radziunas, M., Scneider, K.R.: Dynamics of multisection lasers. Math. Model. Anal. **9**(1), 51–66 (2004)

63. Sieber, J., Recke, L., Schneider, K.R.: Dynamics of multisection semiconductor lasers. J. Math. Sci. **124**(5), 5298–5309 (2004)

64. Topaz, C.M., Bertozzi, A.L., Lewis, M.A.: A nonlocal continuum model for biological aggregation. Bull. Math. Biol. **68**, 1601–1623 (2006)

65. Topaz, C.M., D'Orsogna, M.R., Edelstein-Keshet, L., Bernoff, A.J.: Locust dynamics: behavioral phase change and swarming. PLoS Comput. Biol. **8**, e1002642 (2012)

66. Topaz, C.M., D'Orsogna, M.R., Edelstein-Keshet, L., Bernoff, A.J.: Locust dynamics: behavioural phase change and swarming. PLoS Comput. Biol. **8**(8), e1002642 (2012)

67. Vanderbauwhede, A., Iooss, G.: Center manifold theory in infinite dimensions. In: Jones, C., Kirchgraber, U., Walther, H.O. (eds.) Dynamics Reported, vol. 1, pp. 125–163. Springer, Berlin (1992)

68. Witten, M. (ed.) Hyperbolic Partial Differential Equations. Populations, Reactors, Tides and Waves: Theory and Applications. Pergamon, Elmsford, N.Y. (1983)

69. Wollkind, D.J.: Applications of linear hyperbolic partial equations: predator–prey systems and gravitational instability of nebulae. Math. Model. **7**, 413–428 (1986)

Chapter 4
Canard Cycles with Three Breaking Mechanisms

Magdalena Caubergh and Robert Roussarie

Abstract This article deals with relaxation oscillations from a generic balanced canard cycle Γ subject to three breaking parameters of Hopf or jump type. We prove that in a rescaled layer of Γ there bifurcate at most five relaxation oscillations.

Keywords Balanced • n-multi-layer canard cycle • Breaking parameter • Rescaled layer • Cyclicity • Bifurcating limit cycle • Relaxation oscillations

4.1 Introduction

We consider slow fast systems of the form

$$X_{\lambda,\varepsilon} : \begin{cases} \dot{x} = f(x, y, \lambda, \varepsilon) \\ \dot{y} = \varepsilon g(x, y, \lambda, \varepsilon), \end{cases} \tag{4.1}$$

where f, g are smooth functions. In the study of relaxation oscillations we follow the general framework as introduced in [2, 3].

Each canard cycle is associated with one or more breaking mechanisms. As in [5] we consider only canard cycles with n generic breaking mechanisms, that may be Hopf breaking mechanisms and jump breaking mechanisms. Each mechanism depends on a so-called breaking parameter, in fact a function $a(\lambda)$ of the parameter λ. The assumed genericity is that the map $\lambda \to (a_1(\lambda), \ldots, a_n(\lambda))$ is a local diffeomorphism. Then, we will suppose that $\lambda = a = (a_1, \ldots, a_n)$. The canard cycle exists when $a = 0 \in \mathbb{R}^n$ and we want to study the system for $a \sim 0 \in \mathbb{R}^n$. A canard cycle with n breaking mechanisms is associated with n

M. Caubergh
Departament de Matemàtiques, Edifici C, Facultat de Ciències, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain

R. Roussarie (✉)
I.M.B., Université de Bourgogne, B.P. 47870, 21078 Dijon, France
e-mail: roussari@u-bourgogne.fr

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_4

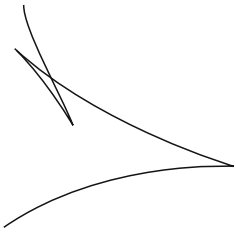


Fig. 4.1 A planar section of the bifurcation diagram of a three-layer canard cycle; in the bounded region five limit cycles are found

(horizontal fast) layers. For this reason we call such a canard cycle, indifferently: n-multi-layer canard cycle or canard cycle with n breaking mechanisms.

Canard cycles with one breaking mechanism were largely investigated and a general result in finite codimension was obtained in [3]. Canard cycles with two breaking mechanisms were introduced in [4] and their study was completed in [7]. Canard cycles with an arbitrarily large number n of mechanisms were introduced in [5]. There it is shown that bounding the limit cycles bifurcating from a generic balanced canard cycle Γ with n canard mechanisms in a *rescaled layer* is reduced to investigate the fixed points of a composition of translated power functions of the form

$$\phi^{\mathbf{r}}_{\boldsymbol{\alpha}}(\xi) = \phi^{r_n}_{\alpha_n} \circ \phi^{r_{n-1}}_{\alpha_{n-1}} \circ \ldots \circ \phi^{r_1}_{\alpha_1}(\xi), \tag{4.2}$$

where $\mathbf{r} = (r_1, \ldots, r_n)$, with $r_i \in \mathbb{R} \setminus \{0\}$ defined in terms of divergence quantities and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, with $\alpha_i \in \mathbb{R}$ obtained by rescaling a_i. The $\phi^{r_i}_{\alpha_i}$ in (4.2) are translated power functions given by

$$\phi^{r_i}_{\alpha_i}(\xi) = \alpha_i + \xi^{r_i}, i = 1, \ldots, n.$$

The maximal number of limit cycles bifurcating from Γ, i.e its cyclicity, in a rescaled layer, is equal to the maximal number of fixed points of $\phi^{\mathbf{r}}_{\boldsymbol{\alpha}}(\xi)$. For $n = 1$, respectively, $n = 2$, the cyclicity is equal to 2, respectively, 3. This is the bound expected for an elementary catastrophe (fold resp. cusp catastrophes), although the catastrophe theory does not apply here. For $n = 3$ an example by Panazzolo exhibits on generic sections in parameter space a bifurcation with three cusp points (see Fig. 4.1; this example was reported in [5]).

In this example one finds values of the parameter with five fixed points for $\phi^{\mathbf{r}}_{\boldsymbol{\alpha}}(\xi)$. Therefore, although this family of maps depends on a mere three dimensional parameter, its bifurcation diagram does not globally reduce to a unique elementary catastrophe. On the other hand, the cyclicity of Γ was not obtained for $n = 3$ in [5]. We obtain such a bound in this paper:

Theorem 1. *Let Γ be a balanced canard cycle with three breaking mechanisms, verifying the generic condition (G). Then there bifurcate at most five limit cycles in any rescaled layer of Γ.*

What is a balanced canard cycle Γ is explained in Definition 4, in terms of the slow divergence integrals (4.4) associated with this canard cycle. The generic condition (G) needed in Theorem 1 is specified below in (4.6) in terms of the divergence quantities (6). What is a rescaled layer is explained in Definition 5. Such a layer is a neighborhood of order ε in the layer variables. These layer variables are used to parameterize the canard cycles near Γ as well as the bifurcating limit cycles. Clearly, a rescaled layer does not cover a whole neighborhood of Γ; in Sect. 4.4, we further discuss this restriction.

Taking into account the Panazzolo's example, we see that the bound obtained in Theorem 1 is optimal. During the preparation of this paper, Panazzolo communicates us an article in preparation [8], where he announces that Eq. (4.2), for $n = 3$, has at most five roots. This implies of course Theorem 1. Nevertheless, the method which is used in our paper seems to be more simple. The method of Panazzolo [8], using the Khovanskii theory of fewnomials [6], allows to obtain for an arbitrary n the following bound:

$$M_n = 2^{n(2n-1)}(n+1)^{2n}.$$

Notice that this general formula does not give the accurate bound $M_3 = 5$, that is obtained in [8] by a direct study.

4.2 General Setting

Here we recall briefly the general setting for slow fast systems and canard cycles with an arbitrary number n of breaking mechanisms (see [5]).

4.2.1 Some Basic Definitions

The following assumptions are made on (4.1):

$$\frac{\partial f}{\partial y}(x, y, \lambda, 0) \neq 0, \forall (x, y, \lambda),$$

and

$$\text{if } f(x, y, \lambda, 0) = \frac{\partial f}{\partial x}(x, y, \lambda, 0) = 0, \text{ then } \frac{\partial^2 f}{\partial x^2}(x, y, \lambda, 0) \neq 0.$$

For $\varepsilon = 0$ we obtain the layer equation $X_{\lambda,0}$. The set $L_\lambda = \{f(x, y, \lambda, 0) = 0\}$ is referred to as the slow curve (of the layer equation). By the assumptions above it follows that the slow curve is a regular curve. Contact points are points

where the slow curve is tangent to the horizontal direction. Let C_λ be the set of these contact points. The set $L_\lambda \setminus C_\lambda$ is the union of normally hyperbolic arcs which may be of attracting type or of repelling type. Limit periodic sets appearing for $\varepsilon \to 0$ and not reduced to a singular point are called slow fast cycles (as they are the union of slow arcs on L_λ and fast orbits). They are compact invariant sets of $X_{\lambda,0}$. Periodic orbits bifurcating from these slow fast cycles are called relaxation oscillations. A distinction is made between canard and common relaxation oscillations. The one we are interested in are the canard relaxation oscillations, which bifurcate from a slow fast cycle containing attracting as well as repelling slow arcs. Such a slow fast cycle is called canard cycle.

4.2.2 Multi-Layer Canard Cycles

Let us recall that a n-multi-layer canard cycle is a canard cycle with n layers or indifferently with n breaking parameters. We suppose that the slow dynamics has no zeros on the slow arcs contained in Γ.

We have two different types of breaking mechanism:

1. The Hopf mechanism, occurring at degenerate contact point (x_0, y_0) where $g(x_0, y_0, 0, 0) = 0$ but $\frac{\partial g}{\partial x}(x_0, y_0, 0, 0) \neq 0$. The breaking parameter is the displacement of this root of g.
2. The jump mechanism where a fast orbit contained in Γ jumps from a non-degenerate contact point to another one. The breaking parameter is the vertical distance between the two contact points after perturbation.

More details about these two mechanisms can be found in [5].

Now, let $\mathcal{T}_1, \cdots, \mathcal{T}_n$ be the n breaking mechanisms which are labeled in the order compatible with the orientation of Γ (each $\mathcal{T}_i$ is situated either at a degenerate contact point or at a fast orbit between two jump points). To each $\mathcal{T}_i$ is associated a breaking function $a_i(\lambda), i = 1, \ldots, n$. We suppose the generic condition:

The map $\lambda \to (a_1(\lambda), \cdots, a_n(\lambda))$ *is a local diffeomorphism at* $\lambda = \lambda_0$. From now on we will assume that $\lambda = (a_1, \ldots, a_n)$.

The orientation of Γ induces a cyclic order on the breaking mechanisms and related loci; we denote them: $\mathcal{T}_1, \cdots, \mathcal{T}_i, \cdots, \mathcal{T}_n$, where i is a *cyclic index* which belongs to $\mathbb{Z}/n\mathbb{Z}$.

In between two breaking mechanisms we suppose to have exactly one fast orbit (in the positive direction) having both as α-limit and as ω-limit a point in $L_0 \setminus C_0$. Of course such a fast orbit has to belong to a one-parameter family of fast orbits having both as α-limit and as ω-limit a point in $L_0 - C_0$; we can call it a *layer of fast orbits* or *fast layer*.

In between a fast layer and a breaking mechanism we admit that Γ consists of a union of attracting slow curves and fast orbits, called *attracting sequence*.

A fast orbit in an attracting sequence necessarily has as α-limit a (jump) point in C_0, while we require that the ω-limit be situated in $L_0 \setminus C_0$.

We also require the same on Γ when we reverse time, implying similar conditions on a succession of repelling slow arcs, as we have on a succession of attracting slow arcs. The related succession of repelling slow curves and intermediate fast orbits will be called a *repelling sequence*.

We return now to the layer orbits. As described before, one has a unique layer orbit l_i in Γ, for each $i \in \mathbb{Z}/n\mathbb{Z}$ (see the convention for the index i introduced above). This layer orbit links the repelling sequence R_i to the attracting sequence A_{i+1}. As already observed, each l_i belongs to a one-parameter family of such fast orbits (a fast layer), and as a consequence the canard cycle is a member of an n-parameter family of similar canard cycles. To make this point more precise, we consider a transverse section Σ_i to l_i, transverse to the field $X_{0,0}$, for each $i \in \mathbb{Z}/n\mathbb{Z}$. Let u_i be a smooth regular parametrization of Σ_i, such that $\Sigma_i \cap l_i$ corresponds to $u_i = 0$. We choose the $|u_i|$ sufficiently small, let us say $u_i \in]-\beta, \beta[$ for some $\beta > 0$ small enough; we can replace l_i by $l_i(u_i)$, the fast orbit passing through the point $u_i \in \Sigma_i$ ($l_i = l_i(0)$) (in what follows, we will reduce each Σ_i to its part parameterized by $]-\beta, \beta[$ and we will write indifferently $u_i \in \Sigma_i$ or $u_i \in]-\beta, \beta[$). So, we have an n-parameter family of canard cycles Γ_u, parameterized by $u = (u_1, \cdots, u_n) \in]-\beta, \beta[^n$. The canard Γ_u is the one containing the fast layer orbits $l_i(u_i)$, for $i \in \mathbb{Z}/n\mathbb{Z}$. To emphasize the dependence on u_i, we will write $n_i(u_i), m_i(u_i)$ for the end points of the layer orbit $l_i(u_i)$, and also $A_i(u_{i-1}), R_i(u_i)$ for the attracting and repelling sequences associated with the transition $\mathcal{T}_i$. We can assume that our canard cycle Γ is just Γ_0. Parameters u_i are called the *layer variables.*

4.2.3 Equation of Bifurcating Limit Cycles

Let us consider an open connected arc $\sigma \subset L_0 \setminus C_0$. Along such an arc one can consider the slow divergence integral $\mathrm{Int}(\sigma)$, as defined in [1], for instance. For the system (4.1) and for an arc σ, above an interval $[x_1, x_2]$ without zero of g nor contact point in its interior, we have that

$$\mathrm{Int}(\sigma) = \mathrm{Int}(x_1, x_2) = -\int_{x_1}^{x_2} \frac{1}{g(x, y(x), 0, 0)} \Big(\frac{\partial f}{\partial x}(x, y(x), 0, 0) \Big)^2 dx, \tag{4.3}$$

where $y(x)$ is the implicit function defined by $f(x, y(x), 0, 0) = 0$ along σ. The end points x_1 and x_2 may be contact points.

Let us consider now the $2n$ integrals $I_{i,j}(u_j)$, defined for $i \in \mathbb{Z}/n\mathbb{Z}$, $j = i, i-1$:

$$I_{i,i-1}(u_{i-1}) = \mathrm{Int}(\sigma(A_i(u_{i-1})), \quad I_{i,i}(u_i) = -\mathrm{Int}(\sigma(R_i(u_i)) \tag{4.4}$$

where $\sigma(A_i(u_{i-1}))$ is the union of the slow arcs which constitute the attracting sequence $A_i(u_{i-1})$ and $\sigma(R_i(u_i))$ is the union of the slow arcs which constitute the repelling sequence $R_i(u_i)$.

Remark 2. For each breaking mechanism $\mathcal{T}_i$ we choose one section T_i. For the Hopf mechanisms, we have to introduce as breaking parameter a rescaled parameter: $\bar{a}_i =$

$\varepsilon^{-\delta} a_i$ for some $\delta > 0$, in consequence of the blow-up needed at this point. To keep the notations homogeneous, we will also write $\bar{a}_i$ for the breaking parameter at a jump breaking mechanism (i.e., we write $a_i = \bar{a}_i$ for a jump breaking parameter). We globally write : $\bar{a} = (\bar{a}_1, \cdots, \bar{a}_n)$ (see [5] for more details).

We recall now an important definition:

Definition 3. We say that a function $f(z, \varepsilon)$, with $z \in \mathbb{R}^p$ for some p, is *ε-regularly smooth in z* (or *ε-regularly $\mathcal{C}^\infty$ in z)* if f is continuous and all partial derivatives of f with respect to z exist and are continuous in (z, ε).

We want to recall now from [5] expressions for the transition maps for $\varepsilon > 0$, from the section Σ_{i-1} to the section T_i, along the flow of $X_{a,\varepsilon}$, and from Σ_i to T_i along the flow of $-X_{a,\varepsilon}$ (reversing time). There exist functions $\tilde{I}_{i,j}(u_j, \bar{a}, \varepsilon)$ which are ε-regularly $\mathcal{C}^\infty$ in $(u_j, \bar{a})$, such that

$$\tilde{I}_{i,j}(u_j, 0, 0) = I_{i,j}(u_j) \text{ for } i \in \mathbb{Z}/n\mathbb{Z}, \ j = i-1, i$$

and such that the *transition maps* have the following expressions:

1. From Σ_{i-1} to T_i : $u_{i-1} \to \exp \frac{\tilde{I}_{i,i-1}(u_{i-1}, \bar{a}, \varepsilon)}{\varepsilon} + f_{i,i-1}(\bar{a}, \varepsilon)$,
2. From Σ_i to T_i : $u_i \to \exp \frac{\tilde{I}_{i,i}(u_i, \bar{a}, \varepsilon)}{\varepsilon} + f_{i,i}(\bar{a}, \varepsilon)$,

with $f_{i,j}$ functions that are ε-regularly smooth in $\bar{a}$. One deduces in [5] the following system of n equations for the limit cycles:

$$\exp \frac{\tilde{I}_{i,i-1}(u_{i-1}, \bar{a}, \varepsilon)}{\varepsilon} - \exp \frac{\tilde{I}_{i,i}(u_i, \bar{a}, \varepsilon)}{\varepsilon} = \bar{a}_i \text{ for } i = 1, \cdots, n \tag{4.5}$$

with new functions $\tilde{I}_{i,j}$ which differ from the previous ones by terms of order $O(\varepsilon)$, which are ε-regularly $\mathcal{C}^\infty$ in $(u, \bar{a})$.

In Figs. 4.2 and 4.3 the transition maps are indicated by dotted lines.

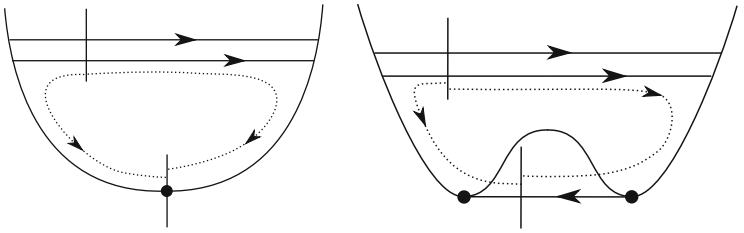


Fig. 4.2 Canards with one breaking parameter; Hopf breaking mechanism on the *left*, jump breaking mechanism on the *right*

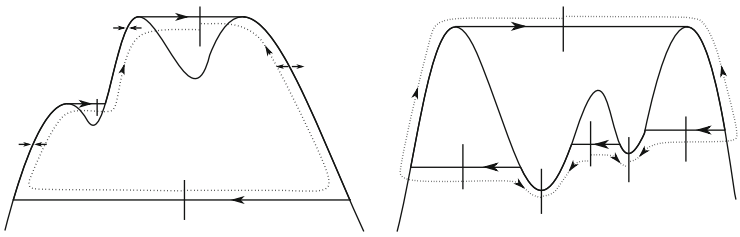


Fig. 4.3 On the *left* a canard cycle with one breaking mechanism (jump) and exhibiting an attracting sequence. On the *right* a canard cycle with three breaking mechanisms (one jump and two Hopf)

4.2.4 Rescaling Generic Balanced Canard Cycles

Recall Γ is the canard cycle Γ_u associated with $u = 0$.

Definition 4. The canard cycle Γ is said to be *balanced* if the integrals $I_{i,j}$ verify the following conditions:

$$I_{i,i}(0) = I_{i,i-1}(0) \text{ for } i \in \mathbb{Z}/n\mathbb{Z}.$$

Let us suppose that Γ is a balanced canard cycle. Then Γ is said to be generic if it verifies the generic condition

$$(G) : \prod_{i=1}^{n} I'_{i,i}(0) \neq \prod_{i=1}^{n} I'_{i,i-1}(0) \tag{4.6}$$

We assume from now on that Γ is a generic balanced canard cycle. It is proven in [5] that there exists an ε-regularly function $u_i(\bar{a}, \varepsilon)$ such that

$$\tilde{I}_{i,i}(u_i(\bar{a}, \varepsilon), \bar{a}, \varepsilon) = \tilde{I}_{i,i-1}(u_{i-1}(\bar{a}, \varepsilon), \bar{a}, \varepsilon),$$

for all $\varepsilon > 0$, small enough. We write $u(\bar{a}, \varepsilon) = (u_1(\bar{a}, \varepsilon), \ldots, u_n(\bar{a}, \varepsilon))$.

We can introduce now the *rescaled layer variables:*

Definition 5. Let us suppose that Γ is a generic balanced canard cycle and let $u(\bar{a}, \varepsilon) = (u_1(\bar{a}, \varepsilon), \ldots, u_n(\bar{a}, \varepsilon))$ the application defined above. For each $i = 1, \cdots, n$, the rescaled layer variable U_i is defined by

$$u_i = u_i(\bar{a}, \varepsilon) + \varepsilon U_i.$$

Taking $K_i > 0$, for $i = 1, \ldots, n$ arbitrarily large constants, we define a **rescaled layer** by taking $U_i \in [-K_i, K_i]$, for $i = 1, \ldots, n$, and ε small enough.

Introduce

$$I_i^0(\bar{a}, \varepsilon) = \tilde{I}_{i,i}(u_i(\bar{a}, \varepsilon), \bar{a}, \varepsilon) = \tilde{I}_{i,i-1}(u_{i-1}(\bar{a}, \varepsilon), \bar{a}, \varepsilon) \text{ and } I_{i,j}^1 = I'_{i,j}(0),$$

then we have that

$$\tilde{I}_{i,j}(u_j, \bar{a}, \varepsilon) = I_i^0(\bar{a}, \varepsilon) + \varepsilon I_{i,j}^1 U_j(1 + O(\varepsilon)), \tag{4.7}$$

where $O(\varepsilon)$ is ε-regularly smooth in U_j.

Substituting (4.7) in Eq. (4.5) we obtain, for $i = 1, \ldots, n$, the rescaled equations:

$$\exp\left(I_{i,i}^1 U_i(1 + O(\varepsilon))\right) - \exp\left(I_{i,i-1}^1 U_{i-1}(1 + O(\varepsilon))\right) = \alpha_i, \tag{4.8}$$

for rescaled parameter variables $\alpha_i = \bar{a}_i \exp(-I_i^0(\bar{a}, \varepsilon)/\varepsilon)$. To simplify the notation further, we also write: $I_{i,i}^1(0) = \tau_i$, $I_{i,i-1}^1(0) = \nu_{i-1}$ and $r_i = \frac{\nu_{i-1}}{\tau_{i-1}}$ for $i \in \mathbb{Z}/n\mathbb{Z}$. At the parameter (α, r) one associates in [5] the translated power function:

$$\Phi_\alpha^r(\xi) = \alpha + \xi^r.$$

Now, putting $\xi = \exp U_n$, it is proven in [5] that the system of equations (4.8) reduces to a one-dimensional fixed point equation for a map: $\xi \to \varphi_\alpha^r(\xi) + O(\varepsilon)$ where

$$\varphi_\alpha^r = \phi_{\alpha_n}^{r_n} \circ \cdots \circ \phi_{\alpha_1}^{r_1},$$

and $O(\varepsilon)$ is ε-regularly smooth in (ξ, α, r).

4.3 System with Three Breaking Parameters

In this section we particularize the general setting to systems with three breaking parameter mechanisms. Each of these mechanisms may be of Hopf or jump type. Figures 4.3 and 4.4 present examples of such system.

We will denote by u, v, w the layer variables, by $I(u), J(v), K(u), L(w), M(v)$ and $N(w)$ the 6 involved slow fast integrals, and by a, b, c the three breaking parameters. As above, we change the parameter (a, b, c) for the new parameter $(\bar{a}, \bar{b}, \bar{c})$ to take into account the possibility of Hopf type mechanisms (see Remark 2). We assume that $\lambda = (\bar{a}, \bar{b}, \bar{c})$ is the whole parameter of $X_{\lambda,\varepsilon}$.

As a consequence of their basic properties, the slow divergence integrals are strictly negative and with strictly non-zero derivative:

$$I'(u) \neq 0,\ K'(u) \neq 0,\ J'(v) \neq 0,\ M'(v) \neq 0,\ L'(w) \neq 0,\ N'(w) \neq 0.$$

We assume that the given canard cycle Γ is $\Gamma_{(0,0,0)}$, i.e. it corresponds to $u = v = w = 0$. It is supposed to be balanced, which here reads as

$$I(0) = J(0),\ K(0) = L(0) \text{ and } M(0) = N(0). \tag{4.9}$$

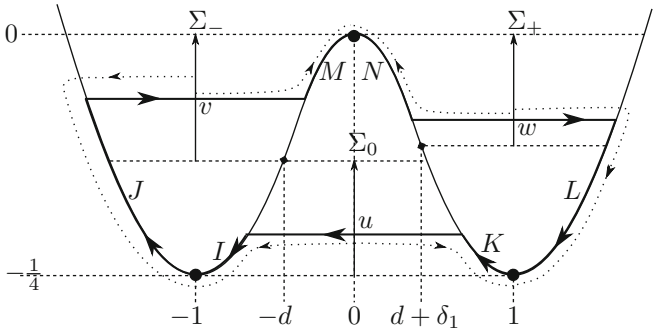


Fig. 4.4 Canard cycle Γ_{uvw} for (4.12) with three Hopf breaking parameters

Definition 6. The divergence quantities are the derivatives of the six divergence integrals, computed along the canard cycle:

$$\begin{cases} I_1 = I'(0), J_1 = J'(0), K_1 = K'(0), \\ L_1 = L'(0), M_1 = M'(0), N_1 = N'(0). \end{cases} \tag{4.10}$$

The canard cycle Γ is also supposed to be generic. This means that it verifies the property (G) in (4.6), which here reads as

$$G = \frac{I_1 L_1 M_1}{J_1 K_1 N_1} \neq 1, \tag{4.11}$$

where G is smooth in λ.

4.3.1 An Example

Consider the slow fast system in the Liénard plane

$$\begin{cases} \dot{x} = y + \frac{1}{2}x^2 - \frac{1}{4}x^4 \\ \dot{y} = \varepsilon g_{d\delta_1\delta_2}(x, a, b, c), \end{cases} \tag{4.12}$$

with

$$g_{d\delta_1\delta_2}(x, a, b, c) = \frac{(x-a)(x+1-b)(x-1-c)(x-d-\delta_1)(x+d)}{1+\delta_2 x}. \tag{4.13}$$

We see that the slow curve has two minima at $x = -1$ and $x = 1$ and one maximum at $x = 0$. The values of the minima is equal to $-\frac{1}{4}$ and the value of the maximum is equal to 0. The breaking parameters are (a, b, c) and for $(a, b, c) = (0, 0, 0)$ we have three Hopf mechanisms of canard cycles. The parameters (d, δ_1, δ_2) are constants that have to be chosen such that there exists a generic balanced canard cycle. We will not give a complete proof of this claim. We content ourselves in giving some indications which support it.

We write $F(x) = -\frac{1}{2}x^2 + \frac{1}{4}x^4$. We will choose $d \in]0, 1[$ and δ_1, δ_2 small enough such that $d + \delta_1 \in]0, 1[$ and $|\delta_2| < 1$. There are two singularities of the slow dynamics at points $(-d, F(d))$ and $(d+\delta_1, F(d+\delta_1))$ on the interior branches of the slow curve. The orientation of the slow dynamics is shown in Fig. 4.4. We choose the three layer sections to be $\Sigma_0 = \{x = 0\}$, $\Sigma_{-1}\{x = -1\}$ and $\Sigma_+ = \{x = 1\}$, with parametrization u, v, w, respectively, being equal to the coordinate y. We take $u \in]-\frac{1}{4}, \text{Inf}\{F(d), F(d+\delta_1)\}[$, $v \in]F(d), 0[$ and $w \in]F(d+\delta_1), 0[$. Then, for any convenient value (u, v, w) we have a canard cycle Γ_{uvw}.

We write $g_{d\delta_1\delta_2}(x) = g_{d\delta_1\delta_2}(x, 0, 0, 0)$. If $f(x) = \frac{\partial F}{\partial x}$, we have that

$$g_{d\delta_1\delta_2}(x) = f(x)\frac{(x - d + \delta_1)(x + d)}{1 + \delta_2 x}$$

and from (4.3) we obtain the following expression for the slow divergence integral:

$$\text{Int}(x_1, x_2) = \int_{x_1}^{x_2} \frac{1 + \delta_2 x}{(x - d + \delta_1)(x + d)} f(x)dx. \tag{4.14}$$

For any $y \in]-\frac{1}{4}, 0[$ we let $-x_+(y) < -x_0(y) < x_0(y) < x_+(y)$ be the four roots of the equation $\{F(x) = y\}$. The six slow divergence integrals are given by (see Fig. 4.4):

$$I(u) = \text{Int}(-x_0(u), -1),\ K(u) = \text{Int}(x_0(u), 1),\ J(v) = \text{Int}(-x_+(v), -1)$$

and

$$M(v) = \text{Int}(-x_0(v), 0),\ L(w) = \text{Int}(x_+(w), 1),\ N(w) = \text{Int}(x_0(w), 0).$$

We now explain how to find a generic balanced canard cycle. First, for $\delta_1 = \delta_2 = 0$, the system is symmetric with respect to the Oy-axis. Then the integrals I, K are identical and also the pairs J, L and M, N. Moreover the four integrals I, K, M, N vary from $-\infty$ to 0 and the two integrals J, L have a bounded variation. Taking, for instance, any value for v we have a unique value $u(v)$ such that the symmetric canard cycle $\Gamma_{u(v)vv}$ is balanced. Of course, as this canard cycle belongs to a one-parameter family of balanced canard cycles, it cannot verify the condition (G). However, it seems reasonable to think that there exist choices of the constants (d, δ_1, δ_2) which break this symmetry and for which there exists a generic balanced canard cycle.

4.3.2 System of Equations for Relaxation Oscillations

The system of governing equations for relaxation oscillations is given by

$$\begin{cases} \exp(\tilde{I}(u,\lambda,\varepsilon)/\varepsilon) - \exp(\tilde{J}(v,\lambda,\varepsilon)/\varepsilon) & = \bar{a}, \\ \exp(\tilde{K}(u,\lambda,\varepsilon)/\varepsilon) - \exp(\tilde{L}(w,\lambda,\varepsilon)/\varepsilon) & = \bar{b}, \\ \exp(\tilde{M}(v,\lambda,\varepsilon)/\varepsilon) - \exp(\tilde{N}(w,\lambda,\varepsilon)/\varepsilon) & = \bar{c}, \end{cases} \tag{4.15}$$

where $\lambda = (\bar{a}, \bar{b}, \bar{c})$ is near $(0,0,0)$ and the solutions (u,v,w) that we are looking for are near $(0,0,0)$. Since Γ is a generic balanced canard cycle [i.e., we have (4.9) and (4.11)], there exist ε-regularly smooth functions $u(\lambda,\varepsilon), v(\lambda,\varepsilon), w(\lambda,\varepsilon)$, with $u(\lambda,0) = v(\lambda,0) = w(\lambda,0) = 0$, such that

$$\begin{aligned} \tilde{I}(u(\lambda,\varepsilon),\lambda,\varepsilon) &\equiv \tilde{J}(v(\lambda,\varepsilon),\lambda,\varepsilon) \\ \tilde{K}(u(\lambda,\varepsilon),\lambda,\varepsilon) &\equiv \tilde{L}(w(\lambda,\varepsilon),\lambda,\varepsilon) \\ \tilde{M}(v(\lambda,\varepsilon),\lambda,\varepsilon) &\equiv \tilde{N}(w(\lambda,\varepsilon),\lambda,\varepsilon). \end{aligned}$$

We introduce the translated layer variables:

$$\bar{u} = u - u(\lambda,\varepsilon),\ \bar{v} = v - v(\lambda,\varepsilon),\ \bar{w} = w - w(\lambda,\varepsilon),$$

and we can expand

$$\tilde{I}(u,\lambda,\varepsilon) = \tilde{I}_0(\lambda,\varepsilon) + \tilde{I}_1(\lambda,\varepsilon)\bar{u}(1+\bar{u}^2),$$

and also the other functions $\tilde{J}, \tilde{K}, \ldots$. Let us notice that we have

$$\tilde{I}_0(\lambda,0) = I(0),\ \tilde{I}_1(\lambda,0) = I_1 \tag{4.16}$$

and similar ε-limits for the other functions $\tilde{J}, \tilde{K}, \ldots$. Next we introduce the rescaled parameter variables

$$\alpha = \bar{a}\exp(-\tilde{I}(0,\lambda,\varepsilon)/\varepsilon),\ \beta = \bar{b}\exp(-\tilde{K}(0,\lambda,\varepsilon)/\varepsilon),\ \gamma = \bar{c}\exp(-\tilde{M}(0,\lambda,\varepsilon)/\varepsilon),$$

and we write $\tilde{I}_1(\lambda,\varepsilon) = \tilde{I}'(0,\lambda,\varepsilon),\ \tilde{J}_1(\lambda,\varepsilon) = \tilde{J}'(0,\lambda,\varepsilon), \ldots$; to simplify reading, in the sequel we shortly write $\tilde{I}_1 = \tilde{I}_1(\lambda,\varepsilon), \tilde{J}_1 = \tilde{J}_1(\lambda,\varepsilon), \ldots$ although they do depend on (λ,ε). Then the system of governing equations for limit cycles (4.15) for $\bar{u}, \bar{v}, \bar{w} \to 0$ is reduced to

$$\begin{cases} \exp\left(\tilde{I}_1\bar{u}(1+O(\bar{u}))/\varepsilon\right) - \exp\left(\tilde{J}_1\bar{v}(1+O(\bar{v}))/\varepsilon\right) & = \alpha, \\ \exp\left(\tilde{K}_1\bar{u}(1+O(\bar{u}))/\varepsilon\right) - \exp\left(\tilde{L}_1\bar{w}(1+O(\bar{v}))/\varepsilon\right) & = \beta, \\ \exp\left(\tilde{M}_1\bar{v}(1+O(\bar{v}))/\varepsilon\right) - \exp\left(\tilde{N}_1\bar{w}(1+O(\bar{w}))/\varepsilon\right) & = \gamma, \end{cases} \tag{4.17}$$

4.3.3 Khovanskii's Reduction of the System of Equations

To control the number of solutions of (4.17) we use a Khovanskii's method. This approach is quite similar to the first step of the method used in [7] for canard cycles with two breaking parameters. In the system (4.17) we replace one equation by an equation equivalent to $D(\bar{u},\bar{v},\bar{w},\lambda,\varepsilon)=0$, where D is the Jacobian determinant of the left-hand side of (4.17) with respect to $(\bar{u},\bar{v},\bar{w})$. Since we have

$$\frac{\partial}{\partial\bar{u}}\exp\frac{\tilde{I}_1\bar{u}(1+O(\bar{u}))}{\varepsilon}=\frac{\tilde{I}_1(1+O(\bar{u}))}{\varepsilon}\exp\frac{\tilde{I}_1\bar{u}(1+O(\bar{u}))}{\varepsilon}=\frac{\tilde{I}_1}{\varepsilon}\exp\frac{\tilde{I}_1\upsilon\tilde{u}(1+O(\bar{u}))}{\varepsilon},$$

and analogous expressions for the other derivatives, we obtain that

$$D=\frac{\tilde{I}_1\tilde{L}_1\tilde{M}_1}{\varepsilon^3}\exp\Big[\frac{\tilde{I}_1\bar{u}(1+O(\bar{u}))+\tilde{M}_1\bar{v}(1+O(\bar{v}))+\tilde{L}_1\bar{w}(1+O(\bar{w}))}{\varepsilon}\Big]$$
$$-\frac{\tilde{J}_1\tilde{K}_1\tilde{N}_1}{\varepsilon^3}\exp\Big[\frac{\tilde{K}_1\bar{u}(1+O(\bar{u}))+\tilde{J}_1\bar{v}(1+O(\bar{v}))+\tilde{N}_1\bar{w}(1+O(\bar{w}))}{\varepsilon}\Big],$$

for $\bar{u},\bar{v},\bar{w}\to 0$. We write $\tilde{G}=\frac{\tilde{I}_1\tilde{L}_1\tilde{M}_1}{\tilde{J}_1\tilde{K}_1\tilde{N}_1}=G+O(\varepsilon)$, with a term $O(\varepsilon)$ which is ε-regularly smooth in λ and with G as defined in (4.11). If $G<0$, the Jacobian determinant is locally non-zero. Then, the system (4.17) has locally at most one solution. *From now on we will suppose that* $G>0$. In this case, for $\|(\bar{u},\bar{v},\bar{w})\|\to 0$, the equation $D=0$ is equivalent to

$$(\tilde{I}_1-\tilde{K}_1)\bar{u}+(\tilde{M}_1-\tilde{J}_1)\bar{v}+(\tilde{L}_1-\tilde{N}_1)\bar{w}+\varepsilon\ln\tilde{G}+O\Big(\|(\bar{u},\bar{v},\bar{w})\|^2\Big)=0. \quad (4.18)$$

Under the generic condition (G), as given in (4.11), at least one of the three coefficients I_1-K_1, J_1-M_1 or L_1-N_1 is different from 0. Without loss of generality, we can assume that $L_1-N_1\neq 0$. Then, using Implicit Function Theorem, (4.18) can be solved for $\bar{w}$. We thus obtain a function $\bar{w}$, that is ε-regularly smooth in $(\bar{u},\bar{v},\lambda)$ and from (4.18) we find for $\|(\bar{u},\bar{v})\|\to 0$:

$$\bar{w}(\bar{u},\bar{v},\lambda,\varepsilon)=-\frac{\tilde{I}_1-\tilde{K}_1}{\tilde{L}_1-\tilde{N}_1}\bar{u}-\frac{\tilde{M}_1-\tilde{J}_1}{\tilde{L}_1-\tilde{N}_1}\bar{v}-\varepsilon\ln\tilde{G}+O\Big(\|(\bar{u},\bar{v})\|^2\Big), \quad (4.19)$$

In order to simplify the system of equations, we consider a coordinate transformation of the layer variables

$$(\bar{u},\bar{v})\mapsto(\tilde{u},\tilde{v}),$$

where for $\bar{u},\bar{v}\to 0$, $\tilde{u}=\tilde{I}_1\bar{u}(1+O(\bar{u}))$, $\tilde{v}=\tilde{J}_1\bar{v}(1+O(\bar{v}))$, are the arguments of the exponential functions appearing in the first equation in (4.17). After this change of variables, the function (4.19) is replaced by the function $\tilde{w}$ with

$$\tilde{w}(\tilde{u},\tilde{v},\lambda,\varepsilon)=-\frac{\tilde{I}_1-\tilde{K}_1}{\tilde{L}_1-\tilde{N}_1}\frac{\tilde{u}}{\tilde{I}_1}-\frac{\tilde{M}_1-\tilde{J}_1}{\tilde{L}_1-\tilde{N}_1}\frac{\tilde{v}}{\tilde{J}_1}-\varepsilon\ln\tilde{G}+O\Big(\|(\tilde{u},\tilde{v})\|^2\Big),\ \|(\tilde{u},\tilde{v})\|\to 0.$$

The Khovanskii's method consists in replacing one of the three equations in (4.17), we choose the last one, by the equation $D = 0$, which for $\|(\bar{u}, \bar{v}, \bar{w})\| \to 0$ is equivalent to the equation $\bar{w} = \tilde{w}(\tilde{u}, \tilde{v}, \lambda, \varepsilon)$ for $\|(\tilde{u}, \tilde{v})\| \to 0$. In this way, we can eliminate $\bar{w}$ in the two first equations of (4.17) to obtain the following system of two equations in $(\tilde{u}, \tilde{v})$ for $\|(\tilde{u}, \tilde{v})\| \to 0$,

$$\begin{cases} \exp\dfrac{\tilde{u}}{\varepsilon} - \exp\dfrac{\tilde{v}}{\varepsilon} & = \alpha, \\ \exp\dfrac{\tilde{\sigma}\tilde{u}(1+O(\tilde{u}))}{\varepsilon} - \exp\dfrac{\tilde{\sigma}_1\tilde{u} + \tilde{\sigma}_2\tilde{v} - \varepsilon\ln\tilde{G} + O(\|(\tilde{u},\tilde{v})\|^2)}{\varepsilon} & = \beta, \end{cases} \tag{4.20}$$

where $\tilde{\sigma}, \tilde{\sigma}_1, \tilde{\sigma}_2$ are ε-regularly functions in λ given by

$$\tilde{\sigma} = \frac{\tilde{K}_1}{\tilde{I}_1},\ \tilde{\sigma}_1 = -\frac{\tilde{L}_1(\tilde{I}_1 - \tilde{K}_1)}{\tilde{I}_1(\tilde{L}_1 - \tilde{N}_1)},\ \tilde{\sigma}_2 = -\frac{\tilde{L}_1(\tilde{M}_1 - \tilde{J}_1)}{\tilde{J}_1(\tilde{L}_1 - \tilde{N}_1)}.$$

The system of equations (4.20) counts the number of contact points between the foliation defined by the last equation in (4.17) and the curves defined by the two first equations of (4.17). Therefore, for a given value of the parameter (λ, ε), the maximal number of solutions (u, v, w) of (4.17) is bounded by 1 + the number of solutions $(\tilde{u}, \tilde{v})$ of (4.20).

Notice that system (4.20) is very similar to the one encountered in [7] for the case of two breaking parameters. The only difference is that the second term in the second equation of (4.20) depends on two variables $\tilde{u}$ and $\tilde{v}$ and not just on the single variable $\tilde{v}$. This simple fact prevents us to proceed to further steps of Khovanskii's method as it was possible in [7]. For this reason we now have to restrict the study to a rescaled layer.

4.3.4 *Rescaled System of Equations*

As we announced in Sect. 4.1, the rescaling of Eq. (4.17) reduces the question of bounding the number of limit cycles bifurcating in a rescaled layer of Γ to the question of the number of fixed points for a fewnomial type map, here composed of three translation power functions:

$$\xi \mapsto -\alpha + (\beta + (-\gamma + \xi^{r_1})^{r_2})^{r_3}, \tag{4.21}$$

where $r_1 = \frac{M_1}{J_1}, r_2 = \frac{L_1}{N_1}, r_3 = \frac{I_1}{K_1}$. As we commented in Sect. 4.1 a direct approach of this question is announced in [8], to obtain 5 as bound. In the present paper we will obtain this bound by rescaling system (4.20). As this system is much simpler than (4.17), we believe that our proof is also much simpler than a direct study of (4.21).

We now enter in the proof of Theorem 1. To this end, we rescale the variables $\tilde{u}, \tilde{v}$ by

$$\tilde{u} = \varepsilon U, \quad \tilde{v} = \varepsilon V,$$

with U, V in arbitrarily large compact intervals. Next, we make the change of variables

$$\xi = \exp U, \eta = \exp V,$$

where now ξ, η are to be considered in arbitrarily compact intervals in $]0, +\infty[$. Recalling the notation of the divergence quantities in (4.10) and (4.16) we write

$$\sigma = \frac{K_1}{I_1}, \sigma_1 = -\frac{L_1(I_1 - K_1)}{I_1(L_1 - N_1)}, \sigma_2 = -\frac{L_1(M_1 - J_1)}{J_1(L_1 - N_1)};$$

then system (4.20) reads as

$$\begin{cases} \xi - \eta = \alpha \\ \xi^{\sigma} - G^{-1}\xi^{\sigma_1}\eta^{\sigma_2} + O(\varepsilon) = \beta, \end{cases} \tag{4.22}$$

where the uniformity of the term $O(\varepsilon)$ is relative to the choice of the compact domain for $(\xi, \eta, \alpha, \beta)$. Moreover this term is ε-regularly smooth in $(\xi, \eta, \alpha, \beta)$. Hence, by substitution of $\eta = \xi - \alpha$ in the second equation we obtain a one-dimensional equation:

$$\xi^{\sigma} - G^{-1}\xi^{\sigma_1}(\xi - \alpha)^{\sigma_2} - \beta + O(\varepsilon) = 0, \tag{4.23}$$

where again the term $O(\varepsilon)$ is uniform with respect to the choice of the compact domain for (ξ, α, β) and it is ε-regularly smooth in (ξ, α, β).

Theorem 1 follows from next claim:

Proposition 7. *For any fixed* $(\sigma, \sigma_1, \sigma_2)$ *and* $(\alpha, \beta) \in \mathbb{R}^2$, *the fewnomial type function*

$$\varphi_{\sigma,\sigma_1,\sigma_2}(\xi, \alpha, \beta) = \xi^{\sigma} - G^{-1}\xi^{\sigma_1}(\xi - \alpha)^{\sigma_2} - \beta, \tag{4.24}$$

has at most 4 *roots in* ξ *counted with their multiplicities in* $]\alpha, \infty[\cap]0, \infty[$.

Before proving Proposition 7 we first show how to deduce Theorem 1 from it. As we are looking for solutions (ξ, η) for system (4.22), in some compact set, we see that we can also restrict (α, β) to some compact set of $\mathbb{R}^2$. Now the term $O(\varepsilon)$ in (4.23) is uniform in compact sets for (ξ, α, β) and ε-regularly smooth.

Let A be a compact interval in $]0, +\infty[$. As the property to have at most four roots in ξ counted with their multiplicities is stable under smooth perturbations on compact domains, it follows that the left-hand side of (4.23) is a function with less than four roots in A for ε small enough. This implies that system (4.22) has less than four solutions on a given compact domain for ε small enough, from which Theorem 1 is proven.

Proof of Proposition 7. First we consider the case $\alpha = 0$. Then

$$\varphi_{\sigma,\sigma_1,\sigma_2}(\xi,\alpha,\beta) = \xi^\sigma - G^{-1}\xi^{\sigma_1+\sigma_2} - \beta.$$

This function has at most two roots counted with their multiplicities if $\sigma \neq \sigma_1 + \sigma_2$, and, as $G \neq 1$, has at most a simple root if $\sigma = \sigma_1 + \sigma_2$.

Next we suppose that $\alpha \neq 0$. To study the zeroes ξ of (4.24) in function of (α, β), we distinguish the case $\alpha > 0$ and $\alpha < 0$. As $(\sigma, \sigma_1, \sigma_2)$ is fixed, we denote the function $\varphi_{\sigma,\sigma_1,\sigma_2}$ simply by φ.

1. *Case* $\alpha > 0$. We introduce the variable μ by $\xi = \alpha(1+\mu)$ with $\mu > 0$ (since $a\mu = \xi - \alpha > 0$). Then φ transforms into

$$\varphi_+(\mu) \equiv \varphi(\alpha(1+\mu),\alpha,\beta) = \alpha^\sigma(1+\mu)^\sigma - G^{-1}\alpha^{\sigma_1+\sigma_2}\mu^{\sigma_2}(1+\mu)^{\sigma_1} - \beta. \quad (4.25)$$

To bound the zeroes of φ_+ we apply a division-derivation algorithm. Hence,

$$\frac{\partial\varphi_+}{\partial\mu}(\mu) = \sigma\alpha^\sigma(1+\mu)^{\sigma-1} - G^{-1}\alpha^{\sigma_1+\sigma_2}[\sigma_1(1+\mu)^{\sigma_1-1}\mu^{\sigma_2} + \sigma_2(1+\mu)^{\sigma_1}\mu^{\sigma_2-1}]$$

and so

$$(1+\mu)^{1-\sigma}\frac{\partial\varphi_+}{\partial\mu}(\mu) = \sigma\alpha^\sigma - G^{-1}\alpha^{\sigma_1+\sigma_2}\varphi_+^1(\mu),$$

where $\varphi_+^1(\mu) = (1+\mu)^{1-\sigma}[\sigma_1(1+\mu)^{\sigma_1-1}\mu^{\sigma_2} + \sigma_2(1+\mu)^{\sigma_1}\mu^{\sigma_2-1}]$. Then

$$\frac{\partial\varphi_+^1}{\partial\mu}(\mu) = (1+\mu)^{\sigma_1-\sigma-1}\mu^{\sigma_2-2}\varphi_+^2(\mu), \text{ where}$$

$$\varphi_+^2(\mu) = \sigma_2(\sigma-2\sigma_1-1)(1+\mu)\mu + \sigma_1(\sigma-\sigma_1)\mu^2 + \sigma_2(1-\sigma_2)(1+\mu)^2. \quad (4.26)$$

This last function is a polynomial of degree 2 in μ, more precisely:

$$\varphi_+^2(\mu) = (\sigma_1+\sigma_2)(\sigma_1+\sigma_2-\sigma)\mu^2 + \sigma_2(2\sigma_1+2\sigma_2-\sigma-1)\mu + \sigma_2(\sigma_2-1).$$

The number of positive zeroes μ for φ_+^2 corresponds to the one for $\frac{\partial^2\varphi_+}{\partial\mu^2}$. A direct and easy analysis shows that this polynomial is identically to zero if and only if the triple $(\sigma,\sigma_1,\sigma_2)$ is equal to $(\sigma,0,0)$, $(1,0,1)$ or to $(\sigma,\sigma,0)$ for some $\sigma \in \mathbb{R}$. In the case $(\sigma,0,0)$, we have that $\varphi_+(\mu) = \alpha^\sigma(1+\mu)^\sigma - G^{-1} - \beta$. In the case $(1,0,1)$, we have that $\varphi_+(\mu) = \alpha(1-G^{-1})\mu + \alpha - \beta$. In the case $(\sigma,\sigma,0)$, we have that $\varphi_+(\mu) = (1-G^{-1})\alpha^\sigma(1+\mu)^\sigma - \beta$. In all three cases the function φ_+ has at most a single root, which is simple. As a consequence φ for $\alpha > 0$ and $\beta \in \mathbb{R}$ has at most four zeroes, counted with their multiplicity.

2. *Case* $\alpha < 0$. Consider now the case $\alpha < 0$ and introduce the variable $\xi = -\alpha\mu = |\alpha|\mu$. We have that $\mu > 0$ for $\xi > 0$. Then φ transforms into

$$\varphi_-(\mu) = \varphi(|\alpha|\mu, \alpha, \beta) = |\alpha|^\sigma \mu^\sigma - G^{-1}|\alpha|^{\sigma_1+\sigma_2}(1+\mu)^{\sigma_2} - \beta.$$

The expression for $\varphi_-(\mu)$ is similar to the one for $\varphi_+(\mu)$ in (4.25), up to permutation of μ with $1+\mu$, and replacing α in φ_+ by $|\alpha|$. Then applying to φ_- two steps of division-derivation procedure as we did to φ_+ in the case $\alpha > 0$ leads to the quadratic polynomial φ_-^2, defined by $\varphi_2^-(\mu) = \sigma_2(\sigma - 2\sigma_1 - 1)\mu(1+\mu) + \sigma_1(\sigma-\sigma_1)(1+\mu)^2+\sigma_2(1-\sigma_2)\mu^2$, which is similar to (4.26), up to the permutation of μ with $1+\mu$. Therefore, also for $\alpha < 0, \beta \in \mathbb{R}$, there are at most four zeroes ξ for φ.

4.4 Open Questions

(1) Theorem 1 computes the cyclicity of a generic balanced canard cycle Γ in restriction to rescaled layers. Such a rescaled layer does not define a whole neighborhood of Γ. It remains to compute the true cyclicity of Γ, i.e. to find a bound of the number of bifurcating limit cycles in a whole neighborhood of Γ. A method for achieving this result would be to blow up the system of equations (4.17). In such blowing up the rescaled domain may be seen as a chart of the blown-up space (the so-called family chart). To complete the study of the cyclicity it thus would remain to study the blown-up system in the other charts (the parameter charts). This does not seem to be a too difficult task.

(2) In [7], the genericity is not assumed and a result was obtained for any finite codimension for canard cycles with two breaking mechanisms (besides the two breaking parameters one considers other parameters to unfold the situation). Moreover the result was obtained in a whole neighborhood (and not just in rescaled layers), by using the Khovanskii's method directly for the non-rescaled system. The idea was to "reduce the number of exponentials." Using this procedure for canard cycles with any number of breaking mechanisms, the first step works, as it produces an equation, Det $= 0$, without exponentials. Unfortunately, the number of exponentials does not decrease at the second step, as soon as there are more than three breaking mechanisms. It would be very interesting to find a general method to tackle the system of equations (4.5) in non-generic cases and for an arbitrarily number of equations.

Acknowledgements The first author is supported by Ramon y Cajal grant RYC-2011-07730, and also partially by grants MINECO/FEDER MTM2008-03437, MINECO MTM2013-40998-P, and AGAUR 2014SGR-568.

References

1. De Maesschalk, P., Dumortier, F.: Time analysis and entry-exit relation near planar turning point. J. Differ. Equ. **215**, 225–267 (2005)
2. Dumortier, F., Roussarie, R.: Canard cycles and centre manifolds. Mem. Am. Math. Soc. **121**(577), 1–100 (1996)
3. Dumortier, F., Roussarie, R.: Multiple canard cycles in generalized Liénard equations. J. Differ. Equ. **174**, 1–29 (2001)
4. Dumortier, F., Roussarie, R.: Canard cycles with two breaking parameters. Discrete Continuous Dyn. Syst. **17**(4), 787–806 (2007)
5. Dumortier, F., Roussarie, R.: Multi-layer canard cycles and translated power functions. J. Differ. Equ. **244**, 1329–1358 (2008)
6. Khovanskii, A.: Fewnomials. Translated from the Russian by Smilka Zdravkovska. Translations of Mathematical Monographs, vol. 88, viii + 139 pp. American Mathematical Society, Providence (1991)
7. Mahmoudi, L., Roussarie, R.: Canard cycles of finite codimension with two breaking parameters. Qual. Theory Dyn. Syst. **11**, 167–198 (2012)
8. Panazzolo, D.: Solutions of the equation $a_n + (a_{n-1} + \ldots (a_2 + (a_1 + x^{r_1})^{r_2} \ldots)^{r_n} = x$, oral communication (2015)

Chapter 5
On the Integrability of Lotka–Volterra Equations: An Update

Colin Christopher, Wuria M.A. Hussein, and Zhaoxia Wang

To Christiane—with thanks for being able to share in your mathematical interests in a small way.

Abstract In 2004, Christopher and Rousseau considered various results around the integrability of the origin for the Lotka–Volterra equations

$$\dot{x} = x(1 + ax + by), \qquad \dot{y} = y(-\lambda + cx + dy),$$

for rational values of λ. In particular, for $\lambda = p/q$ with $p + q \leq 12$, they showed that all the integrability conditions were given by either the Darboux method or a monodromy argument.

In this paper we consider the integrability of the critical points which do not lie at the origin. For those on one of the axes, we classify all integrable critical points with ratio of eigenvalues $-p'/q'$ with $p' + q' \leq 17$; and for those not on the axes, we consider all critical points with ratio of eigenvalues $-p''/q''$ with $p'' + q'' \leq 10$. We also extend the classification of integrable critical points at the origin for $p + q \leq 20$.

In all these cases, we are able to show that the monodromy method is sufficient to prove integrability except when $\lambda ab + (1-\lambda)ad - cd = 0$, for which the system has

C. Christopher (✉)
School of Computing, Electronics and Mathematics, Faculty of Science and Engineering, Plymouth University, Plymouth, UK
e-mail: C.Christopher@plymouth.ac.uk

W.M.A. Hussein
Department of Mathematics, College of Science, University of Salahaddin, Kurdistan Region, Erbil, Iraq

Z. Wang
School of Mathematical Sciences, University of Electronic Science and Technology of China, Qingshuihe Campus, 2006 Xiyuan Avenue, West Hi-Tech Zone Chengdu, Sichuan 611731, P.R. China

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_5

an invariant line. However, to do this, we need to extend the monodromy method to include the monodromy about some of the invariant algebraic curves of the system as well as the axes.

Keywords Lotka-Volterra equations • Monodromy method • Integrable critical points • Invariant algebraic curves • Darboux method

5.1 Introduction

In [2] the authors considered various results around the integrability and linearizability of the origin for the Lotka–Volterra equations

$$\dot{x} = x(1 + ax + by), \qquad \dot{y} = y(-\lambda + cx + dy), \tag{5.1}$$

for rational values of λ.

In particular, for $\lambda = p/q$ with $p + q \leq 12$ they showed that all the integrability conditions were generated by two mechanisms. First, when

$$\lambda ab + (1 - \lambda)ad - cd = 0, \tag{5.2}$$

there is an invariant line $L = 0$. Using the classical theory of Darboux, a first integral of the form $x^p y^q L^\alpha$ could be found. Furthermore, the conditions for linearizability can be given explicitly.

Second, the integrability of the origin could be explained by a monodromy argument. That is, the ratio of eigenvalues of the critical points on the axes and the line at infinity implied that the monodromy map around the origin was linearizable, and hence the critical point was in fact integrable. When (5.2) does not hold, it was shown that the conditions for integrability automatically imply linearizability.

Several results for more general values of λ were given. In addition, by comparison with the results of Moulin-Ollagnier [5] on Liouvillian integrability of Lotka–Volterra systems, two exceptional cases were found when $\lambda = 8/7$ and $13/7$ which turned out to support invariant algebraic curves and were hence solvable by the Darboux method. Some further results were announced in [4] and [3].

Our aim in this paper is to extend this investigation to the critical points of (5.1) which do not lie at the origin. In particular, if the critical point with ratio of eigenvalues $-p/q$ lies on one of the axes (the "side" case) we have show that for $p + q \leq 17$ the critical point is integrable if either there is an invariant line passing through the point (and the system is reducible to the conditions found above via a projective transformation) or there is a monodromy argument involving the monodromy group of the axes and the line at infinity and possibly an invariant algebraic curve passing through the critical point.

If, on the other hand, the critical point does not lie on the axes (the "face" case) we have shown that for $p + q \leq 10$ the critical point is integrable if either there is

an invariant line passing through the point (and again the system is reducible to the conditions found in [2]) or there is a monodromy argument involving an invariant algebraic curve passing through the critical point.

We also return to the origin case considered in [2] and extend the classification to $p + q \leq 20$. We have shown that no new cases appear, and furthermore, that the two exceptional cases mentioned above can be considered as arising from monodromy arguments involving the invariant curves and the axes.

The paper is arranged as follows. In the next section we give a brief summary of the monodromy method in the form that we use here. We will also explain how we can extend the monodromy method to some of the invariant algebraic curves found by Moulin-Ollagnier. The geometric classification of these curves and their singularities is part of a more extensive investigation to be published elsewhere [1]. Here, however, we want to show that the monodromy method can still be applied in some cases where the invariant curve has singularities.

Finally, in Sect. 5.3, we state our results. Since the computation of integrability conditions is now a well-trodden area, we do not give extensive sets of conditions, but merely indicate the classes of systems involved and state the examples arising from invariant algebraic curves.

5.2 The Monodromy Method

Recall that a polynomial vector field,

$$\dot{x} = P(x, y), \qquad \dot{y} = Q(x, y),$$

gives rise to an analytic foliation (with singularities)

$$P(x, y)\, dy - Q(x, y)\, dx = 0.$$

Such a foliation extends in a natural way to $\mathbf{P}^2(\mathbb{C})$. If P and Q have degree n, then the line at infinity is invariant if $xQ_n - yP_n \neq 0$, where P_n and Q_n are the terms of highest degree in P and Q, respectively.

For the Lotka–Volterra system (5.1) we therefore have three invariant lines: the x and y-axes and the line at infinity. Since we work over $\mathbf{P}^2(\mathbb{C})$, these lines are really copies of the Riemann Sphere.

In the neighbourhood of each line we can consider the monodromy group as follows. We fix a family of transversals to the line which pass through all points on the line which are non-singular points of the vector field (call this set of points S). We also fix a point $p \in S$ on the line and denote its transversal by Σ.

For each closed path, γ, starting at p and each point, q, on Σ sufficiently close to p we can lift the path to a unique curve on the leaf of the foliation through q. On returning to Σ this curve will intersect Σ at a new point q'. If we are given a local parameter z for Σ with $z(p) = 0$, then the map from q to q' will define the

germ of a local analytic function $M_\gamma : (\mathbb{C}, 0) \mapsto (\mathbb{C}, 0)$. This map is called the monodromy map associated with the path γ. The map M_γ only depends on the path up to homotopy in S. Furthermore, a change in the transversal or its parameterization will give a monodromy map conjugate to the original one. Finally, the monodromy map for a composition of paths is just the compositions of the monodromies ($M_{\alpha\circ\beta} = M_\alpha \circ M_\beta$).

It is well known that there is a close connection between the conjugacy class of the monodromy map about a path surrounding a single critical point and the analytic classification of the critical point itself. In particular, a critical point which is of saddle type is integrable (i.e. can be orbitally linearized) if and only if the monodromy map is linearizable. The monodromy method consists of finding simple conditions which guarantee that the monodromy around a critical point is linearizable by considering the monodromy maps of the other points on the line.

Since we are working on the Riemann Sphere, the monodromy M_1 about a critical point, c_1, is just the inverse of the composition of the monodromies, M_k, about the other critical points, c_k, on the sphere. Thus, if M_2 is linearizable and M_k is the identity map for $k > 2$, then it is clear that M_1 is also linearizable and hence the critical point c_1 must be integrable.

The power of the method lies in the fact that in many cases it is easy to give conditions for a critical point to have identity or linearizable monodromy.

Consider a critical point on the Riemann Sphere whose ratio of eigenvalues is λ (calculated so that the eigenvalue associated with the tangential space of the Riemann Sphere forms the denominator). If λ is a positive rational number which is not an integer or the reciprocal of an integer, then the critical point is a linearizable node, and hence has linearizable monodromy.

In the case when λ or $1/\lambda$ is a positive integer, then the node may have at most one resonant term. If this resonant term is zero, the node must be linearizable as above. This can be established by a simple computation. However, in most cases the linearizability can be seen geometrically: if the node is resonant, then there is no analytic separatrix passing through the critical point tangent to the eigenvector with smaller (in absolute terms) eigenvalue. Thus, if λ is a positive integer greater than one, then the fact that Riemann Sphere itself is such a separatrix shows that the node must be resonant, and the monodromy just the identity. Similarly, if the node occurs at the crossing of two invariant lines, it must also be non-resonant.

What we have said about lines will also work for any smooth invariant curve of genus 0 (i.e. conics). More generally, since the monodromy only "sees" the branches of the curve, we can also apply the method to genus 0 curves whose singularities only have smooth branches (that is, the curve has at most ordinary multiple points).

If the curve has singularities, then a further investigation needs to take place. However, for our investigation we need only consider one such case: when the curve has a cusp and the associated vector field is non-degenerate at this point. In this case, the associated critical point at the cusp is a node with ratio of eigenvalues $2/3$. Since such a node is linearizable, we can locally find an analytic transformation bringing it to the form $\dot{x} = 2x$, $\dot{y} = 3y$ with invariant curve $y^2 = x^3$. The monodromy can be calculated directly from the parameterization of a loop around the critical

point, $(x, y) = (e^{2i\theta}, e^{3i\theta})$, which shows that the monodromy is in fact the identity. Alternatively, and more geometrically, the monodromy will be preserved under blowing up. If we blow up the cusp singularity, we get a smooth branch with ratio of eigenvalues 2. We can therefore conclude that the monodromy is the identity.

In what follows, we shall say that we can apply the monodromy method if all linearizability and identity monodromies are deduced in exactly the ways described above.

We now give the details of the five cases of invariant curve which appear in our classification. Since the integrability of a critical point is independent of a projective transformation of $\mathbf{P}^2(\mathbb{C})$, it is sufficient to present the examples in just one projectively equivalent configuration. That is, each example can be transformed by any projective transformation which maps the axes and the line at infinity between themselves.

5.2.1 Case A

Here, we have an invariant cubic curve (Case 11 in [5]). Figure 5.1 shows the geometric behaviour of the system in one choice of projective coordinates. The eigenvalue ratios of the critical points on the smooth branches of the curve are 2, 3, 6, and -8. There is also a cusp with eigenvalue ratio $2/3$ but, as explained above, this has identity monodromy. Thus, the critical point with eigenvalue ratio -8 must be integrable. It also follows that the critical point at P_4 must also be integrable by considering the monodromy on the x-axis.

5.2.2 Case B

Here, we have an invariant quartic curve (Case 15 in [5]). Figure 5.2 shows the geometric behaviour of the system in one choice of projective coordinates. The dotted curve represents complex branches through the critical points. The eigenvalue ratios of the critical points on the smooth branches of the curve are 2, 2, 3, 6, and -7. There is also a cusp with eigenvalue ratio $2/3$ but, as explained above, this has identity monodromy. Thus, the critical point with eigenvalue ratio -7 must be integrable. It follows that the critical point at P_4 must also be integrable by considering the monodromy on the line at infinity. Finally, the critical point at P_3 must also be integrable.

5.2.3 Case C

Here, we have an invariant conic (Case 4 in [5]). Figure 5.3 shows the geometric behaviour of the system in one choice of projective coordinates. The eigenvalue

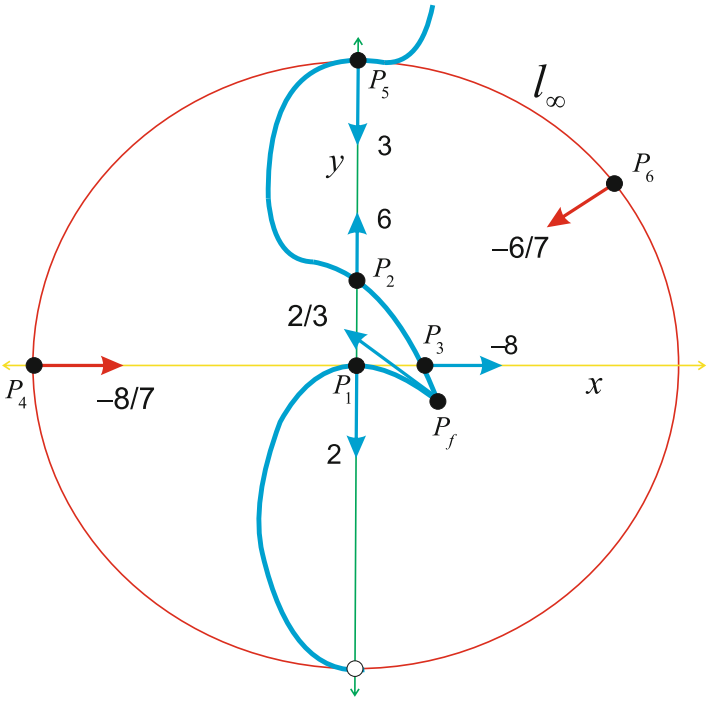


Fig. 5.1 Case A

ratios of the critical points on the smooth branches of the curve are 2, 2, α, and $-\alpha$. If α is positive but not the reciprocal of a positive integer, then the critical point at P_f has linearizable monodromy and hence the critical point at P_2 is integrable. Conversely, if $-\alpha$ is positive but not the reciprocal of a positive integer, then the critical point at P_2 has linearizable monodromy and hence the critical point at P_f is integrable.

5.2.4 Case D

Here, we have an invariant cubic curve (Case 5 in [5]). Figure 5.4 shows the geometric behaviour of the system in one choice of projective coordinates. The dotted curve represents complex branches through the critical points. The eigenvalue ratios of the critical points on the smooth branches of the curve are $1/2$, 2, 3, 3 and $-3/2$ (not a cusp). Thus the critical point with eigenvalue ratio $-3/2$ must be integrable.

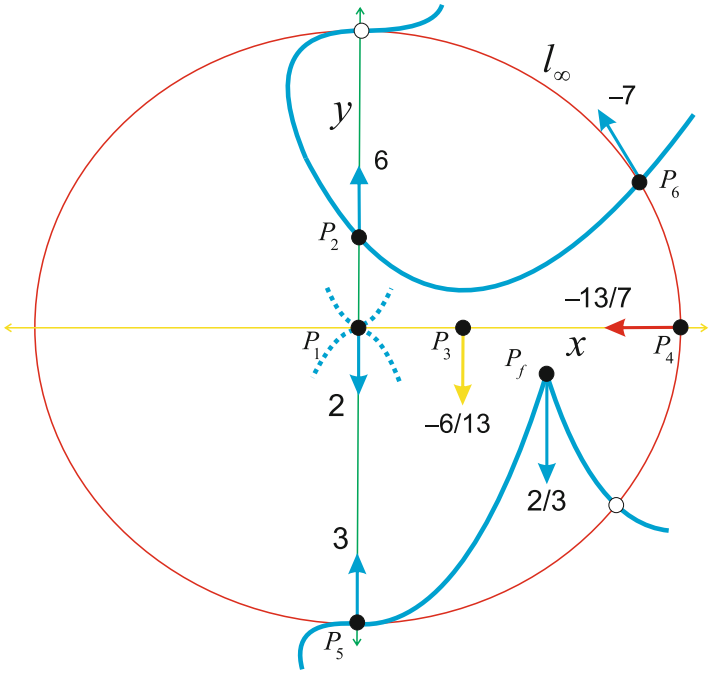


Fig. 5.2 Case B

5.2.5 Case E

Here, we have an invariant quartic curve (Case 6 in [5]). Figure 5.5 shows the geometric behaviour of the system in one choice of projective coordinates. The dotted curve represents complex branches through the critical points. The eigenvalue ratios of the critical points on the smooth branches of the curve are 1/3, 3, 2, 2, 4 and $-4/3$. Thus the critical point with eigenvalue ratio $-4/3$ must be integrable.

5.3 Results for Lotka–Volterra Systems

Now, we return to consider the Lotka–Volterra equation in $\mathbf{P}^2(\mathbb{C})$. On each invariant line (including the one at infinity) we have three critical points. If one of these critical points has identity monodromy and the other monodromy is linearizable, we can conclude that the third critical point also has linearizable monodromy and is hence integrable.

In more elaborate cases we might need to iterate this construction. That is, we apply the monodromy method on a line to show that a certain critical point has

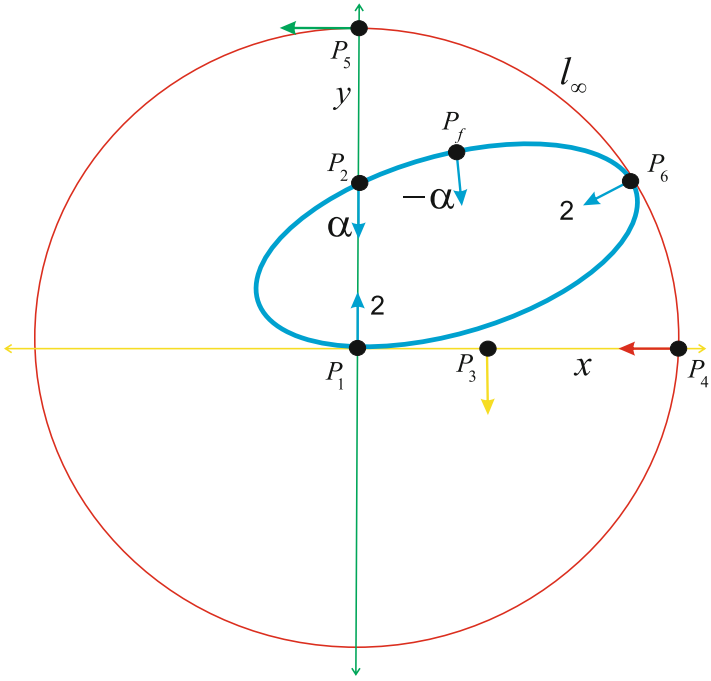


Fig. 5.3 Case C

linearizable monodromy, and then use this knowledge to apply the monodromy method on a second line on which the critical point lies. In the "side" case mentioned below, a third iteration is sometimes needed.

We now describe our results. We will split our consideration into three cases. The first considers the integrability of a saddle critical point at the origin. This is the case considered in [2]. For each $p, q > 0$ with $p + q \leq 20$, we take the general Lotka–Volterra system with a saddle of ratio of eigenvalues $-p/q$ and calculate the first three resonant terms of the normal form. From these calculations we obtain necessary conditions for integrability of the saddle. We then prove the sufficiency of these conditions: either by showing that (5.2) holds and hence there is a Darboux first integral, or by establishing that the critical point is integrable by a monodromy argument. The same technique is then applied to the cases where the critical point considered is on an axis but not at the origin and finally the case where the critical point is not on either axis. Our results are given below.

As an indicative example of the monodromy arguments used we consider the following case, which falls under Theorem 2 (2),

$$\dot{x} = x(1 - x + 9y), \qquad \dot{y} = (-7/12 + x + 7y). \tag{5.3}$$

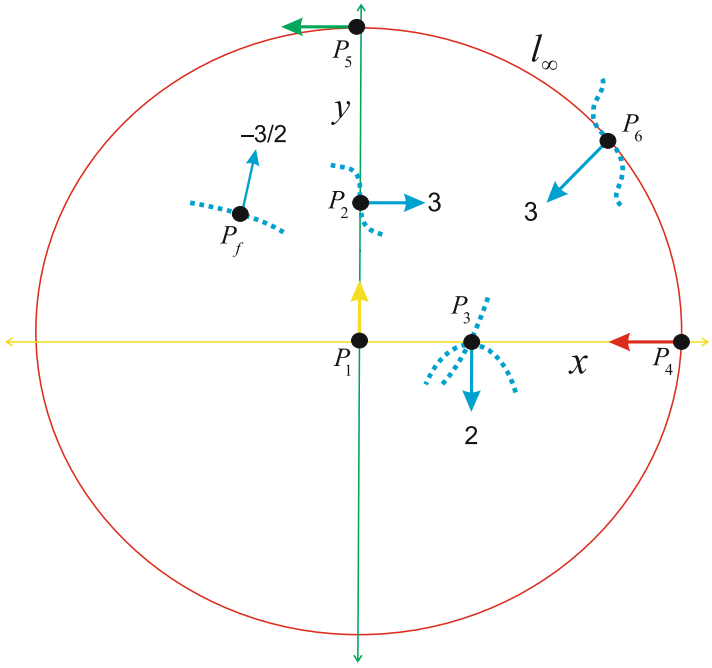


Fig. 5.4 Case D

The line at infinity has ratio of eigenvalues $1/2$, 4 and $-7/2$. The critical point with eigenvalue ratio $1/2$ also lies on the x-axis and therefore has two analytic separatrices and hence has linearizable monodromy. The critical point with eigenvalue ratio 4 clearly has identity monodromy and hence the critical point with eigenvalue ratio $-7/2$ must be integrable. This critical point also lies on the y-axis. With respect to this axis it has eigenvalue ratio $-2/7$, and the other two critical points have eigenvalue ratios of 3 and $-12/7$. Thus, the critical point with eigenvalue ratio $-12/7$, which lies at the origin, must also be integrable. Finally, along the x-axis, the critical point has ratio of eigenvalues $-7/12$ (the origin), $-5/12$ and 2 (at infinity). This shows that the critical point with eigenvalue ratio $-5/12$ on the x-axis is integrable.

Theorem 1. *If a Lotka–Volterra system has an integrable saddle at the origin with ratio of eigenvalues $-p/q$ with $p + q \leq 20$, then it falls into one of the following categories:*

1. *The condition (5.2) holds and the system has a Darboux first integral.*
2. *The monodromy method can be applied using one or two of the axes of the system.*
3. *We have $p/q = 8/7$ or $7/8$ and we use the monodromy method about the invariant cubic described in Case A.*
4. *We have $p/q = 13/7$ or $7/13$ and we use the monodromy method about the invariant quartic described in Case B.*

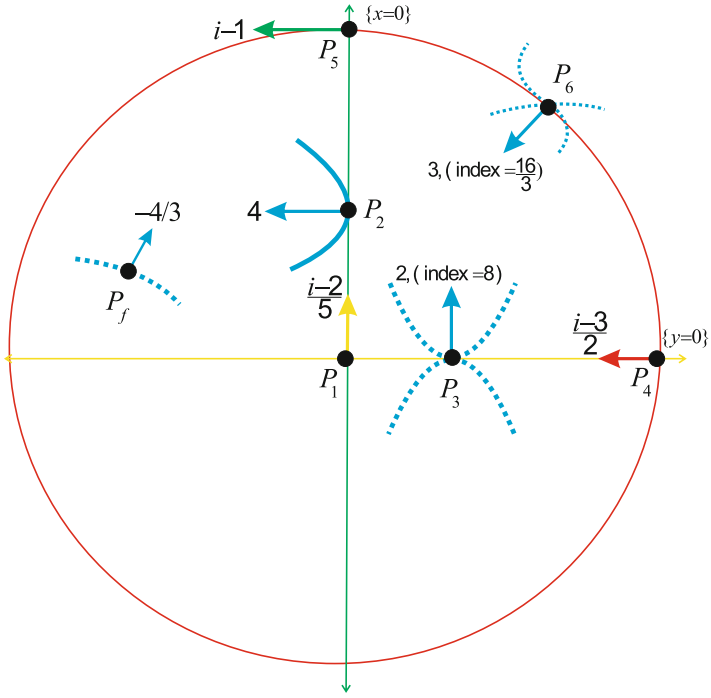


Fig. 5.5 Case E

Remark 1. 1. A more detailed enumeration of all the possible types of monodromy argument can be found in [2], where they were given the names $(A_n), \ldots, (H_{n,m})$. The same list is valid here.
2. The final two cases were found in [2], where it was shown that they gave Darboux first integrals. However, in the light of the other results below, it is probably better to consider them as examples of the monodromy method.

The second case is where the critical point lies on one of the invariant lines but not at the origin. We proceed as above and find the following result.

Theorem 2. *If a Lotka–Volterra system has an integrable saddle on one of its axes but not at the origin with ratio of eigenvalues $-p'/q'$ with $p' + q' \leq 17$ then it falls into one of the following categories:*

1. *The condition (5.2) holds and the system has a Darboux first integral.*
2. *The monodromy method can be applied using one, two, or three of the invariant lines of the system.*
3. *The monodromy method can be applied using the invariant conic given in Case C ($\alpha < 0$).*
4. *We have $p'/q' = 1/8$ or 8 and we use the monodromy method about the invariant cubic described in Case A.*

5. *We have $p'/q' = 1/7$ or 7 (or $6/13$ or $13/6$) and we use the monodromy method about the invariant quartic described in Case B.*

Remark 2. 1. It is possible to need monodromy arguments around all three invariant lines, as shown in the system (5.3) above. This also can arise because a critical point on two axes with an integer ratio of eigenvalues contributes an identity monodromy to one axis, but only a linearizable monodromy to the other axis.
2. It will be seen from Fig. 5.2 that the ratio of eigenvalue $6/13$ or $13/6$ is also a possibility. We have included it in the statement of the theorem although it lies outside the $p' + q' \leq 17$ threshold.

Finally, we consider the case where the saddle is not on the axes. In this case we find the following result.

Theorem 3. *If a Lotka–Volterra system has an integrable saddle which does not lie on one of its axes with ratio of eigenvalues $-p''/q''$ with $p'' + q'' \leq 10$, then it falls into one of the following categories:*

1. *The condition (5.2) holds and the system has a Darboux first integral.*
2. *The monodromy method can be applied using the invariant conic given in Case C ($\alpha > 0$), or by using an invariant line passing through the critical point.*
3. *We have $p''/q'' = 3/2$ or $2/3$ and we use the monodromy method about the invariant cubic described in Case D.*
4. *We have $p''/q'' = 3/4$ or $4/3$ and we use the monodromy method about the invariant quartic described in Case E.*

Remark 3. We believe that the results of Theorems 1, 2 will hold for all rational ratios of eigenvalues but were not able to establish this yet. We also conjecture the same is true for Theorem 3. However, for higher values of $p'' + q''$ it is necessary to add special cases for $p''/q'' = 6/5,\ 5/6,\ 6/7$ and $7/6$ which arise from invariant curves (cases 7, 10, 12, 13 and 14 of [5]).

References

1. Christopher, C., Hussein, W.: A geometric approach to Moulin-Ollagnier's classification of algebraic solutions of Lotka-Volterra systems. Preprint, (2015)
2. Christopher, C., Rousseau, C.: Normalizable, integrable and linearizable saddle points in the Lotka-Volterra system. Qual. Theory Dyn. Syst. **5**, 11–61 (2004)
3. Gravel, S., Thibault, P.: Integrability and linearizability of the Lotka-Volterra system with a saddle point with rational hyperbolicity ratio. J. Differ. Equ. **184**, 20–47 (2002)
4. Liu, C., Chen, G., Li, C.: Integrability and linearizability of the Lotka-Volterra systems. J. Differ. Equ. **198**, 301–320 (2004)
5. Moulin-Ollagnier, J.: Liouvillian integration of the Lotka-Volterra systems. Qual. Theory Dyn. Syst. **2**, 307–358 (2001)

Chapter 6
Impact of Pharmacokinetic Variability on a Mechanistic Physiological Pharmacokinetic/Pharmacodynamic Model: A Case Study of Neutrophil Development, PM00104, and Filgrastim

Morgan Craig, Mario González-Sales, Jun Li, and Fahima Nekka

Abstract Interindividual variability (IIV) is considered a crucial factor for the general use of mathematical modelling in physiology. However, mechanistic models of physiological systems are commonly built for an average patient, raising the question of their applicability at the population level. Using our previously developed physiological model of neutrophil regulation, which accounts for the detailed hematopoietic mechanisms as well as the pharmacokinetics (PKs) of a chemotherapeutic agent (PM00104) and a granulostimulant (filgrastim), we incorporated the reported population pharmacokinetic (PopPK) models of each drug to investigate the impact of PK variability on fully mechanistic models. A variety of scenarios, including multiple doses of PM00104, were simulated for cohorts of 500 in silico patients to analyse the model's predictability in terms of several pharmacological indicators, such as the time to neutrophil nadir, the value of the nadir, and the area under the effect curve. Our results indicate the robustness of our model's predictions in all considered scenarios. Based on our findings, we conclude that for drugs with short-lived PKs in comparison with their pharmacodynamics (PDs), models that "sufficiently" account for physiological mechanisms inherently assimilate PK deviations, making the further inclusion of PK variability unnecessary.

Keywords Physiological modelling • Interindividual variability • Neutropenia • Pharmacometrics

M. Craig (✉) • J. Li • F. Nekka
Faculté de Pharmacie, Université de Montréal, Montréal, QC, Canada H3T 1J4
e-mail: morgan.craig@umontreal.ca; li@crm.umontreal.ca; fahima.nekka@umontreal.ca

M. González-Sales
Faculté de Pharmacie, Université de Montréal, Montréal, QC, Canada H3T 1J4

InVentiv Health Clinical, 5160 Décarie Blvd., Suite 800, Montréal, QC, Canada H3X 2H9
e-mail: mario.gonzalez.sales@umontreal.ca

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_6

6.1 Introduction

One of the most important considerations in modern pharmacometrics is the determination of the dose–response relationship. This can be obtained through data-driven models [13, 14, 18] but can also be achieved through techniques stemming from mathematical biology using physiologically driven mechanistic modelling [2, 5, 6, 9, 10, 12]. While both approaches use mathematics to describe the disposition of drugs in the body, each handles the problem from a unique vantage point. On the one hand, data-driven models are based on various components: the structural model (typically compartmental), a set of statistical models (with assumptions for probabilistic distributions around model parameters as well as error structure), and covariate models (when relevant). To this end, population pharmacokinetic/pharmacodynamic (Pop-PK/PD) modelling is the most representative approach and is now widely used in drug research and development. Even so, the construction of data-driven models highly depends on the available data and the required statistical optimisation procedures. The poor quality of data can induce model misspecification and hamper the generalisability of the models outside of the context in which they were built.

On the other hand, mechanistic models of human processes tackle the problem using the so-called bottom-up strategy. These models are constructed directly from the system being studied by applying the available physiological knowledge to drive their predictions. During this process, a number of hypotheses are generated and translated using various mathematical techniques. Generally, the model parameters are derived or estimated from experimentally determined values available from a diverse cross-section of fields (physiological, chemical, physical, etc.), and utilise patients' average values [3]. Physiological models are used to explore a variety of complex situations. For example, the model of neutrophil development studied herein can be applied in oncological settings or in the study of hereditary disorders like cyclical neutropenia. Mechanistic models are useful for explaining an observed effect in relation to its components as a result of their physiologically detailed construction. However, constructing a physiological model can be time-consuming and requires advanced mathematical knowledge to ensure the models' validity. This complexity makes this approach more popular in academia but work still needs to be done to expand its use in routine data analysis.

Since physiological models are frequently developed for an average patient, an investigation of the impact of PK variability on their predictions is crucial to extend their applicability to patient populations. Considering the complexity of these models, testing their robustness by simulating credible scenarios of patient variability will determine their suitability for general use. For instance, it is pertinent to know whether drug regimens identified for an average patient through these models can be extended to the population level. We have previously published a physiological model of granulopoiesis [6], with integrated PK models of both PM00104 [18] and filgrastim (adjuvant recombinant human granulocyte

colony-stimulating factor (G-CSF), used to increase neutrophil counts to prevent and/or recover from neutropenia) [14], which studied the impact of the time of administration of supportive filgrastim during chemotherapy. Using published or well-derived parameters from the literature for an average patient, our model successfully predicted clinical data and identified beneficial regimens. Therein, we determined that delaying the administration of rhG-CSF after PM00104 by 7 days mitigated the neutropenic impact of anti-cancer treatment, resulting in a reduction from ten administrations per cycle to 3 or 4 and a reduction in the burden to the patient [6]. The current study will address the extendibility of this regimen to a population by investigating the impact of PK interindividual variability (IIV), and of the reported interoccasion variability (IOV) of PM00104, on relevant indicators, such as the time to neutrophil nadir and the nadir level.

From a systems pharmacology point of view, the PKs of the previously mentioned drugs are short-lived in comparison with their PDs. Indeed, the PK half-lives of both PM00104 and filgrastim are on the order of hours (24 and from 6–10, respectively) whereas it can take several days to observe their effects on cells in circulation due to the production time of neutrophils in the bone marrow (up to 14 days) [6, 18]. In the current work, we focused on the impact that IIV components of the data-driven models of [14, 18] can have on our physiological model [6]. Based on numerical simulations, the sensitivity of the physiological model to the impact of IIV was quantified and statistically analysed. Using a variety of scenarios that cover a large number of clinical situations, our physiological PK/PD model, though developed for an average patient, proved to be robust in terms of PK IIV when clinically relevant PD criteria are tested, advocating its general applicability to a large population.

6.2 A Hypothesis-Driven Physiological/PK/PD Model of Granulopoiesis During Chemotherapy with Supportive Adjuvant

A mechanistic physiological model of myelopoiesis was constructed [6] by extending the previous work of Brooks et al. [2], Colijn and Mackey [5], Foley and Mackey [9], and Foley et al. [10] through the addition of neutrophil reservoir pools in the bone marrow and other tissues, and then subsequently incorporating comprehensive PK/PD models for PM00104 (Zalypsis®), a chemotherapeutic drug, and filgrastim (rhG-CSF), a supportive adjuvant, to determine dosing schemes that provide the most benefit (least harm) for patients. The physiological model translates the physiological mechanisms of neutrophil production mathematically using delay differential equations (DDEs) to characterise the cellular transition delays.

The neutrophil model is a three-dimensional set of DDEs with variable aging rate and general delays obtained from an age-structured partial differential equation

model with appropriate boundary conditions. A schematic diagram of the model is given in Fig. 1 in [6]. Beginning in a quiescent state, a hematopoietic stem cell (HSC-population $Q(t)$ in units 10^6 cells/kg), which is capable of self-renewal at rate $\beta(Q)$ (in units days^{-1}) and which is subject to apoptosis at rate γ_S (in units days^{-1}), undergoes differentiation into one of three blood cell lines. In this model, we consider any differentiation into the erythrocyte or platelet lineages to occur at rate κ_δ (in units days^{-1}) whereas differentiation into the neutrophil line occurs at rate $\kappa_N(N)$ (in units days^{-1}). Note that while κ_δ is taken herein to be constant, the rate of entry of the HSCs into the neutrophil lineage depends on the concentration of circulating neutrophils (population $N(t)$ in units 10^9 cells/kg). Once committed to the neutrophil line, cells undergo proliferation—a period of successive divisions—at a rate of η_{NP} (in units days^{-1}) for a total of τ_{NP} days. Cells then cease division and mature at a velocity of V_N (in units days^{-1}) for a total of $\tau_{NM}(t)$ days. During this maturation period, cells are subject to random cell death at a rate of γ_{NM} (in units days^{-1}). Newly mature neutrophils are then sequestered within the bone marrow in the mature neutrophil reservoir (population $N_r(t)$ in units 10^9 cells/kg). These reserved cells are mobilised from the bone marrow into circulation at a rate of $f_{trans}(G(t))$ (in units days^{-1}) or, failing to reach the circulation, die from the reservoir at rate γ_r (in units days^{-1}). The mature pool is a crucial aspect of the neutrophil lineage, as it contains ten times the number of circulating neutrophils and is necessary for the rapid restocking of the blood neutrophils in case of falling ANCs or infection [11, 21]. Cells reaching the circulation subsequently disappear from the blood at a rate of γ_N (in units days^{-1}). Beginning with a quiescent hematopoietic stem cell (HSC) differentiating into the neutrophil lineage, we model the proliferation and maturation of neutrophilic cells in the bone marrow. The mature neutrophils then settle into the marrow reservoir before appearing in circulation (release from the reservoir can be steady, or homeostatic, or rapid mobilisation in the case of emergency). Once a mature neutrophil reaches the circulation, it disappears fairly rapidly (half-life of around 7 h) through apoptosis or margination into the tissues. Equations (6.1)–(6.3) below highlight the primary model equations.

$$\frac{dQ(t)}{dt} = -\left(\kappa_N(N(t)) + \kappa_\delta + \beta(Q(t))\right) Q(t) + A_Q(t)\beta\left(Q(t-\tau_S)\right) Q(t-\tau_S) \tag{6.1}$$

$$\frac{dN_r(t)}{dt} = A_N(t)\kappa_N(N(t-\tau_N))Q(t-\tau_N)\left(\frac{V_N(G(t))}{V_N(G(t-\tau_{NM}(t))}\right) \tag{6.2}$$

$$\frac{dN(t)}{dt} = f_{trans}(G(t))N_r(t) - \gamma_N N(t). \tag{6.3}$$

Delays are indicated by $t-\tau$, where τ is a physiologically present delay in the system (time of HSC self-renewal, time of proliferation, time of maturation, time of residence in the marrow reservoir, and the total time it takes to produce a neutrophil

from differentiation to appearance in the circulation). Equations (6.1)–(6.3) are subject to the initial condition of homeostasis ($Q(t) = Q^{homeo}$, $N_r(t) = N_r^{homeo}$, $N(t) = N^{homeo}$, for all $t \leq t_0$, where t_0 marks the beginning of treatment). In [6], parameter estimation for an average patient was carried out in a consistent way using data available in the literature. Typical values of the PK models of PM00104 and G-CSF were adapted from [18] and [14], respectively. In this study, IIV and IOV components of the PK parameters were added where necessary. All parameter values in the current work were kept as in [6] and any exceptions will be indicated explicitly below. Particular attention was paid to capturing the dominant processes implicated in the development of a circulating neutrophil within the bone marrow.

As in our previous work, multiplying $N(t)$ by the fraction of circulating cells is necessary for comparison to data since $N(t)$ herein represents the total blood neutrophil pool (TBNP). The HSC's feedback rate and amplification rates of both the HSCs and the blood neutrophils are modelled as

$$\beta(Q) = f_Q \frac{\theta_2^{s_2}}{(\theta_2^{s_2} + Q^{s_2})} \tag{6.4}$$

$$\kappa_N(N) = f_N \frac{\theta_1^{s_1}}{(\theta_1^{s_1} + N^{s_1})} \tag{6.5}$$

$$A_Q(t) = 2 \exp\left[- \int_{t-\tau_S}^{t} \gamma_S(s) ds \right] \tag{6.6}$$

$$A_N(t) = \exp\left[\int_{t-\tau_N(t)}^{t-\tau_N(t)+\tau_{NP}} \eta_{NP}(s) ds - \int_{t-\tau_N(t)+\tau_{NP}}^{t} \gamma_{NM}(s) ds \right]. \tag{6.7}$$

The entire process of neutrophil development is regulated by the concentration of G-CSF, $G(t)$ in units ng/ml, which acts in negative feedback with the blood neutrophil numbers in that its concentration falls when neutrophil numbers increase and vice versa. G-CSF acts during the entire neutrophil development cycle to maintain neutrophil counts at homeostatic levels. It is implicated in the recruitment of HSCs into the neutrophil line, in the regulation of the rates of proliferation and maturation, and controls the release of mature neutrophils from the bone marrow reservoir into circulation [9]. Details on the PD effects modelled herein are given in Fig. 6.1 and in the following sections.

6.3 Pharmacokinetics and Pharmacodynamics of PM00104

The pharmacokinetics of PM00104 were characterised using a catenary four compartment disposition model with linear elimination [18]. The differential equations describing the system were as follows:

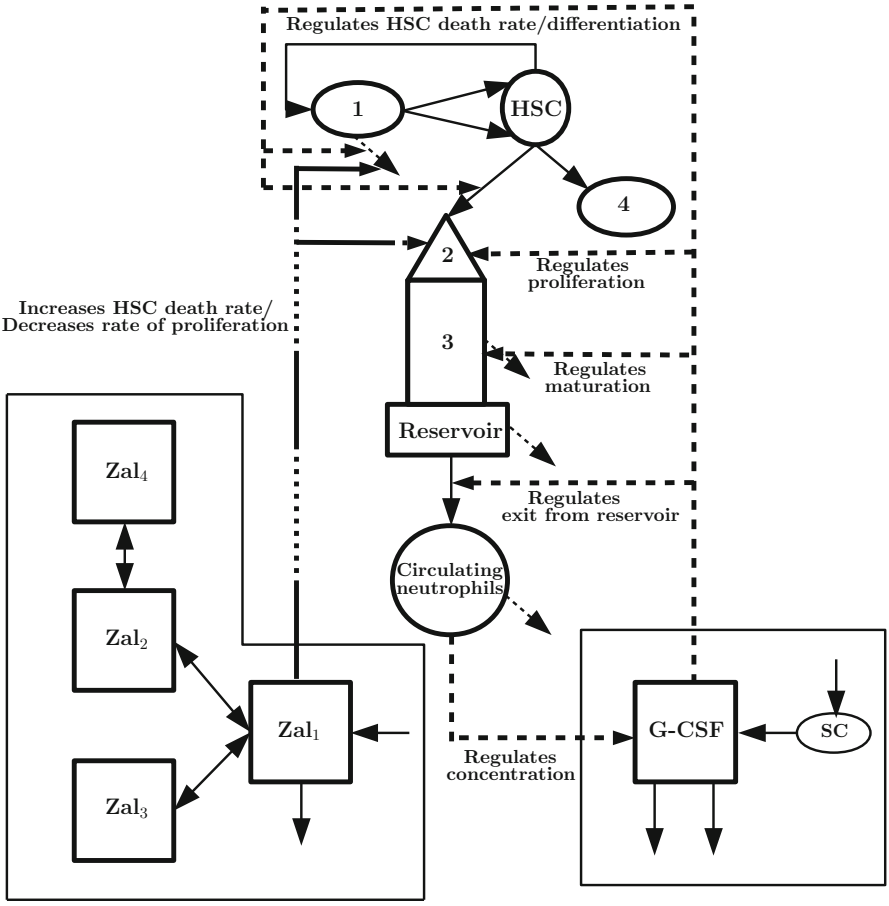


Fig. 6.1 Schematic representation of the effects of PM00104 (*solid/dotted lines*) and G-CSF (*dashed*). Model summary as in Fig. 1 in [6] in Sect. 6.2. PM00104 (Zal) acts to disrupt cellular division resulting in a higher rate of death out of the hematopoietic stem cell (HSC) compartment, and a decrease in the rate of neutrophil proliferation in the bone marrow. G-CSF acts throughout the neutrophil lineage to stabilise the numbers of circulating neutrophils by regulating their exit out of the marrow reservoir, their proliferation, and their maturation. Concurrently, it acts upon the HSCs by regulating their differentiation into the neutrophil lineage and their death rate (to stabilise their population numbers). The complete model is taken from [6]. HSC: hematopoietic stem cells at rest, 1: dividing HSCs, 2: proliferating marrow neutrophils, 3: maturing marrow neutrophils, 4: other blood cell lines, Reservoir: mature marrow neutrophil reservoir, Zal_1: central compartment of PM00104, Zal_2: second compartment of PM00104, Zal_3: third compartment of PM00104, Zal_4: fourth compartment of PM00104, G-CSF: granulocyte colony-stimulating factor, SC: subcutaneous pool

$$\frac{dA_1}{dt} = -\left(\frac{CL}{V_1} + \frac{Q_2}{V_1} + \frac{Q_3}{V_1}\right)A_1 + \frac{Q_2}{V_2}A_2 + \frac{Q_3}{V_3}A_3 \tag{6.8}$$

$$\frac{dA_2}{dt} = -\frac{Q_2}{V_2}A_2 - \frac{Q_4}{V_2}A_2 + \frac{Q_2}{V_1}A_1 + \frac{A_4}{V_4}A_4 \tag{6.9}$$

$$\frac{dA_3}{dt} = -\frac{Q_3}{V_3}A_3 + \frac{Q_3}{V_1}A_1 \tag{6.10}$$

$$\frac{dA_4}{dt} = -\frac{Q_4}{V_4}A_4 + \frac{Q_4}{V_2}A_2, \tag{6.11}$$

where A_1, A_2, A_3 and A_4 represent the amount of PM00104 in compartments 1, 2, 3, and 4, CL represents the clearance (in units L/h), Q_2 is the intercompartmental clearance between compartments 1 and 2 (in units L/h), Q_3 is the intercompartmental clearance between compartments 1 and 3 (in units L/h), Q_4, is the intercompartmental clearance between compartments 2 and 4 (in units L/h), V_1 is the volume of distribution in the central compartment (in units L), V_2 is the volume of distribution in compartment 2 (in units L), V_3 is the volume of distribution in compartment 3 (in units L), and V_4 is the volume of distribution in compartment 4 (in units L). Concentrations in each compartment are given by dividing the amount of the drug A_j $(j = 1, \ldots, 4)$ by the volume in the respective compartment V_j $(j = 1, \ldots, 4)$.

The above model and results of González-Sales et al. [13] and Pérez-Ruixo et al. [18] were incorporated into [6] in a physiologically consistent way. The main function of a chemotherapeutic agent is to quell the uncontrolled division of cells by reducing/destroying their ability to replicate. In the blood system, the effects of this reproductive cessation are assumed to be twofold: first, the HSCs experience an increase in the rate of cell death in the proliferative phase (effectively reducing their proliferative capabilities) and second, the rate at which the neutrophils undergo successive divisions is greatly reduced. These two effects are modelled, respectively, as

$$\gamma_S^{chemo}(C_p(t)) = \gamma_S^{homeo} + h_S C_p, \tag{6.12}$$

where C_p is the concentration of PM00104 in the first, or plasmatic, compartment, γ_S^{chemo} is the rate of apoptosis of the proliferative HSCs during chemotherapy, γ_S^{homeo} is their rate of apoptosis at homeostasis. Due in large part to the absence of data relating the effects of chemotherapy on the HSCs, we took a linear effect from chemotherapy upon γ_S, modulated by the effects parameter h_S, and

$$\eta_{NP}^{chemo}(C_p(t)) = \eta_{NP}^{homeo}\left(\frac{(EC_{50})^h}{(EC_{50})^h + \left(C_p(t)\right)^h}\right). \tag{6.13}$$

Here η_{NP}^{chemo} is the rate of neutrophil proliferation during chemotherapeutic treatment, η_{NP}^{homeo} is the homeostatic rate of proliferation, EC_{50} is the usual half-effect constant and h is the Hill coefficient of the effect.

6.4 Pharmacokinetic and Pharmacodynamic Model of G-CSF

As previously alluded to, G-CSF is an endogenous cytokine which stimulates the production of neutrophils. It is also used in an exogenous form as an adjuvant to help patients with low neutrophil counts rescue their circulating ANCs [9]. In terms of its PK properties, G-CSF is believed to have a constant production rate [14] and to have two modes of elimination, namely an unsaturable process from renal elimination, and a saturable process driven by internalisation by the neutrophils [15]. Accordingly, as in [6], we model the endogenous concentration of G-CSF as:

$$\frac{dG}{dt} = G_{prod} - k_{ren}G - \chi k_{int}\frac{G^2}{G^2 + k_d^2}N, \tag{6.14}$$

where G_{prod} is the endogenous constant production rate of G-CSF (in units ng/ml/day, k_{ren} is the rate of renal elimination (in units days^{-1}), k_{int} is the rate of internalisation by the neutrophils (in units days^{-1}), k_d is the usual dissociation constant (in units ng/ml), and χ is a scaling factor to correct for the units of Eq. (6.14). The choice of Hill coefficient is due to the 2:2 stoichiometry of G-CSF binding to its G-CSFR receptor on the neutrophils [16]. When G-CSF is given exogenously, primarily in subcutaneous form, we model its administration as in [6], and originally in [14] as

$$\frac{dG}{dt} = \frac{F(Dose)k_a}{V_d}e^{-k_a t} + G_{prod} - k_{ren}G - \chi k_{int}\frac{G^2}{G^2 + k_d^2}N, \tag{6.15}$$

where F is the bioavailable fraction, $Dose$ is the administered subcutaneous dose (in ng), k_a is the rate of absorption from the subcutaneous pool (in units days^{-1}), and V_d is the volume of distribution (in units ml).

The pharmacodynamic action of G-CSF is multifaceted. From the beginning of a neutrophil as a stem cell, G-CSF reduces the rate of cell death in the proliferating HSC compartment (decreasing γ_S), increases the rate of neutrophil proliferation in the bone marrow (increases η_{NP}), increases the speed of neutrophil maturation (increases $V_N(N)$ or, equivalently, decreases τ_{NM} [21]), decreases neutrophil death out of the marrow maturation compartment (decreases γ_{NM}), and modulates the rate of transfer between the mature neutrophil reservoir and the circulation in function of the ANC (modulates f_{trans}) [2, 9, 14, 20]. These effects are modelled as follows, with $b_i,\ i = S, N, NP, V$ the parameters relating the half-maximal concentration of G-CSF. In the HSC compartment,

$$\gamma_S(G(t), C_p(t))) = \gamma_S^{min} - \frac{(\gamma_S^{min} - \gamma_S^{chemo})b_S}{G(t) - G^{homeo} + b_s}, \tag{6.16}$$

where γ_S^{min} is the minimal rate of apoptosis in the HSCs proliferative phase. Note that the rate of cell death of the HSCs is dependent both on the concentration of G-CSF and on the concentration of the chemotherapeutic agent. The details of the latter dependency are given in Eq. (6.18) below.

Neutrophils undergoing proliferation are also subject to the effects of chemotherapeutic drugs (Eq. (6.13) and details above) and to the concentration of G-CSF

$$\eta_{NP}(G(t), C_p(t)) = \eta_{NP}^{chemo}(C_p(t)) + \frac{(\eta_{NP}^{max} - \eta_{NP}^{chemo}(C_p(t)))(G(t) - G^{homeo})}{G(t) - G^{homeo} + b_{NP}} \tag{6.17}$$

$$\gamma_{NM}(G(t)) = \gamma_{NM}^{min} - \frac{(\gamma_{NM}^{min} - \gamma_{NM}^{homeo})b_{NM}}{G(t) - G^{homeo} + b_{NM}}. \tag{6.18}$$

Here η_{NP}^{max} is the maximal proliferation rate of the neutrophils and γ_{NM}^{min} is the minimal rate of random cell loss of the maturing neutrophils.

When G-CSF concentrations are high, the speed with which the neutrophils in the marrow age increases, thereby decreasing the time they spend maturing. These simultaneous effects are given by

$$V_N(G(t)) = 1 + (V_{max} - 1)\frac{G(t) - G^{homeo}}{G(t) - G^{homeo} + b_V}, \tag{6.19}$$

where V_{max} is the maximal aging velocity of the maturing neutrophils, and

$$\frac{d\tau_N(t)}{dt} = \frac{d\tau_{NM}(t)}{dt} = 1 - \frac{V_N(G(t))}{V_N(G(t - \tau_{NM}(t)))}. \tag{6.20}$$

The details of the derivation of Eq. (6.20) are given in full in [6].

Finally, the recruitment of a reserved neutrophil to the blood given as

$$f_{trans}(G(t)) = trans^{homeo}\frac{trans^{ratio}(G(t) - G^{homeo}) + b_G}{G(t) - G^{homeo} + b_G}, \tag{6.21}$$

where $trans^{homeo}$ relates the homeostatic rate of transit from the neutrophil bone marrow reservoir into the circulation, and $trans^{ratio} = \frac{trans^{max}}{trans^{homeo}}$ is an empirically determined ratio modulating the fraction of neutrophils released from the reservoir [22, 25].

6.5 Incorporating Variability into the Physiological PK/PD Model

By incorporating the PK variability reported for PM00104 and filgrastim in [18] and [14] into our average patient model [6], we now study the impact of PK variability on our model's predictions. In the case of PM00104, the fixed effects were assumed to follow a lognormal distribution according to the following equation:

$$P_{j,k} = P^* e^{\eta_j \tau_k}, \tag{6.22}$$

Table 6.1 Summary of the PopPK model parameters of PM00104 reported in [18]

Parameter (units)	Interpretation	Estimate	%RSE
Fixed effect		θ	
Cl (L/h)	Clearance	43.7	3.43
V_1 (L)	Volume of central compartment	32.7	12.4
Q_2 (L)	Transit rate (compartments 1 and 2)	123	5.76
V_2 (L)	Volume of second compartment	162	8.33
Q_3 (L/h)	Transit rate (compartments 1 and 3)	11.3	13.2
V_3 (L)	Volume of third compartment	388	11.8
Q_4 (L/h)	Transit rate (compartments 2 and 4)	62.3	9.00
V_4 (L)	Volume of fourth compartment	239	9.00
Interindividual variability		*CV%*	
Cl	IIV of Cl	34.1	24.6
V_1	IIV of V_1	82.5	37.7
V_2	IIV in V_2	65.1	41.8
Q_3	IIV of Q_3	87.7	31.2
V_3	IIV of V_3	52.0	25.2
(Cl, V_2)	Correlation between Cl and V_2	0.555	78.6
(Cl, Q_3)	Correlation between Cl and Q_3	0.572	33.6
(V_2, Q_3)	Correlation between V_2 and Q_3	0.522	84.0
Interoccasion variability		*CV%*	
Cl	IOV of Cl	14.1	96.0

Interindividual variability (IIV), correlations between interindividual random effects, and interoccasion variability (IOV) were reported as percentage of coefficient of variation (CV)

where $P_{j,k}$ is the individual jth PK parameter for the kth occasion, P^* is the typical value of the parameter of interest, and η_j and τ_k are independent and normally distributed interindividual and interoccasion (IOV) random variables with zero-mean and variance ω_p^2 and π_p^2, respectively. The magnitude of the IIV and the IOV were expressed as coefficient of variation (CV). The parameter estimates and their associated precisions, measured as percentage of relative standard error (%RSE), are presented in Table 6.1.

The filgrastim model used in this study is given by Eq. (6.15) in Sect. 6.4. Table 6.2 summarises the estimated model parameters of filgrastim and their associated precisions, expressed as %RSE.

As we were primarily concerned with PK variability, and owing to the differences in the current model's PD effects [Eqs. (6.16)–(6.21)], the IIV reported by the SC_{50}, S_{max_1}, and NB_0 parameters were not considered in the present analysis.

Table 6.2 Summary of the PopPK/PD model parameters of filgrastim, adapted from [14]

Parameter (units)	Interpretation	Estimate	%RSE
Fixed effect		θ	
k_{el} (h^{-1})	Rate of renal elimination	0.152	16.6
V_d (L)	Volume of distribution	2.42	6.8
ξ (fg/cell)	Proportionality constant ([GCSFR] per neutrophil)	0.181	45.5
NB_0 (cells/—L)	Initial number of blood neutrophils	1.55	17.9
SC_{50} (ng/ml)	Serum concentration eliciting 50 % of the maximal effect	3.15	21.0
S_{max_1}	Maximum effect	34.7	36.0
Interindividual variability		ω^2	
k_{el}	IIV of k_{el}	0.194	33.1
V_d	IIV of V_d	0.138	25.9
ξ	IIV of ξ	5.87×10^{-2}	65.6
SC_{50}	IIV of SC_{50}	0.764	25.0
S_{max_1}	IIV of S_{max_1}	1.88×10^{-4}	133
NB_0	IIV of NB_0	0.109	29.1

Interindividual variability (IIV) was reported as variances. Only those values which were found to be impacted by IIV are reported

6.6 Quantification of the Impact of IIV Using Computer Simulation

To rigorously quantify the impact of IIV on the physiological granulopoiesis model, in silico simulations of 500 patients were performed. All simulations were carried out in Matlab 2013a [17]. The physiological model of neutrophil production, consisting of a three-dimensional system of DDEs, and the associated PK/PD models were previously implemented using the *ddesd* solver in Matlab as described in [6]. To incorporate the IIV of the PK models provided in Tables 6.1 and 6.2, each parameter value subject to a random effect was sampled from a normal or multivariate normal distribution and a simulation was performed for these values. This sampling technique was performed 500 times to simulate 500 patients in each scenario. Parameter values were sampled using the *normrnd* and *mvnrnd* functions in Matlab. The following variability scenarios were covered:

(a) 6.892 mg (1 h infusion) PM00104 alone with variability.
(b) 350 mg filgrastim alone with variability.
(c) 6.892 mg (1 h infusion) PM00104 with variability and 350 mg filgrastim without variability.
(d) 6.892 mg (1 h infusion) PM00104 without variability and 350 mg filgrastim with variability.
(e) 6.892 mg (1 h infusion) PM00104 and 350 mg filgrastim, both with variability.

The additional considered scenario without any variability was previously treated in [6] and serves as a reference in the present analysis.

6.6.1 Statistical Analyses

For each variability scenario, a hypothesis test was carried out in Matlab using the *ttest* function; one-sample Student *t*-tests about the mean time to nadir (TNad), mean ANC nadir (Nad), and mean area under the effect curve (AUEC) were performed. These three metrics were chosen as evaluation criteria for the determination of optimal regimens, as previously carried out in [6, 24]. Further, the 95 % asymptotic confidence interval (CI) of the mean differences between the model with (*test*) and without (*reference*) variability was computed to check the narrowness of the CI. The mean difference is then judged significant if 0 is outside of CI, implying the null hypothesis (H_0) cannot be rejected. Finally, in a manner analogous to a bioequivalence trial, the ratios of the *test* to the *reference* for the mean TNad, mean Nad, and AUEC were computed. Accordingly, if the ratio was within the range of 80–125 % [4], both models were considered equivalent implying no difference was observed in terms of this indicator.

6.7 Results

6.7.1 No Statistically Significant Differences in Time to Nadir Between the Model With and Without Variability

Visually, the five simulated scenarios produced TNads close to the mean for most, if not all, of the 500 in silico patients (Fig. 6.2). Indeed, no significant difference in the time to nadir was found when testing for differences in the means between the test and reference models. While the mean TNad varies owing to the particular drug combination being tested, all five of the scenarios examined herein produced no difference to the time to the nadir onset and the asymptotic 95 % CI of the difference in each case was narrow and always included H_0 (see Table 6.3). Further, all ratios of the test model (with IIV) to the reference model (without IIV) were within the 80 % and 125 % range in every scenario. Consequently, the test models can be considered equivalent to the reference model (see Table 6.3).

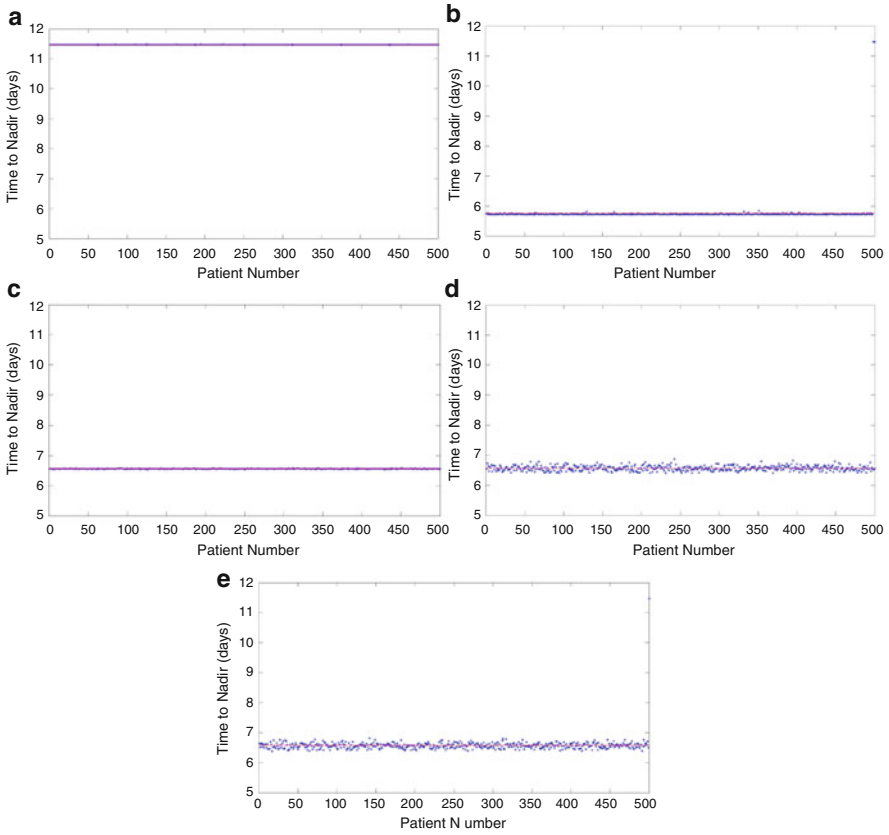


Fig. 6.2 Time to nadir results of each in silico patient in each scenario: (**a**) PM00104 alone with variability, (**b**) Filgrastim alone with variability, (**c**) PM00104 with variability with filgrastim without variability, (**d**) PM00104 without variability and filgrastim with variability, (**e**) PM00104 and filgrastim, both with variability. *Solid horizontal lines* represent the mean of each scenario

6.7.2 No Statistically Significant Differences in the Value of the Nadir Between the Models With or Without Variability

On the other hand, and similar to the time to nadir, both visually (Fig. 6.3) and statistically speaking, no significant differences were found in the nadir values. In all scenarios, the asymptotic 95 % CI of the difference remains narrow, indicating small standard errors, as seen in Table 6.4. In addition, the calculated ratios of the nadir value of the test models versus the reference model were within the interval of 80–125 % and, accordingly, the models with variability were equivalent to the reference model.

Table 6.3 Results of the test for significance in the time to nadir of each of the studied scenarios: (a) PM00104 alone with variability, (b) Filgrastim alone with variability, (c) PM00104 with variability with filgrastim without variability, (d) PM00104 without variability and filgrastim with variability, (e) PM00104 and filgrastim, both with variability

Scenario	Mean $TNad_{ref}$ (days)	Mean $TNad_{test}$ (days)	Ratio (%)	95 % CI of difference
(a)	11.47	11.47	100.0	$10^{-4}\times([-1.77, 1.77])$
(b)	6.26	5.77	92.1	$10^{-2}\times([-3.90, 3.90])$
(c)	6.79	6.56	96.7	$10^{-4}\times([-4.93, 4.93])$
(d)	6.79	6.57	96.8	$10^{-3}\times([-7.30, 7.30])$
(e)	6.79	6.58	96.9	$10^{-2}\times([-2.07, 2.07])$

Ratios computed as (test/reference)×100. In all cases, differences were determined to be statistically insignificant at the $\alpha = 5\,\%$ level (all p-values were 1). $TNad_{ref}$: time to nadir of the reference model, $TNad_{test}$: time to nadir of the test model, CI: confidence interval

6.7.3 No Statistically Significant Differences in the Area Under the Effects-Time Curve Between the Model With or Without Variability

Furthermore, no statistically significant differences in the AUECs were found. While the previous tests of the time to nadir and the nadir value decomposed the results into one direction at a time (x and y, respectively), the AUEC metric provides insight into the simultaneous xy-behaviour of the predictions. Further, as we are no longer simply looking a single nadir point but over the entire 50 simulated days, this last measure synthesises the full temporal nature of the simulated solutions. In the case of the AUECs, as in the previous two tests, all the asymptotic 95 % CI of the difference included 0. Finally, all ratios were within the 80 % and 125 % range and can therefore be considered equivalent (Table 6.5). Figure 6.4 reveals the consistency of the statistical analysis of the AUEC values. That being said, in each scenario involving the full variability model for filgrastim, AUEC values are less uniform. Indeed, because the AUEC investigation shifts the focus to the full time-course studied, the longer-term effects of G-CSF serving to replenish the neutrophil reservoir, such as increased speed of maturation (V_N), increased rate of neutrophil proliferation (η_{NP}), and increased rate of differentiation from the HSCs (κ_N), can be seen.

6.7.4 Full Time Courses of Neutrophil Counts

The full time courses of each patient's ANCs over 50 days were simulated and the results presented in Fig. 6.5. Each subfigure corresponds to one of the five

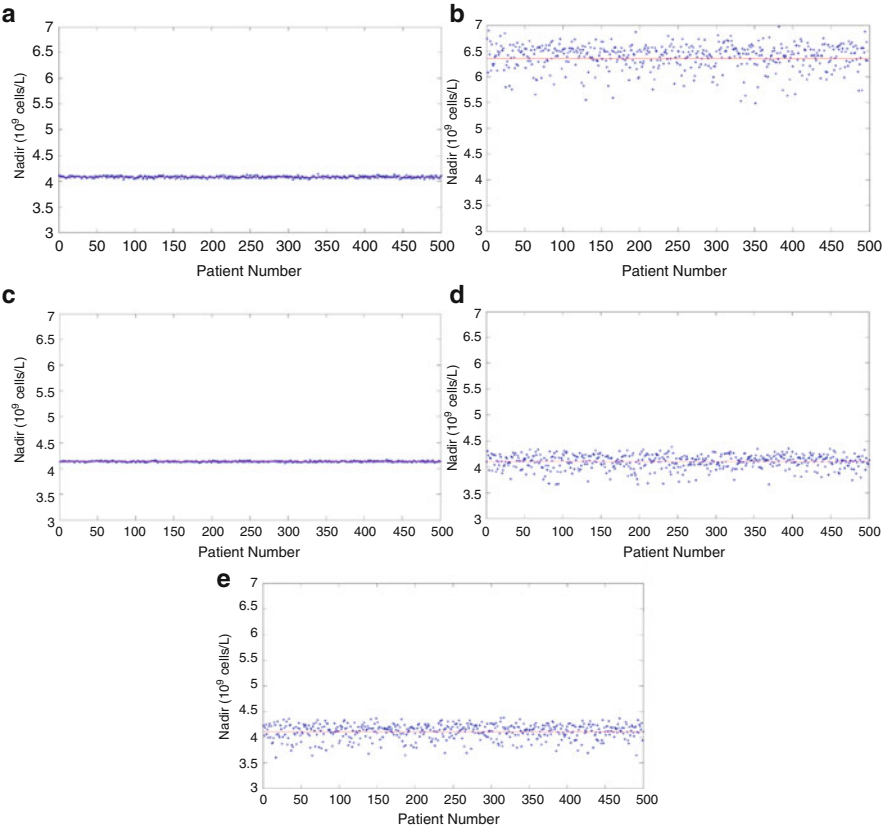


Fig. 6.3 Nadir value per in silico patient in each scenario: (**a**) PM00104 alone with variability, (**b**) Filgrastim alone with variability, (**c**) PM00104 with variability with filgrastim without variability, (**d**) PM00104 without variability and filgrastim with variability, (**e**) PM00104 and filgrastim, both with variability. *Solid horizontal lines* represent the mean of each scenario

different scenarios used to discern the influence of IIV on the prediction. As is consistent with the previously presented results, variability in filgrastim leads to larger variations in the ANC compared to variability in PM00104, which does not cause deviations on the same scale due to its limited involvement in neutrophil development. PM00104 disrupts cellular transcription in a variety of ways, leading to apoptosis through the arrest of the S-phase [19]. We therefore consider that those cells that are no longer dividing, notably neutrophils that have finished proliferation (postmitotic-maturing neutrophils, neutrophils in the marrow reservoir, and circulating and/or marginated neutrophils), are no longer subject to its effects. Accordingly, since these nondividing cells constitute the bulk of the neutrophils in the lineage (postmitotic neutrophils are estimated to be about 77 % of the marrow neutrophils [7]), the PDs of PM00104 will have a more limited role on neutrophils developing in the marrow.

Table 6.4 Results of the test for significance in the nadir value of each of the studied scenarios: (a) PM00104 alone with variability, (b) Filgrastim alone with variability, (c) PM00104 with variability with filgrastim without variability, (d) PM00104 without variability and filgrastim with variability, (e) PM00104 and filgrastim, both with variability

Scenario	Mean Nad_{ref} of (10^9 cells/L)	Mean Nad_{test} (10^9 cells/L)	Ratio (%)	95 % CI of difference
(a)	4.08	4.08	100.7	$10^{-3}\times([-1.60, 1.60])$
(b)	6.88	6.35	92.4	$10^{-2}\times([-4.72, 4.72])$
(c)	4.04	4.14	102.3	$10^{-4}\times([-8.04, 8.04])$
(d)	4.04	4.10	101.5	$10^{-2}\times([-1.20, 1.20])$
(e)	4.04	4.11	101.7	$10^{-2}\times([-1.28, 1.28])$

Ratios computed as (test/reference)×100. In all cases, differences were determined to be statistically insignificant at the $\alpha = 5\,\%$ level (all p-values were 1). Nad_{ref}: nadir of the reference model, Nad_{test}: nadir of the test model, CI: confidence interval

Table 6.5 Results of the test for significance in the area under the effects curve (AUEC) of each of the studied scenarios: (a) PM00104 alone with variability, (b) Filgrastim alone with variability, (c) PM00104 with variability with filgrastim without variability, (d) PM00104 without variability and filgrastim with variability, (e) PM00104 and filgrastim, both with variability

Scenario	Mean $AUEC_{ref}$ [(10^9 cells/L) days]	Mean $AUEC_{test}$ [(10^9 cells/L) days]	Ratio (%)	95 % CI of difference
(a)	404.30	404.30	100.0	$10^{-2}\times([-1.74, 1.74])$
(b)	501.25	478.80	95.5	$10^{-1}\times([-6.91, 6.91])$
(c)	432.73	428.52	99.0	$10^{-2}\times([-2.39, 2.39])$
(d)	432.73	429.92	99.4	$10^{-1}\times([-4.07, 4.07])$
(e)	432.73	429.68	99.3	$10^{-1}\times([-4.08, 4.08])$

Ratios computed as (test/reference)×100. In all cases, differences were determined to be statistically insignificant at the $\alpha = 5\,\%$ level (all p-values were 1). $AUEC_{ref}$: area under the effect-time curve of the reference model, $AUEC_{test}$: area under the effect-time curve of the test model, CI: confidence interval

6.7.5 Assessing the Impact of PK Variability on Regimens Identified by the Physiological Model

The value of the neutrophil nadir following the administration of PM00104 was previously used in [6] to determine those regimens which best mitigated neutropenia and which reduced the number of administrations of filgrastim per 21-day periodic cycle over six cycles. Although the nadir is not affected by the PK IIV, as shown for the single dose scenario reported above, the presence of IOV was reported for PM00104 [18], which may have an impact on dosing regimen decisions. Hence, to extend our findings to the optimal regimens we previously reported, we investigated the impact of IOV on the physiological model by simulating three cycles of 21-day

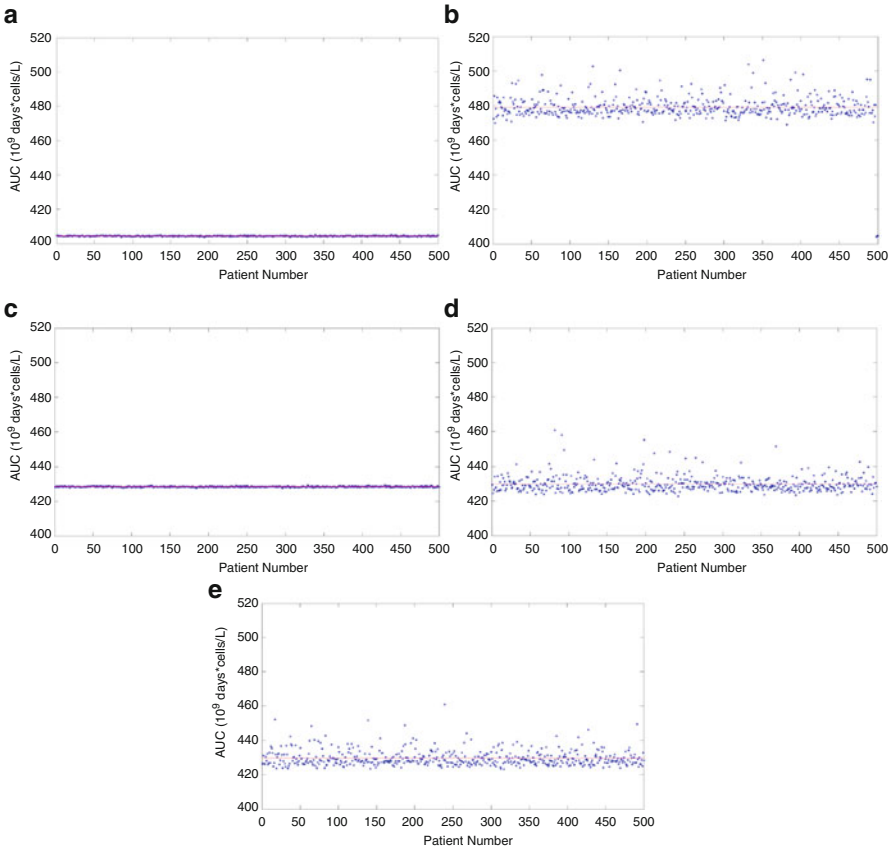


Fig. 6.4 Area under the effect curve (AUEC) results per in silico patient in each scenario: (**a**) PM00104 alone with variability, (**b**) Filgrastim alone with variability, (**c**) PM00104 with variability with filgrastim without variability, (**d**) PM00104 without variability and filgrastim with variability, (**e**) PM00104 and filgrastim, both with variability. *Solid horizontal lines* represent the mean of each scenario

periodic administration of 6.892 mg (1 h infusion) of PM00104 with both IIV and IOV models as in [18] for another group of in silico patients. However, no significant impact on the model's predictions could be observed with this additional source of variability (not shown). Since all clinical markers used in this study are in fact not affected by the presence of PK IIV, and IOV did not have significant impact on the prediction of nadir, it is reasonable for us to extend the regimens identified for the average patient using the physiological model to the population as reported in [18]. Consequently, in line with the findings of [6], it may be prudent to delay the first administration of filgrastim after the administration of chemotherapy to lessen the impact of the anti-cancer drug(s) on the neutrophil lineage.

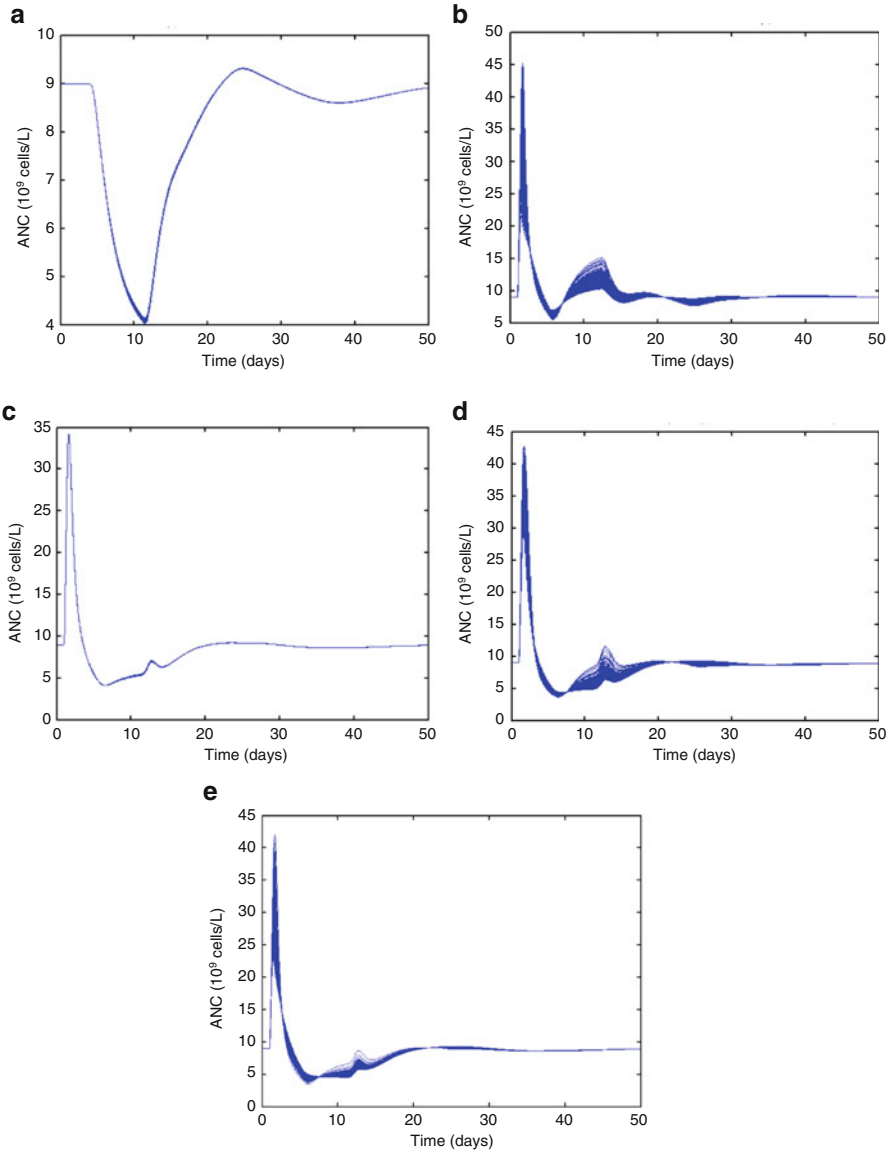


Fig. 6.5 Absolute neutrophil counts of each in silico patient in the five studied scenarios: (**a**) PM00104 alone with variability, (**b**) Filgrastim alone with variability, (**c**) PM00104 with variability with filgrastim without variability, (**d**) PM00104 without variability and filgrastim with variability, (**e**) PM00104 and filgrastim, both with variability

6.8 Discussion

Physiological modelling is increasingly in-demand as a means to deepen our understanding of the mechanisms underlying drug fate and effects and to allow for an increased incorporation of physiology beyond the more popular data-driven physiologically-based models. Although the most widespread model evaluation criteria are based on the goodness of fit of the model to data, this approach can overlook subtle mechanisms behind the physiological system that may be essential to explain outcomes. When these mechanisms become the key to ensuring the model's generalisability, concerns could be raised about the transferability of the model's findings and its applicability to a variety of situations. To bridge this gap, it is natural to incorporate first principles modelling rooted in the most contemporary scientific knowledge; this mechanistic methodology ultimately leads to increases in our understanding of the principle processes of the systems under study. Physiological PK/PD models use rigorous mathematical expressions to characterise processes on the causal path between xenobiotics, the body, and pathologies, and are intended to improve our capacity for extrapolation and prediction [8]. As a result of the complexity of this tri-relationship, these kinds of models frequently use an average representation to reconstruct the entire process. While one can admit that this average portrayal greatly serves to demystify the physiological mechanisms being studied in addition to their interactions with drugs and diseases, its application to a population can be challenged by the presence of different sources of variability. The robustness of these physiological PK/PD models to IIV, a pervasive concern in PopPK/PD, has to be investigated to ensure clinical applicability.

This is precisely the objective of the current study; herein, we were concerned by the impact of IIV on the predictive quality of a physiological PK/PD model that we previously developed to study the production of neutrophils for use in chemotherapeutic contexts [6]. This model was predicated upon detailed hematopoietic mechanisms and incorporated the pharmacokinetics of a chemotherapeutic agent (PM00104) and a granulostimulant (filgrastim) to successfully reproduce the behaviour of this lineage with respect to a variety of oncological protocols. In this study, we have added the reported IIV of the PopPK models of the two previously considered drugs, PM00104 [18] and filgrastim [14], to investigate the impact of IIV on this model. Five variability scenarios were considered to assess the impact of IIV for each separate drug model in addition to combinations thereof. First, single doses of PM00104 and filgrastim were administered alone to two separate cohorts of 500 in silico patients using each drug's respective variability model. Second, both drugs were administered together to two other sets of 500 in silico patients taking into account just the fixed effects of one drug, and the fixed and random effects of the other. Last, the variability models of both drugs were applied together to another set of 500 patients to evaluate the full impact of the IIV of both drugs. In all situations, no significant impact of IIV was observed on any chosen clinical indicator, namely the time to ANC nadir, the mean ANC value of the nadir, and the mean AUEC, nor did IOV significantly impact the predictions. These findings

confirm the applicability of the physiological PK/PD neutrophil model beyond the case of an average patient. For example, on the basis of our current findings, the administration regimen previously judged to be optimal to avoid moderate to severe neutropenia during 21-day periodic administration of PM00104 over 6 cycles (G-CSF given on days 7 through 10 following the administration of PM00104) can now be extended beyond the average patient to a population of patients [6]. We attribute the robustness of the model's predictions to pharmacokinetic IIV to the fact that both drugs (PM00104 and filgrastim) have short elimination half-lives in comparison with the lifespan of the neutrophils they affect, a factor inherently incorporated into the physiological model. Since the physiological PK/PD model was built to reproduce the sequential events leading to the formation of neutrophils (recruitment of HSCs into the neutrophil line, proliferation, maturation, and release of mature neutrophils into circulation) and specifically identifies where each drug has an effect on the appearance of a neutrophil in the blood stream, the model is able to directly relate each drug's concentration in the plasma with the chain of events it will induce over the course of a neutrophil's lifespan. Thus, because both PM00104 and filgrastim have shorter PK timespans than their PDs, the magnitude of their IIV is also shorter lived and these differences will only marginally influence the effects on the physiology.

The results presented in this work provide evidence that physiological modelling is a valuable alternative to the widely used data-driven modelling approach. Once their general applicability has been proved, as it is the case for the present model, physiological models can even transcend the means for which they are intended, thereby justifying the effort required for their construction. A large number of physiological models have been developed for which a combination with PK/PD models can be envisaged. It would be advisable to systematically submit these models to a "variability screening test" to guarantee their general applicability prior to any clinical validation. Designing such standard tests remains a challenge that has to be addressed. Indeed, the National Institutes of Health (NIH) in the USA has identified quantitative systems pharmacology (QSP) as ideally situated to develop quantitative and predictive methods able to identify the impact of individual variability [23]. Fortunately, system biologists and pharmacologists have access to a variety of databases [1] which facilitate both the formation/instatement and the evolution of standardised variability screening tests. Ultimately, a concerted and coordinated effort between industry, academia, and regulatory agencies, as exemplified by the partnership between the NIH and the Food and Drug Administration (FDA), is required to ensure that variability is addressed when performing QSP approaches.

In conclusion, this study not only substantiates and situates the use of physiological modelling in pharmacometrics, it provides incentives to continue to improve our understanding of the underlying physiological mechanisms of a given system. In a broader sense, this work testifies to the necessity of building bridges between diverse actors from different backgrounds (pharmaceutical scientists, clinicians, biomathematicians, statisticians, engineers, etc.) in the pharmacometrics community to best serve patients and their needs.

Notes

This work makes up a portion of the doctoral thesis of MC.

References

1. Bai, J.P.F., Abernathy, D.R.: Systems pharmacology to predict drug toxicity: integration across levels of biological organization. Annu. Rev. Pharmacol. Toxicol. **53**, 451–473 (2013)
2. Brooks, G., Langlois, G., Lei, J., Mackey, M.C.: Neutrophil dynamics after chemotherapy and G-CSF: the role of pharmacokinetics in shaping the response. J. Theor. Biol. **315**, 97–109 (2012)
3. Brown, R., Delp, M., Lindstedt, S., Rhomberg, L., Beliles, R.: Physiological parameter values for physiologically based pharmacokinetic models. Toxicol. Ind. Health **13**, 407–484 (1997)
4. Center for Drug Evaluation and Research (CDER): U.S. Department of Health and Human Services Food and Drug Administration. Guidance for industry. Bioavailability and bioequivalence studies submitted in NDAs or INDs—general considerations. Tech. rep. (2014)
5. Colijn, C., Mackey, M.C.: A mathematical model of hematopoiesis: II. Cyclical neutropenia. J. Theor. Biol. **237**, 133–146 (2005)
6. Craig, M., Humphries, A., Bélair, J., Li, J., Nekka, F., Mackey, M.C.: Neutrophil dynamics during concurrent chemotherapy and g-csf administration: mathematical modelling guides dose optimisation to minimise neutropenia. J. Theor. Biol. **385**, 77–89 (2015)
7. Dancey, J., Deubelbeiss, K., Harker, L., Finch, C.: Neutrophil kinetics in man. J. Clin. Invest. **58**, 705–715 (1976)
8. Danhof, M., DeLange, E., Della Pasqua, O., Ploeger, B., Voskuyl, R.: Mechanism-based pharmacokinetic-pharmacodynamic (pkpd) modeling in translational drug research. Trends Pharmacol. Sci. **29**, 186–191 (2008)
9. Foley, C., Mackey, M.C.: Mathematical model for G-CSF administration after chemotherapy. J. Theor. Biol. **257**, 27–44 (2009)
10. Foley, C., Bernard, S., Mackey, M.C.: Cost-effective G-CSF therapy strategies for cyclical neutropenia: mathematical modelling based hypotheses. J. Theor. Biol. **238**, 756–763 (2006)
11. Furze, R.C., Rankin, S.M.: Neutrophil mobilization and clearance in the bone marrow. Immunology **125**, 281–288 (2008)
12. Gobburu, J., Agersø, H., Jusko, W., Ynddal, L.: Pharmacokinetic-pharmacodynamic modeling of ipamorelin, a growth hormone releasing peptide, in human volunteers. Pharm. Res. **16**, 1412–1416 (1999)
13. González-Sales, M., Valenzuela, B., Pérez-Ruixo, C., Fernández Teruel, C., Miguel-Lillo, B., Matos, A.S., et al.: Population pharmacokinetic-pharmacodynamic analysis of neutropenia in cancer patients receiving PM00104 (Zalypsis). Clin. Pharmacokinet. **51**, 751–764 (2012)
14. Krzyzanski, W., Wiczling, P., Lowe, P., Pigeolet, E., Fink, M., Berghout, A., et al.: Population modeling of filgrastim PK-PD in healthy adults following intravenous and subcutaneous administrations. J. Clin. Pharmacol. **9**(Suppl.), 101S–112S (2010)
15. Kuwabara, T., Kato, Y., Kobayashi, S., Suzuki, H., Sugiyama, Y.: Nonlinear pharmacokinetics of a recombinant human granulocyte colony-stimulating factor derivative (Nartograstim): species differences among rats, monkeys and humans. J. Pharmacol. Exp. Ther. **271**, 1535–1543 (1994)
16. Layton, J.E., Hall, N.E.: The interaction of G-CSF with its receptor. Front. Biosci. **31**, 177–199 (2006)
17. Mathworks: MATLAB 2013a. Mathworks, Natick (2013)

18. Pérez-Ruixo, C., Valenzuela, B., Fernández Teruel, C., González-Sales, M., Miguel-Lillo, B., Soto-Matos, A., et al.: Population pharmacokinetics of PM00104 (Zalypsis) in cancer patients. Cancer Chemother. Pharmacol. **69**, 15–24 (2012)

19. Petek, B., Jones, R.: PM00104 (Zalypsis®): a marine derived alkylating agent. Molecules **19**, 12328–12335 (2014). doi:10.3390/molecules190812328

20. Price, T.H., Chatta, G.S., Dale, D.C.: Effect of recombinant granulocyte colony-stimulating factor on neutrophil kinetics in normal young and elderly humans. Blood **88**, 335–340 (1996)

21. Rankin, S.M.: The bone marrow: a site of neutrophil clearance. J. Leukoc. Biol. **88**, 241–251 (2010)

22. Scholz, M., Schirm, S., Wetzler, M., Engel, C., Loeffler, M.: Pharmacokinetic and -dynamic modelling of G-CSF derivatives in humans. Theor. Biol. Med. Model. **9**, 1497–1502 (2012)

23. Sorger, D.R., Allerheiligen, S.R.B., Abernethy, R.B., Altman, K.L.R., Brouwer, A.C., Califano, A., D'Argenio, D.Z., Iyengar, R., Jusko, W.J., Lalonde, R., Lauffenburger, D.A., Shoichet, B., Stevens, J.L., Subramaniam, S., van der Graaf, P., Vincini, P.: Quantitative and systems pharmacology in the post-genomic era: new approaches to discovering drugs and understanding therapeutic mechanisms. An NIH white paper by the QSP Workshop Group – October 2011, pp. 1–47. National Institutes of Health of the United States of America, Bethesda (2011)

24. Vainas, O., Ariad, S., Amir, O., Mermershtain, W., Vainstein, V., Kleiman, M., Inbar, O., Ben-Av, R., Mukherjee, A., Chan, S., Agur, Z.: Personalising docetaxel and G-CSF schedules in cancer patients by a clinically validated computational model. Br. J. Cancer **107**, 814–822 (2012)

25. Wang, B., Ludden, T.M., Cheung, E.N., Schwab, G.G., Roskos, L.K.: Population pharmacokinetic-pharmacodynamic modeling of filgrastim (r-metHuG-CSF) in healthy volunteers. J. Pharmacokinet. Pharmacodyn. **28**, 321–342 (2001)

Chapter 7
Asymptotic Behavior of Linear Almost Periodic Differential Equations

Bui Xuan Dieu, Luu Hoang Duc, Stefan Siegmund, and Nguyen Van Minh

Abstract The present paper is concerned with strong stability of solutions of non-autonomous equations of the form $\dot{u}(t) = A(t)u(t)$, where $A(t)$ is an unbounded operator in a Banach space depending almost periodically on t. A general condition on strong stability is given in terms of Perron conditions on the solvability of the associated inhomogeneous equation.

Keywords Strong stability • Non-autonomous equation • Almost periodicity • Evolution semigroup • Perron type conditions

2010 Mathematics Subject Classification. Primary: 34C27; Secondary: 34D05, 34D20, 47D06

B.X. Dieu
Department of Mathematics, Dresden University of Technology, 01062 Dresden, Germany

School of Applied Mathematics & Informatics, Hanoi University of Science and Technology, 01 Dai Co Viet Road, Ha Noi, Viet Nam
e-mail: Bui.Xuan_Dieu@mailbox.tu-dresden.de; dieu.buixuan@hust.edu.vn

L.H. Duc
Department of Mathematics, Dresden University of Technology, 01062 Dresden, Germany

Institute of Mathematics, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet Road, 10307 Ha Noi, Viet Nam
e-mail: Hoang_Duc.Luu@tu-dresden.de; lhduc@math.ac.vn

S. Siegmund
Department of Mathematics, Dresden University of Technology, 01062 Dresden, Germany
e-mail: stefan.siegmund@tu-dresden.de

N. Van Minh (✉)
Department of Mathematics, Columbus State University, 4225 University Avenue, Columbus, GA 31907, USA

Department of Mathematics & Statistics, University of Arkansas at Little Rock, Little Rock, AR 72204, USA
e-mail: mvnguyen1@ualr.edu

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_7

7.1 Introduction

In this paper we consider the asymptotic behavior of solutions of evolution equations of the form

$$\frac{dx}{dt} = A(t)x, \quad t \in \mathbb{R}, \tag{7.1}$$

in terms of the existence and uniqueness of bounded solutions to the inhomogeneous equations

$$\frac{dx}{dt} = A(t)x + f(t), \quad t \in \mathbb{R}, \tag{7.2}$$

where $A(t)$ is a family of unbounded linear operators on a (complex) Banach space $\mathbb{X}$ that depends almost periodically on t, and f is an almost periodic function taking values in $\mathbb{X}$. Throughout the paper we assume that the homogeneous equation (7.1) generates an almost periodic evolutionary process $(U(t,s)_{t\geq s})$ (see Definition 2.5). If (7.1) is autonomous (that is, $A(t) = A$ for all t), and $\dim(\mathbb{X}) < \infty$, the classical Lyapunov Theorem states that (7.2) is strongly stable if all real parts of the eigenvalues of A are negative. In the infinite dimensional case this condition is no longer valid. In fact, one needs more conditions to guarantee the (strong) stability of solutions to (7.1). We refer the reader to [3, 5, 9, 26, 28] and the references therein. If $A(t)$ depends on t, the spectra $\sigma(A(t))$, in general, do not play any role in determining the behavior of solutions to (7.1). If $A(t)$ depends periodically on t with period τ, the period map $P := U(\tau, 0)$ can be used to study the problem via discrete analogs, results that can be found in [1, 21, 25].

The problem becomes much more complicated when $A(t)$ depends on t almost periodically (but not periodically). The idea of linear skew products has been used extensively to study the stability and exponential dichotomy of non-autonomous equations (see, e.g., [10, 13, 22] and the references therein). In this direction, the concept of evolution semigroups, as a variation of the aforementioned idea, proves to be a very effective analytic tool (see [4, 8, 27]). However, since typically evolution semigroups are considered in the function spaces $L^p(\mathbb{X})$, $C_0(\mathbb{X})$ or $\mathrm{AP}(\mathbb{X})$, the spectrum of the evolution semigroup associated with (7.1) is too coarse to characterize finer properties of the system like strong stability. Indeed, the spectrum of the generator of the evolution semigroup associated with an evolution equation in one of these function spaces consists of a union of vertical strips in the complex plane, hence the imaginary axis is either contained completely in the spectrum or does not intersect it. On the other hand, the well-known ABLV Theorem (see Theorem 2.4) for stability of C_0-semigroups allows the generator's spectrum to intersect the imaginary axis.

The main purpose of this paper is to provide a setting in which the idea of evolution semigroups combined with the spectral theory of functions can be further used to study the asymptotic behavior of solutions of (7.1). We consider

the evolution semigroup associated with (7.1) in the smallest invariant subspace of the space of all almost periodic functions AP($\mathbb{X}$) which we call *minimal evolution semigroup* of (7.1). In general, this smallest invariant function space is determined by the Bohr spectrum of the coefficient operator $A(t)$. The main results (cf. Theorem 3.6) we obtain in the paper are extensions of results known in the autonomous and periodic cases. Our conditions for strong stability are stated in terms of Perron conditions which are very popular in recent studies on stability and dichotomy (see, for example, [18–20, 23, 24, 27]). We analyze some particular cases as examples of how the obtained results can be applied to equations with almost periodic coefficients.

7.2 Preliminaries

7.2.1 Almost Periodic Functions

In this paper we use the concept of almost periodicity in Bohr's sense. The reader is referred to [3, 14] for some standard definitions and properties of almost periodic functions taking values in a Banach space $\mathbb{X}$.

Definition 2.1. A bounded and continuous function $g : \mathbb{R} \to \mathbb{X}$ is said to be almost periodic in the sense of Bohr (or simply almost periodic) if for each given sequence $\{\tau_n\}_{n=1}^{\infty} \subset \mathbb{R}$ there exists a subsequence $\{\tau_{n_k}\}_{k=1}^{\infty}$ such that the limit

$$\lim_{k\to\infty} g(t + \tau_{n_k})$$

exists uniformly in $t \in \mathbb{R}$.

Given an almost periodic function g, for each $\lambda \in \mathbb{R}$ the following is shown to exist (see [14])

$$M_{\lambda,g} := \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{-i\lambda t} g(t)dt.$$

And, except for at most a countable set $\sigma_b(g)$ of values of λ, this limit $M_{\lambda,g}$ is always equal to zero.

Definition 2.2. Let $T \subset \mathbb{R}$. The semi-module generated by T, denoted by $sm(T)$, is the set of all real numbers μ of the form

$$\mu := n_1\lambda_1 + n_2\lambda_2 + \cdots + n_k\lambda_k$$

where k is a positive integer, $\lambda_1, \lambda_2, \ldots, \lambda_k \in T$, and $n_1, n_2, \ldots, n_k$ are non-negative integers.

Note that by this definition $0 \in sm(T)$.

Let $\sigma_b(A)$ be the Bohr spectrum of the almost periodic $A : \mathbb{R} \to \mathbb{X}$. We will denote the semi-module generated by this spectrum by $\Lambda := sm(\sigma_b(A))$. We introduce the following notation

$$\mathrm{AP}_\Lambda(\mathbb{X}) := \{g \in \mathrm{AP}(\mathbb{X}) : \sigma_b(g) \subset \Lambda\}.$$

Note that $\mathrm{AP}_\Lambda(\mathbb{X})$ is a closed subspace of $\mathrm{AP}(\mathbb{X})$ (see [7, Lemma 2.1]).

Let A be a closed operator in a Banach space $\mathbb{X}$, and let A generate a uniformly bounded C_0-semigroup of linear operators $(T(t))_{t\geq 0}$, i.e., $\sup_{t\geq 0} \|A(t)\| < \infty$. The following lemma is proved in [6, p. 2073].

Lemma 2.3. *Let $x \in \mathbb{X}$ be fixed, then the map*

$$R : \{\mathrm{Re}\lambda > 0\} \to \mathbb{X}, \lambda \mapsto R(\lambda, A)x$$

is holomorphic. Furthermore, denote by $\sigma_u(A, x)$ the local unitary spectrum of x, i.e. the set of points $\lambda \in i\mathbb{R}$ to which R cannot be extended holomorphically. Then $\sigma_u(A, x) \subset \sigma(A) \cap i\mathbb{R}$, and

$$\sigma(A) \cap i\mathbb{R} = \bigcup_{x\in\mathbb{X}} \sigma_u(A, x).$$

We also restate a well-known result from [6, Theorem 3.4].

Theorem 2.4 (ABLV Theorem). *Suppose $(T(t))_{t\geq 0}$ is a bounded C_0-semigroup in a Banach space $\mathbb{X}$ with generator A, $x \in \mathbb{X}$ is fixed. Denote by $\sigma_u(A, x)$ the set of $i\beta \in i\mathbb{R}$ such that the local resolvent, $R(\alpha + i\beta, A)x, \alpha > 0$, does not extend analytically in some neighborhood of $i\beta$. If*

(i) *$\sigma_u(A, x)$ is countable,*
(ii) *$\lim_{\alpha\downarrow 0} \alpha R(\alpha + i\beta, A)x = 0$, for all β with $i\beta \in \sigma_u(A, x)$,*

then

$$\lim_{t\to\infty} \|T(t)x\| = 0.$$

7.2.2 Evolution Semigroups Associated with Evolutionary Processes

Definition 2.5. A two-parameter family $(U(t, s))_{t\geq s}$ of bounded linear operators acting in a Banach space $\mathbb{X}$ is said to be an *evolutionary process* if the following conditions are satisfied:

(i) $U(t, t) = Id$, for all $t \in \mathbb{R}$, where Id is the identity operator of $\mathbb{X}$,
(ii) $U(t, r)U(r, s) = U(t, s)$ for all $s \leq r \leq t$,

(iii) There are non-negative numbers M, α such that $\|U(t,s)\| \leq Me^{\alpha(t-s)}$ for all $t \geq s$,
(iv) The map $(t,s) \mapsto U(t,s)x$ is continuous for each $x \in \mathbb{X}$.

An evolutionary process $(U(t,s))_{t\geq s}$ in a Banach space $\mathbb{X}$ is said to be *almost periodic* if

(v) for each $x \in \mathbb{X}, s \in \mathbb{R}$ the function $\mathbb{R} \ni t \mapsto U(t+s,t)x \in \mathbb{X}$ is almost periodic.

Definition 2.6. Given a function space F as a subspace of $BC(\mathbb{R},\mathbb{X})$. Assume that $(U(t,s))_{t\geq s}$ is an almost periodic evolutionary process generated by (7.1) and, for each $h \geq 0$ and $g \in F$, the function $\mathbb{R} \ni t \mapsto U(t,t-h)g(t-h)$ belongs to F. Then, the *evolution semigroup* $(T^h)_{h\geq 0}$ associated with (7.1) in the function space F is defined as the family of bounded operators $T^h, h \geq 0$, defined by

$$[T^h g](t) = U(t,t-h)g(t-h), \quad g \in F, t \in \mathbb{R}, h \geq 0.$$

From our assumption that $(U(t,s))_{t\geq s}$ is almost periodic evolutionary process, the function $\mathbb{R} \ni t \mapsto U(t,t-h)g(t-h)$ belongs to $\mathrm{AP}(\mathbb{X})$ for every $g \in \mathrm{AP}(\mathbb{X})$ (see [4, Lemma 3]). Hence by choosing $F := \mathrm{AP}(\mathbb{X})$ in the worst case, we see that the evolution semigroup $(T^h)_{h\geq 0}$ associated with (7.1) is well defined.

7.3 Asymptotic Behavior of Solutions

Consider evolution equations of the form

$$\dot{x} = [A_0 + A(t)]x, \quad t \in \mathbb{R}, x \in \mathbb{X}, \tag{7.3}$$

where A_0 generates a C_0-semigroup denoted by $e^{tA_0}, t \geq 0$, and $A : \mathbb{R} \to L(\mathbb{X})$ is almost periodic in the norm topology.

To Eq. (7.3) we associate the following integral equation

$$x(t) = e^{(t-s)A_0}x(s) + \int_s^t e^{(t-\xi)A_0}A(\xi)x(\xi)d\xi, \quad t \geq s; t,s \in \mathbb{R}. \tag{7.4}$$

Every continuous function $x(\cdot)$ on an interval J, which is of the form $[a,b]$, $(a,b]$, (a,b) or $[a,b)$, is said to be a mild solution of (7.3) on J.

It is well known that (7.3) generates an evolutionary process in $\mathbb{X}$ which is determined by the integral equation (7.4). The following theorem shows that this process is almost periodic, and its associated evolution semigroup $(T^h)_{h\geq 0}$ in $\mathrm{AP}(\mathbb{X})$ leaves the function space $\mathrm{AP}_\Lambda(\mathbb{X})$ invariant, where Λ is the semi-module generated by the Bohr spectrum of $A : \mathbb{R} \to L(\mathbb{X})$.

Theorem 3.1. *Under the above assumptions and notation, the evolution semigroup associated with Eq.* (7.3) *in the function space* $\mathrm{AP}_\Lambda(\mathbb{X})$ *is well defined as a* C_0*-semigroup.*

Proof. First, note that the evolution semigroup associated with the evolution equation $\dot{x} = A_0 x$ is well defined as a C_0-semigroup in $\mathrm{AP}_\Lambda(\mathbb{X})$. In fact, this follows from the fact that $e^{hA_0}g(\cdot)$ is in $\mathrm{AP}(\mathbb{X})$ whenever g is in $\mathrm{AP}(\mathbb{X})$ and h is fixed. Moreover, $\sigma_b(e^{hA}g(\cdot)) \subset \sigma_b(g)$. Next, since $A(\cdot)$ is almost periodic in the norm topology, in $\mathrm{AP}(\mathbb{X})$ we can define the operator M_A of multiplication by $A(t)$, that is,

$$M_A : \mathrm{AP}(\mathbb{X}) \ni g \mapsto A(\cdot)g(\cdot) \in \mathrm{AP}(\mathbb{X}).$$

Note that M_A leaves $\mathrm{AP}_\Lambda(\mathbb{X})$ invariant. In fact, this can be checked by using the Approximation Theorem of almost periodic functions [12, Theorem 1.19, p. 27] as follows. Since $g \in \mathrm{AP}_\Lambda(\mathbb{X})$, there is a sequence of trigonometric polynomials g_n with exponents in Λ that approximates g. Similarly, we construct a sequence of trigonometric polynomials $A_n(\cdot)$ that approximates $A(\cdot)$ in norm topology with exponents also in Λ. Then, $A_n(\cdot)g_n(\cdot)$ approximates $A(\cdot)g(\cdot)$. As Λ is the semimodule generated by $\sigma_b(A(\cdot))$ we get that the exponents of $A_n(\cdot)g_n(\cdot)$ lie in Λ.

Let G_{A_0} be the generator of the evolution semigroup (T_0^h) associated with $\dot{x} = A_0 x$ in $\mathrm{AP}_\Lambda(\mathbb{X})$. Then, since M_A is a bounded linear operator in $\mathrm{AP}_\Lambda(\mathbb{X})$, $G_{A_0} + M_A$ generates a C_0-semigroup in $\mathrm{AP}_\Lambda(\mathbb{X})$. We now show that this semigroup in $\mathrm{AP}_\Lambda(\mathbb{X})$ is nothing but the evolution semigroup of (7.4) in $\mathrm{AP}_\Lambda(\mathbb{X})$ associated with Eq. (7.3). In fact, let us denote by (S^h) the semigroup that is generated by $G_{A_0} + M_A$ in $\mathrm{AP}_\Lambda(\mathbb{X})$. Then, this semigroup (S^h) satisfies the equation

$$S^h v = T_0^h v + \int_0^h T_0^{h-\xi} M_A S^\xi v \, d\xi, \quad \text{for all } h \geq 0, v \in \mathrm{AP}_\Lambda(\mathbb{X}).$$

Therefore, for each $t \in \mathbb{R}$, by the definition of the evolution semigroup (T_0^h) associated with the equation $\dot{x} = A_0 x$, we have

$$\begin{aligned}[S^h v](t) &= (T_0^h v)(t) + \int_0^h (T_0^{h-\xi}(M_A S^\xi v))(t) d\xi \\ &= T_0(h)v(t-h) + \int_0^h T_0(h-\xi)(M_A S^\xi v)(t-h+\xi) d\xi.\end{aligned}$$

Since h and t are arbitrary, we may set $h = t - s$, so the above equation becomes

$$\begin{aligned}[S^{t-s} v](t) &= T_0(t-s)v(s) + \int_0^h T_0(t-s-\xi)(M_A S^\xi v)(s+\xi) d\xi \\ &= T_0(t-s)v(s) + \int_s^t T_0(t-\eta)(M_A S^{\eta-s} v)(\eta) d\xi.\end{aligned}$$

Define $w(t) := [S^{t-s}v](t)$, then w is the unique solution of the equation

$$w(t) = T_0(t-s)v(s) + \int_0^h T_0(t-s-\xi)A(s+\xi)S^{\xi}v(s+\xi)d\xi$$
$$= T_0(t-s)x + \int_s^t T_0(t-\eta)A(\eta)w(\eta)d\eta.$$

However, this is the equation that defines $U(t,s)x$. Therefore, we have for all $t \geq s$, $v \in \mathrm{AP}_\Lambda(\mathbb{X})$

$$[S^{t-s}v](t) = U(t,s)v(s).$$

In particular, when $h := t-s$ we have that $S^h v = U(t,t-h)v(t-h)$ for all $h \geq 0, t \in \mathbb{R}, v \in \mathrm{AP}_\Lambda(\mathbb{X})$, i.e., $(S^h)_{h\geq 0}$ is the evolution semigroup $(T^h)_{h\geq 0}$ for each $h \geq 0$. This completes the proof of Theorem. □

Definition 3.2. The evolution semigroup $(T^h)_{h\geq 0}$ associated with (7.3) in the function space $\mathrm{AP}_\Lambda(\mathbb{X})$ is called the *minimal evolution semigroup* associated with (7.3) and $G := G_{A_0} + M_A$ is called the *infinitesimal generator* of T^h.

Definition 3.3. A well-posed equation Eq.(7.1) is said to be *strongly stable* if the evolution process $(U(t,s))_{t\geq s}$ associated with it satisfies:

$$\lim_{t\to\infty} U(t,s)x = 0 \tag{7.5}$$

for all fixed $x \in \mathbb{X}$ and $s \in \mathbb{R}$.

If $A(t)$ is independent of t and generates a C_0-semigroup, the strong stability of the evolution equation (7.1) means that $\lim_{t\to\infty} T(t)x = 0$ for all $x \in \mathbb{X}$.

Theorem 3.4. *Equation (7.3) is strongly stable if its minimal evolution semigroup $(T^h)_{h\geq 0}$ is strongly stable.*

Proof. Let $0 \neq x_0 \in \mathbb{X}$. Define $g(t) = x_0$ for all $t \in \mathbb{X}$. Obviously, $g \in \mathrm{AP}_\Lambda(\mathbb{X})$ because $\sigma_b(g) \subset \{0\}$. Since $(T^h)_{h\geq 0}$ is strongly stable,

$$\lim_{h\to\infty} T^h z = 0$$

for each $z \in \mathrm{AP}_\Lambda(\mathbb{X})$. In particular,

$$\lim_{h\to\infty} T^h g = 0.$$

That means,

$$0 = \lim_{h\to\infty} \sup_{t\in\mathbb{R}} \|U(t,t-h)x_0\| \geq \lim_{h\to\infty} \|U(h,0)x_0\| \geq 0.$$

Therefore, every mild solution of (7.3) is convergent to the origin, proving strong stability of (7.3). □

Definition 3.5. Let us denote by Σ the following set

$$\Sigma := \sigma(G) \cap i\mathbb{R},$$

and call it *the spectrum of equation* (7.3).

It can be checked (see, e.g., [16]) that the generator $G := G_{A_0} + M_A$ is well defined. The domain $D(G)$ of G consists of all functions u in $\mathrm{AP}_\Lambda(\mathbb{X})$ such that there exists a function $f \in \mathrm{AP}_\Lambda(\mathbb{X})$ for which

$$u(t) = U(t,s)u(s) + \int_s^t U(t,\xi)f(\xi)d\xi, \text{ for all } t \geq s, t, s \in \mathbb{R}, f \in \mathrm{AP}_\Lambda(\mathbb{X}).$$

And, in this case, for such f and u, $Gu = -f$. In the same way, $u \in D(G_{A_0})$ and $G_{A_0}u = -f$ if and only if for all $t \geq s$, $t, s \in \mathbb{R}, f \in \mathrm{AP}_\Lambda(\mathbb{X})$

$$u(t) = T_0(t-s)u(s) + \int_s^t T_0(t-\xi)f(\xi)d\xi.$$

The operator $G - \lambda$ generates a semigroup (R^h) in $\mathrm{AP}_\Lambda(\mathbb{X})$ which is uniquely determined by the equation

$$R^h v = T^h v + \int_0^h T^{h-\xi}(-\lambda)R^\xi v d\xi, \quad h \geq 0, v \in \mathrm{AP}_\Lambda(\mathbb{X}).$$

Without difficulty we can check that $R^h = e^{-\lambda h}T^h$ which is exactly the evolution semigroup associated with the process $V(t,s) := e^{-\lambda(t-s)}U(t,s)$. Therefore, $u \in D(G-\lambda)$ and $(G-\lambda)u = -f$ if and only if

$$u(t) = e^{-\lambda(t-s)}U(t,s)u(s) + \int_s^t e^{-\lambda(t-\xi)}U(t,\xi)u(\xi)d\xi, \quad (t \geq s).$$

Therefore, λ belongs to $\rho(G)$ if and only if for each $f \in \mathrm{AP}_\Lambda(\mathbb{X})$ there is a unique solution $u_{\lambda,f}$ to the equation

$$u_{\lambda,f}(t) = e^{-\lambda(t-s)}U(t,s)u_{\lambda,f}(s) + \int_s^t e^{-\lambda(t-\xi)}U(t,\xi)f(\xi)d\xi \tag{7.6}$$

for all $t, s \in \mathbb{R}$ with $t \geq s$.

With this preparation we are ready to prove the main result of the paper.

Theorem 3.6. *Assume that*

$$\sup_{t \geq s} \|U(t,s)\| < \infty, \tag{7.7}$$

and for all $\lambda \in i\mathbb{R}$ but at most a countable set Σ, Eq. (7.6) has a unique solution $u_{\lambda,f}$ for each given $f \in \mathrm{AP}_\Lambda(\mathbb{X})$. Moreover, assume that for each $\lambda \in \Sigma$ and each fixed $f \in \mathrm{AP}_\Lambda(\mathbb{X})$,

$$\lim_{\alpha \downarrow 0} \alpha u_{\lambda+\alpha,f} = 0. \tag{7.8}$$

Then Eq. (7.3) is strongly stable.

Proof. First, by (7.7) the semigroup (T^h) is uniformly bounded. By the Spectral Inclusion of C_0-semigroups (see Pazy [17]), $\sigma(G) \subset \{z \in \mathbb{C} : Rez \leq 0\}$. Since $Re(\alpha + \lambda) = \alpha > 0$, by Lemma 2.3, $\alpha + \lambda \in \rho(G)$, and thus, (7.8) makes sense. Since Σ is countable, also $\sigma_u(G, u)$ is countable. The theorem is obtained by applying directly the ABLV Theorem 2.4 to the evolution semigroup (T^h) in $\mathrm{AP}_\Lambda(\mathbb{X})$. In fact, as shown above, (7.8) is exactly the condition that

$$\lim_{\alpha \downarrow 0} \alpha R(\alpha + \lambda, G)u = 0$$

for each $u \in \mathrm{AP}_\Lambda(\mathbb{X})$, $\lambda \in \Sigma$. This means that the evolution semigroup (T^h) is strongly stable in $\mathrm{AP}_\Lambda(\mathbb{X})$. By Theorem 3.4, this yields the strong stability of (7.3).□

7.3.1 Special Cases of Theorem 3.6

Below we will discuss several special cases of the above theorem.

Example 3.7. If $A(t) = 0$ for all t, then $\Lambda = \{0\}$. Therefore, $\mathrm{AP}_\Lambda(\mathbb{X})$ is nothing but the space of all constant functions, hence it can be identified with $\mathbb{X}$. The evolution semigroup associated with (7.3) is actually the semigroup e^{tA_0} generated by the operator A_0.

Therefore, the following corollary is obvious, and is the ABLV Theorem.

Corollary 3.8. *Let A_0 generate a uniformly bounded semigroup $T(t)$, and let $\sigma(A_0) \cap i\mathbb{R}$ be countable. Moreover, let for each $i\lambda \in \sigma(A_0) \cap i\mathbb{R}$ and $x \in \mathbb{X}$*

$$\lim_{\alpha \downarrow 0} \alpha R(i\lambda + \alpha, A_0)x = 0.$$

Then, Eq. (7.3) is strongly stable.

Example 3.9. Consider the linear evolution equation

$$\frac{dx}{dt} = A(t)x,$$

where $x \in \mathbb{X}$, $A(t)$ is a (not necessarily bounded) linear operator acting on $\mathbb{X}$ for every fixed t and $A(t+1) = A(t)$. Suppose further that the semi-module generated by Bohr spectrum of $A(\cdot)$ is actually a module. Then $\mathrm{AP}_\Lambda(\mathbb{X})$ is nothing but the space

of all 1-periodic functions, and the process $(U(t,s))_{t\geq s}$ is 1-periodic (see [16] for related concepts and results). By [16, Proposition 1] Eq. (7.6) has a unique solution in $\mathrm{AP}_\Lambda(\mathbb{X})$ if and only if $1 \in \rho(e^{-\lambda}U(1,0))$, or equivalently, $e^\lambda \in \rho(U(1,0))$. Let us denote by P the monodromy operator $U(1,0)$. It is well known that the strong stability of the 1-periodic evolutionary process $(U(t,s))_{t\geq s}$ can be studied via the stability of its monodromy operator P (see, for example, [21]). A discrete version of the ABLV Theorem on stability of individual orbits of the monodromy operator P is given in [25, Corollary 3.3]. We now show that the conditions of [25, Corollary 3.3] actually yield the conditions of our Theorem 3.6. In fact, the countability of $\Sigma[(7.3)]$ follows from the countability of $\sigma(P) \cap \Gamma$, where Γ is the unit circle of the complex plane. Next, let $i\lambda_0 \in \Sigma$ (7.3). Then, $z_0 := e^{i\lambda_0} \in \sigma(P) \cap \Gamma$. Therefore,

$$\lim_{z\downarrow z_0}(z - z_0)R(z,P)x_0 = 0. \tag{7.9}$$

We will show that (7.9) yields (7.8). By the definition of limit (7.9) in [25], (7.9) can be re-written as

$$\lim_{\alpha\downarrow 0}(e^{i\lambda_0+\alpha} - e^{i\lambda_0})R(e^{i\lambda_0+\alpha},P)x_0 = 0.$$

Or equivalently,

$$\lim_{\alpha\downarrow 0}(e^{\alpha} - 1)R(e^{i\lambda_0+\alpha},P)x_0 = 0,$$

and hence,

$$\lim_{\alpha\downarrow 0}\alpha R(e^{i\lambda_0+\alpha},P)x_0 = 0.$$

Let $f \in AP_\Lambda(\mathbb{X})$, that is, f is an arbitrary continuous 1-periodic function taking values in $\mathbb{X}$. Then, let

$$x_0 := \int_0^1 e^{-(i\lambda_0+\alpha)(1-\xi)}U(1,\xi)f(\xi)d\xi.$$

And let $u_{i\lambda_0+\alpha,f}$ be the unique solution to Eq. (7.6). Then, it is easy to check that

$$u_{i\lambda_0+\alpha,f}(0) = R(e^{i\lambda_0+\alpha},P)x_0.$$

Therefore, by the definition of the evolutionary process,

$$\begin{aligned}\|u_{i\lambda_0+\alpha,f}\| := \sup_{t\in\mathbb{R}}\|u_{i\lambda_0+\alpha,f}(t)\| &= \sup_{0\leq t\leq 1}\|u_{i\lambda_0+\alpha,f}(t)\| \\ &= Me^{\beta}\|u_{i\lambda_0+\alpha,f}(0)\| \\ &\leq Me^{\beta}\|R(e^{i\lambda_0+\alpha},P)x_0\|,\end{aligned}$$

where the positive numbers M, β depend only on the process $(U(t,s))_{t\geq s}$ (see Definition 2.5). This shows that (7.9) yields (7.8), and thus, the following result is a corollary to Theorem 3.6.

Corollary 3.10. *Let the monodromy operator P of the 1-periodic evolutionary process $(U(t,s))_{t\geq s}$ be a power bounded operator, i.e. $\sup_{n\in\mathbb{N}}\|P^n\| < \infty$, such that $\sigma(P)\cap\Gamma$ is a countable set. Moreover, assume that for each $\xi_0 \in \sigma(P)\cap\Gamma$ the following holds for each $x_0 \in \mathbb{X}$*

$$\lim_{\lambda\downarrow\xi_0}(\lambda-\xi_0)R(\lambda,P)x_0 = 0.$$

Then, for every $x_0 \in \mathbb{X}$ and for every $s \in \mathbb{R}$,

$$\lim_{t\to\infty} U(t,s)x_0 = 0.$$

In summary, Theorem 3.6 covers two well-known special cases of non-autonomous equations, including the autonomous and periodic cases. For the general case of non-autonomous equations the generator G of the evolution semigroup may have a more complicated spectrum, and we will discuss this topic in the next section.

7.4 Analysis of the Spectrum of the Generator G

As shown in the previous section, the spectrum of the generator G of the evolution semigroup $(T^h)_{h\geq 0}$ plays an important role in studying the asymptotic behavior of the equations. Moreover, in the autonomous and periodic cases this spectrum may not be the whole vertical strips. In this section we will give a detailed analysis of the spectrum of the generator G of the evolution semigroup associated with certain non-periodic equations and applications of the results obtained from the previous section.

Proposition 4.1. *Let (T^h) be the minimal evolution semigroup associated with (7.3), and G be its generator. Assume further that the semi-module generated by the set of frequencies of the function $A(\cdot)$ is actually a module. Then, for each $\lambda \in \Lambda$*

$$i\lambda + \sigma(G) \subset \sigma(G)$$
$$i\lambda + \rho(G) \subset \rho(G).$$

Proof. Let $\lambda \in \Lambda$. The above inclusions are actually equivalent to the claim that $\mu \in \rho(G)$ if and only if $i\lambda + \mu \in \rho(G)$. By the argument that precedes Theorem 3.6, $\mu \in \rho(G)$ if and only if for each $f \in \mathrm{AP}_\Lambda(\mathbb{X})$ the integral equation

$$x(t) = e^{-\mu(t-s)}U(t,s)x(s) + \int_s^t e^{-\mu(t-\xi)}U(t,\xi)f(\xi)d\xi \quad (t\geq s) \tag{7.10}$$

has a unique solution in $\mathrm{AP}_\Lambda(\mathbb{X})$, denoted by $u_{\mu,f}$. Since Λ is assumed to be a module, $u_{\mu,f} \in \mathrm{AP}_\Lambda(\mathbb{X})$ if and only if the function $v : \mathbb{R} \to \mathbb{R}, t \mapsto e^{i\lambda t}u_{\mu,f}$, belongs to $\mathrm{AP}_\Lambda(\mathbb{X})$. Moreover, $u_{\mu,f}$ is the unique solution of (7.10) if and only if v is the unique solution of the equation

$$y(t) = e^{-(i\lambda+\mu)(t-s)}U(t,s)y(s) + \int_s^t e^{-(i\lambda+\mu)(t-\xi)}U(t,\xi)f(\xi)d\xi \quad (t \geq s),$$

that is, $i\lambda + \mu \in \rho(G)$. □

In order to analyze the spectrum of the generator G of the minimal evolution semigroup associated with (7.3), in case the non-autonomous term $A(t)$ is small, it is useful to consider the generator $G_{0,\Lambda}$ of the evolution semigroup $(T^h_{0,\Lambda})$ associated with the equation $\dot{u} = A_0 u$ in the function space $\mathrm{AP}_\Lambda(\mathbb{X})$.

Lemma 4.2. *Assume that A_0 is a sectorial operator, and the semi-module Λ is a closed subset of the real line. Under the above assumption and notation,*

$$\sigma(G_{0,\Lambda}) = \sigma(A_0) - i\Lambda.$$

Proof. By the main results of [15],

$$\mu \in \rho(G_{0,\Lambda}) \Leftrightarrow \sigma(A_0 - \mu) \cap i\Lambda = \emptyset.$$

This condition means there are no complex numbers $\eta \in \sigma(A_0)$ and $\lambda \in \Lambda$ such that $\eta - \mu = i\lambda$, or, μ cannot be expressed as $\mu = \eta - i\lambda$ with $\eta \in \sigma(A_0)$ and $\lambda \in \Lambda$. In turn, this yields that $\mu \notin \sigma(A_0) - i\Lambda$, or, $\mu \in \mathbb{C}\backslash(\sigma(A_0) - i\Lambda)$. This proves the proposition. □

Recall that we are denoting by G the generator of the evolution semigroup associated with Eq. (7.3).

Proposition 4.3. *Assume that A_0 is a sectorial operator, and the semi-module Λ is a closed subset of the real line. Then, for each compact subset*

$$K \subset \rho(G_{0,\Lambda}) = \mathbb{C}\backslash(\sigma(A_0) - i\Lambda)$$

there exists a number $\delta_0 > 0$ such that if

$$\sup_{t\in\mathbb{R}} \|A(t)\| < \delta_0,$$

then

$$\sigma(G) \cap K = \emptyset. \tag{7.11}$$

Proof. Since K is a compact subset of $\rho(G_{0,\Lambda})$

$$\sup_{\lambda\in K}\|R(\lambda, G_{0,\Lambda})\| = \mu < \infty.$$

Let $\mathcal{A}$ denote the operator of multiplication by $A(t)$ in $\mathrm{AP}_\Lambda(\mathbb{X})$. Note that this multiplication operator is well defined in $\mathrm{AP}_\Lambda(\mathbb{X})$, and moreover, this operator is bounded. Therefore, there exists a positive δ_0 such that the operator

$$(I - R(\lambda, G_{0,\Lambda})\mathcal{A})^{-1}$$

whenever $\|\mathcal{A}\| < \delta_0$ and $\lambda \in K$. Next, we can show that for each $\lambda \in K$

$$(I - R(\lambda, G_{0,\Lambda})\mathcal{A})^{-1}R(\lambda, G_{0,\Lambda}) = R(\lambda, G_{0,\Lambda} + \mathcal{A}) = R(\lambda, G).$$

In fact, set

$$U := (I - R(\lambda, G_{0,\Lambda})\mathcal{A})^{-1}R(\lambda, G_{0,\Lambda}).$$

With this notation we have

$$(I - R(\lambda, G_{0,\Lambda})\mathcal{A})U = R(\lambda, G_{0,\Lambda}).$$

Therefore,

$$U = R(\lambda, G_{0,\Lambda}) + R(\lambda, G_{0,\Lambda})\mathcal{A}U,$$

and hence,

$$(\lambda - G_{0,\Lambda})U = I + \mathcal{A}U.$$

This yields

$$(\lambda - G_{0,\Lambda})U - \mathcal{A}U = I,$$

that is,

$$(\lambda - G_{0,\Lambda} - \mathcal{A})U = (\lambda - G)U = I.$$

In other words, $U = R(\lambda, G)$ whenever $\lambda \in K$ and $\|\mathcal{A}\| < \delta_0$. This yields (7.11). The proposition is proved. □

7.5 Examples

Consider the equation

$$\frac{dx}{dt} = a(t)x,$$

where $a(t)$ is a *numerical* almost periodic function taking values in $\mathbb{X} := \mathbb{R}$. Define the operator $G = -d/dt + a(t)$. We will describe the part of spectrum $\sigma(G) \cap i\mathbb{R}$.

Theorem 5.1. *Suppose that*

(i) *the semi-module generated by the Bohr spectrum of* $a(t)$*, i.e.* $\Lambda := sm(\sigma_b(a))$*, is a discrete countable set and*
(ii) $0 \notin \sigma_b(a)$.

Then Σ *is also a discrete countable set. Moreover*

$$\Sigma \subset -i\Lambda \cup i\Lambda.$$

In order to prove Theorem 5.1, we need the following two lemmas.

Lemma 5.2. *Let* $f \in \mathrm{AP}(\mathbb{R})$ *such that* $|\lambda_n| \geq M > 0$ *for all* $\lambda_n \in \sigma_b(f)$*. Then* $g(\cdot) := \int_0^{\cdot} f(s)ds \in \mathrm{AP}(\mathbb{R})$ *and* $\sigma_b(g) \subset \sigma_b(f) \cup \{0\}$.

Proof. This Lemma is a direct consequence of [11, Theorem 4.12] and [11, Theorem 5.2]. Indeed, [11, Theorem 4.12] stated that $g(t) = \int_0^t f(s)ds$ is an almost periodic function. Therefore $g(t)$ is an almost periodic solution to equation $x'(t) = f(t)$, and by [11, Theorem 5.2], $\sigma_b(g) \subset \sigma_b(f) \cup \{0\}$. □

Lemma 5.3. $e^{\int_0^{\cdot} a(s)ds} \in \mathrm{AP}_\Lambda(\mathbb{R})$.

Proof. First, note that if we replace $sm(\sigma_b(a))$ with $m(\sigma_b(a))$—the module generated by $\sigma_b(a)$, then the claim is clear from [11, Theorem 1.9, p.5]. However, in general $sm(\sigma_b(a))$ may differ from $m(\sigma_b(a))$. Since $sm(\sigma_b(a))$ is a discrete set, so is $\sigma_b(a)$. From the assumption $0 \notin \sigma_b(a)$, we get that $\sigma_b(a)$ is bounded away from zero. Applying Lemma 5.2,

$$g(\cdot) := \int_0^{\cdot} a(t)dt \in \mathrm{AP}_\Lambda(\mathbb{R}).$$

We have

$$e^{g(t)} = 1 + g(t) + \frac{g^2(t)}{2} + \frac{g^3(t)}{3!} + \cdots + \frac{g^n(t)}{n!} + \cdots$$

Since $g \in \mathrm{AP}_\Lambda(\mathbb{R})$ we have that g is bounded, so the above infinite sum is uniformly convergent. Also $g^n \in \mathrm{AP}_\Lambda(\mathbb{R})$ due to the fact that $\mathrm{AP}_\Lambda(\mathbb{R})$ is closed under products (see [11, Theorem 1.9, p. 5]) and $\sigma_b(g^n) \in \Lambda$ as an application of the Approximation Theorem. Since the uniform limit of a sequence of almost periodic functions is also an almost periodic function, $e^{g(\cdot)} \in \mathrm{AP}(\mathbb{R})$. On the other hand, $\mathrm{AP}_\Lambda(\mathbb{R})$ is a closed subspace of $\mathrm{AP}(\mathbb{R})$. Therefore, $e^{g(\cdot)} \in \mathrm{AP}_\Lambda(\mathbb{R})$. □

Proof of Theorem 5.1. It is sufficient to prove that for every real number $\lambda \notin -\Lambda \cup \Lambda$, the equation

$$\frac{dx}{dt} = (a(t) - i\lambda)x + f(t) \tag{7.12}$$

has a unique solution $x \in \mathrm{AP}_\Lambda(\mathbb{R})$ for every $f \in \mathrm{AP}_\Lambda(\mathbb{R})$.

Let $y(t) = e^{i\lambda t}x(t)$ or $x(t) = e^{-i\lambda t}y(t)$, then (7.12) becomes

$$\frac{dy}{dt} = a(t)y + e^{i\lambda t}f(t), \tag{7.13}$$

which has a general solution

$$y(t) = e^{\int_0^t a(s)ds}\left(\int_0^t e^{i\lambda\tau}f(\tau)e^{-\int_0^\tau a(s)ds}d\tau + C_0\right).$$

By applying Lemma 5.3, one has $f(\cdot)e^{-\int_0^\cdot a(s)ds} \in \mathrm{AP}_\Lambda(\mathbb{R})$. Now suppose that

$$\lambda_n \in \sigma_b\left(e^{i\lambda\tau}f(\tau)e^{-\int_0^\tau a(s)ds}\right),$$

then $\lambda_n = \lambda + \mu_n$ for some $\mu_n \in \Lambda$. Since $\lambda \notin -\Lambda$, it follows that $\lambda_n \neq 0$ for all n. Because Λ is a discrete set, $e^{i\lambda\tau}f(\tau)e^{-\int_0^\tau a(s)ds}$ is an almost periodic function with Bohr spectrum bounded away from zero.

By applying Lemma 5.2, it follows that $G(\cdot) = \int_0^\cdot e^{i\lambda\tau}f(\tau)e^{-\int_0^\tau a(s)ds}d\tau \in \mathrm{AP}(\mathbb{R})$. In fact $G \in \mathrm{AP}_{\Lambda+\{\lambda\}}(\mathbb{R})$, where $\Lambda + \{\lambda\}$ is the set of all real numbers μ of the form

$$\mu := m + \lambda, \quad m \in \Lambda.$$

Therefore the unique solution in $\mathrm{AP}_\Lambda(\mathbb{R})$ of (7.12) is

$$x(t) = e^{-i\lambda t}y_0(t),$$

where

$$y(t) = e^{\int_0^t a(s)ds}\int_0^t e^{i\lambda\tau}f(\tau)e^{-\int_0^\tau a(s)ds}d\tau \quad (C_0 = 0).$$

□

Remark 5.4. The condition $0 \notin \sigma_b(a)$ is essential. Otherwise the solution

$$y(t) = e^{\int_0^t a(s)ds}\left(\int_0^t e^{i\lambda\tau} f(\tau) e^{-\int_0^\tau a(s)ds} d\tau + C_0\right)$$

is even unbounded. For instance, choose $a(t) := 1 + e^{it}$, then $\int_0^t a(s)ds = t + \frac{e^{it}}{i}$ is unbounded.

Example 5.5. Consider the equation

$$\frac{dx(t)}{dt} = (e^{it} + e^{i\sqrt{2}t})x(t), \quad x(t) \in \mathbb{C}, t \in \mathbb{R}. \tag{7.14}$$

The frequencies of $a(t) := (e^{it} + e^{i\sqrt{2}t})$ are $\{1, \sqrt{2}\}$. The semi-module generated by the set of frequencies of $a(t)$ is the set $\mathbb{N}_0 + \sqrt{2}\mathbb{N}_0$. With the operator $G = -d/dt + a(t)$ one gets

$$\sigma(G) \cap i\mathbb{R} \subset -i(\mathbb{N}_0 + \mathbb{N}_0\sqrt{2}) \cup i(\mathbb{N}_0 + \mathbb{N}_0\sqrt{2}),$$

where $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. In this example $U(t, s) = e^{\int_s^t a(\tau)d\tau}$, so all solutions $x(t) = U(t, s)x(s)$ are almost periodic, and therefore bounded.

Remark 5.6. Theorem 5.1 holds not only for one dimensional case, but also for finite dimensional case, i.e., for equation

$$\frac{dx}{dt} = A(t)x,$$

where $A(t)$ is an almost periodic matrix and $\mathbb{X} = \mathbb{R}^n$ or $\mathbb{C}^n$.

Example 5.7. In the following we give a numerical example to illustrate the conditions required in our Theorem 3.6. The equation we consider is of the form

$$\frac{dx(t)}{dt} = (\cos t + \cos t\sqrt{2} - 2)x(t), \quad x(t) \in \mathbb{R}, t \in \mathbb{R}_+. \tag{7.15}$$

The frequencies of $a(t) := \cos t + \cos t\sqrt{2} - 2$ are $\{0, -1, 1, -\sqrt{2}, \sqrt{2}\}$, therefore $\Lambda := sm(\sigma_b(a)) = m(\sigma_b(a)) = \mathbb{Z} + \mathbb{Z}\sqrt{2}$. We will show that

$$\Sigma = i\Lambda.$$

We have $M_{0,a} := \lim\limits_{T\to\infty} \frac{1}{T}\int_0^T a(t)dt = -2$, so $\mathrm{Re}M_{0,a} = -2 \neq 0$. It is shown in [11, Theorem 6.6] that for all $\lambda \notin \Lambda$, the equation

$$\frac{dx}{dt} = (a(t) - i\lambda)x + f(t),$$

or with $y = e^{i\lambda t}x$,

$$\frac{dy}{dt} = a(t)y + e^{i\lambda t}f(t)$$

has a unique bounded solution

$$y(t) = -\int_t^\infty e^{\int_s^t a(\tau)d\tau} e^{i\lambda s} f(s)ds$$

which is almost periodic and in $\mathrm{AP}_{\Lambda+\lambda}(\mathbb{X})$. It yields that $\Sigma \subset i\Lambda$. We now verify the condition (7.8). Since $U(t,s) = e^{\int_s^t a(\tau)d\tau}$, (7.6) becomes (for $s = 0$)

$$u_{\lambda,f} = e^{-\lambda t} e^{\int_0^t a(\tau)d\tau} u(0) + \int_0^t e^{-\lambda(t-\xi)} e^{\int_\xi^t a(\tau)d\tau} f(\xi)d\xi.$$

Therefore for each $i\lambda \in \Sigma$,

$$\begin{aligned} u_{\alpha+i\lambda,f} &= e^{-(\alpha+i\lambda)t} e^{\int_0^t a(\tau)d\tau} u(0) + \int_0^t e^{-(\alpha+i\lambda)(t-\xi)} e^{\int_\xi^t a(\tau)d\tau} f(\xi)d\xi \\ &= e^{-(2+\alpha)t} e^{-i\lambda t + \sin t + \frac{\sin\sqrt{2}t}{\sqrt{2}}} \left[u(0) + \int_0^t e^{i\lambda\xi - \sin\xi - \frac{\sin\sqrt{2}\xi}{\sqrt{2}}} f(\xi) e^{2\xi} d\xi \right] \end{aligned}$$

Since $e^{-i\lambda t + \sin t + \frac{\sin\sqrt{2}t}{\sqrt{2}}}$ and $e^{i\lambda\xi - \sin\xi - \frac{\sin\sqrt{2}\xi}{\sqrt{2}}} f(\xi)$ are almost periodic in t and ξ, respectively, there exist positive constants M and N such that

$$|e^{-i\lambda t + \sin t + \frac{\sin\sqrt{2}t}{\sqrt{2}}}| \le M, \quad |e^{i\lambda\xi - \sin\xi - \frac{\sin\sqrt{2}\xi}{\sqrt{2}}} f(\xi)| \le N.$$

We have

$$\begin{aligned} |u_{\alpha+i\lambda,f}| &\le e^{-(2+\alpha)t} M \left[|u_0| + N \int_0^t e^{(2+\alpha)\xi} d\xi \right] \\ &\le e^{-(2+\alpha)t} M|u_0| + e^{-(2+\alpha)t} MN \frac{e^{(2+\alpha)t} - 1}{2+\alpha} \\ &\le M|u_0| + MN \frac{1 - e^{-(2+\alpha)t}}{2+\alpha} \\ &\le M|u_0| + \frac{MN}{2+\alpha}. \end{aligned}$$

Therefore $\lim_{\alpha\downarrow 0} \alpha u_{\alpha+i\lambda,f} = 0$. Obviously, the set Σ is countable, so according to Theorem 3.6, Eq. (7.15) is strongly stable.

Example 5.8. In the following we give an example in the infinite dimensional case in which the spectrum of the generator of the evolution semigroup is a countable set, so it is possible to apply our result in the paper. Assume that A_0 is the generator of an analytic semigroup in a complex Banach space $\mathbb{X}$ (so, it is a sectorial operator), and $a(t) := e^{it} + e^{i\sqrt{2}t}$ is a numerical function. Consider the evolution equation

$$\frac{du(t)}{dt} = A_0u(t) + a(t)u(t), \quad u(t) \in \mathbb{X}, t \in \mathbb{R}. \tag{7.16}$$

We define $\Lambda := \mathbb{N}_0 + \sqrt{2}\mathbb{N}_0$. And as in the previous examples and results we know that an evolution semigroup associated with this equation is well defined and strongly continuous in the function space $\mathrm{AP}_\Lambda(\mathbb{X})$. Let us consider the multiplication operator M_A as defined in the proof of Theorem 3.1, that is, $M_A : \mathrm{AP}_\Lambda(\mathbb{X}) \to \mathrm{AP}_\Lambda(\mathbb{X})$ defined as $M_Af(t) = a(t)f(t)$ for all $t \in \mathbb{R}$. By using the spectral estimates of commuting operators developed in [2] it is easy to see that in this case the spectrum of the generator of the evolution semigroup $\overline{-d/dt + M_A + A_0}$ in $\mathrm{AP}_\Lambda(\mathbb{X})$ can be estimated as

$$\sigma(\overline{-d/dt + M_A + A_0}) \subset \sigma(-d/dt + M_A) + \sigma(A_0). \tag{7.17}$$

If we assume that $\sigma(A_0) \subset i\mathbb{R}$ is countable, then by the result in the previous example, we have

$$\sigma_i(\overline{-d/dt + M_A + A_0}) \subset \mathbb{N}_0 + \sqrt{2}\mathbb{N}_0 + \sigma(A_0) \tag{7.18}$$

is countable. Therefore, the imaginary spectrum of the generator of the minimal evolution semigroup is countable.

Acknowledgements The first author was supported by the Vietnamese Ministry of Education and Training (MOET) Scholarship Scheme (Project 322) and the Graduate Academy (GA) of the TU Dresden (PSPElement: F-00361-553-52A-2330000) in accordance with the funding regulations of the German Research Foundation (DFG). The second author was supported by DFG under grant number Si801/6-1 and NAFOSTED under grant number 101.02-2011.47.

References

1. Arendt, W., Batty, C.J.K.: Asymptotically almost periodic solutions of inhomogeneous Cauchy problems on the half-line. Bull. Lond. Math. Soc. **31**, 291–304 (1999)
2. Arendt, W., Rabiger, F., Sourour, A.: Spectral properties of the operator equation $AX+XB = Y$. Q. J. Math. Oxford Ser. (2) **45** (178), 133–149 (1994)
3. Arendt, W., Batty, C.J.K., Hieber, M., Neubrander, F.: Vector-Valued Laplace Transforms and Cauchy Problems. Monographs in Mathematics, vol. 96. Birkhäuser Verlag, Basel (2001)
4. Ballotti, M.E., Goldstein, J.A., Parrott, M.E.: Almost periodic solutions of evolution equations. J. Math. Anal. Appl. **138**, 522–536 (1989)

5. Basit, B.: Harmonic analysis and asymptotic behavior of solutions to the abstract Cauchy problem. Semigroup Forum **54**, 58–74 (1997)
6. Batty, C.J.K., Van Neerven, J., Räbiger, F.: Local spectra and individual stability of uniform bounded C_0-semigroups. Trans. Am. Math. Soc. **350**, 2071–2085 (1998)
7. Boumenir, A., Van Minh, N., Kim Tuan, V.: Frequency modules and nonexistence of quasi-periodic solutions of nonlinear evolution equations. Semigroup Forum **76**, 58–70 (2008)
8. Chicone, C., Latushkin, Y.: Evolution Semigroups in Dynamical Systems and Differential Equations. Mathematical Surveys and Monographs, vol. 70. American Mathematical Society, Providence, RI (1999)
9. Chill, R., Tomilov, Y.: Stability of operators semigroups: ideas and results. In: Perspectives in Operator Theory, vol. 75, pp. 71–109. Banach Center Publications, Polish Academy Science, Warszawa (2007)
10. Ellis, R., Johnson, R.A.: Topological dynamics and linear differential systems. J. Differ. Equ. **44**, 21–39 (1982)
11. Fink, A.M.: Almost Periodic Differential Equations. Springer, Berlin/Heidelberg/New York (1974)
12. Hino, Y., Naito, T., Van Minh, N., Shin, J.S.: Almost Periodic Solutions of Differential Equations in Banach Spaces. Taylor & Francis, London/New York (2002)
13. Johnson, R.A., Sell, G.R.: Smoothness of spectral subbundles and reducibility of quasiperiodic linear differential systems. J. Differ. Equ. **41**, 262–288 (1981)
14. Levitan, B.M., Zhikov, V.V.: Almost Periodic Functions and Differential Equations. Moscow University Publishing House, Moscow (1978). English translation by Cambridge University Press, Cambridge, UK (1982)
15. Murakami, S., Naito, T., Van Minh, N.: Evolution semigroups and sums of commuting operators: a new approach to the admissibility theory of function spaces. J. Differ. Equ. **164**, 240–285 (2000)
16. Naito, T., Van Minh, N.: Evolutions semigroups and spectral criteria for almost periodic solutions of periodic evolution equations. J. Differ. Equ. **152**, 358–376 (1999)
17. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Springer, New York (1983)
18. Preda, C.: $(L^p(\mathbb{R}_+, X), L^q(\mathbb{R}_+, X))$-admissibility and exponential dichotomies of cocycles. J. Differ. Equ. **249**, 578–598 (2010)
19. Preda, P., Pogan, A., Preda, C.: Schäffer spaces and exponential dichotomy for evolutionary processes. J. Differ. Equ. **230**, 378–391 (2006)
20. Preda, C., Preda, P., Craciunescu, A.: Criterions for detecting the existence of the exponential dichotomies in the asymptotic behavior of the solutions of variational equations. J. Funct. Anal. **258**, 729–757 (2010)
21. Quoc Phong, V.: Stability and almost periodicity of trajectories of periodic processes. J. Differ. Equ. **115**, 402–415 (1995)
22. Sacker, R.J., Sell, G.R.: A spectral theory for linear differential systems. J. Differ. Equ. **27**, 320–358 (1978)
23. Thieu Huy, N.: Exponentially dichotomous operators and exponential dichotomy of evolution equations on the half-line. Int. Equ. Oper. Theory **48**, 497–510 (2004)
24. Thieu Huy, N.: Exponential dichotomy of evolution equations and admissibility of function spaces on a half-line. J. Funct. Anal. **235**, 330–354 (2006)
25. Van Minh, N.: Asymptotic behavior of individual orbits of discrete systems. Proc. Am. Math. Soc. **137**, 3025–3035 (2009)

26. Van Minh, N.: A spectral theory of continuous functions and the Loomis-Arendt-Batty-Vu theory on the asymptotic behavior of solutions of evolution equations. J. Differ. Equ. **247**, 1249–1274 (2009)

27. Van Minh, N., Räbiger, F., Schnaubelt, R.: On the exponential stability, exponential expansiveness and exponential dichotomy of evolution equations on the half line. Int. Equ. Oper. Theory **32**, 332–353 (1998)

28. van Neerven, J.M.A.M.: The asymptotic behaviour of semigroups of linear operator. In: Operator Theory, Advances and Applications, vol. 88. Birkhäuser Verlag, Basel /Boston/Berlin (1996)

Chapter 8
A Preparation Theorem for a Class of Non-differentiable Functions with an Application to Hilbert's 16th Problem

Mohamed El Morsalani and Abderaouf Mourtada

Abstract We consider a class of unfoldings of quasi-regular functions. We assume that such perturbations have asymptotic developments which depend on many unfoldings of the logarithm function. We prove a preparation theorem for such functions; namely, they are "conjugated" to a finite principal part via a "pseudo-isomorphism". This finite principal part is polynomial in the phase variable and these unfoldings of the logarithm function. As an application there exists a uniform bound in the parameter of the numbers of zeros of such class of non-differentiable functions. A finiteness result of the number of the limit cycles bifurcating from a perturbed hyperbolic polycycle is obtained too.

Keywords Hilbert 16th problem • Malgrange preparation theorem • Quasi-regular function • Pseudo-isomorphism • Chebychev expansion • Hyperbolic saddle point

8.1 Introduction

The Malgrange Preparation theorem, in its elementary form, says that if $\delta(x, \lambda)$ is smooth with respect to the phase variable x and the parameter λ and if it satisfies $\delta(x, \lambda_0) = x^k g(x)$ with $g(0) \neq 0$ i.e. $x = 0$ is a regular zero of $\delta(x, \lambda_0)$ of order k then there exist a neighborhood V of $(0, \lambda_0)$ and smooth functions u, a_i where $u(x, \lambda) \neq 0$ such that

$$\delta(x, \lambda) = u(x, \lambda)(x^k + a_{k-1}(\lambda)x^{k-1} + \cdots + a_0(\lambda)) \tag{8.1}$$

M. El Morsalani (✉)
Landesbank Baden-Württemberg, 70173 Stuttgart, Germany
e-mail: mohamed.EL-Morsalani@LBBW.DE

A. Mourtada
(Deceased)

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_8

Such theorem plays an important role in the theory of bifurcations. In this paper we show an equivalent preparation theorem for a class of non-differentiable functions which are unfoldings of the class of quasi-regular functions $\mathcal{QR}$.

Definition 1.1. Let $g(x) : [0, x_0] \longrightarrow \mathbb{R}$ be an analytic function for $x > 0$ and continuous at $x = 0$. we say that g is *quasi-regular* if:

($\mathcal{QR}1$) $g(x)$ has a formal expansion of *Dulac type*. This means that there exists a formal series:

$$\hat{g}(x) = \sum_{i=0}^{\infty} x^{\beta_i} P_i(\ln x)$$

where β_i is strictly increasing sequence of positive real numbers $0 < \beta_0 < \beta_1 \ldots$ tending to infinity and for i, P_i is a polynomial and $\hat{g}$ is a formal expansion of $g(x)$ in the following sense:

$$\forall n \geq 0 \quad g(x) - \sum_{i=0}^{n} x^{\beta_i} P_i(\ln x) = 0(x^{\beta_n}).$$

($\mathcal{QR}2$) Let $G(\xi) = g(\exp^{-\xi})$ for $\xi \in [\xi_0 = -\ln x_0, \infty[$. Then G has a bounded *holomorphic extension* in a complex domain $\Omega(C) = \{\zeta = \xi + i\eta; \quad \xi^4 \geq C(1 + \eta^2)\}$ for some $C > 0$.

We begin by defining the class of functions we will study. Let δ be defined on $[0, x_0] \times W$ into $[0, y_0]$ where $x_0, y_0 > 0$ and W is a neighborhood of $\lambda = 0$ in $\mathbb{R}^{\Lambda}$. Then δ is an unfolding of a quasi-regular function if it satisfies the following:

($\mathcal{H}1$) It is analytic on $(0, x_0] \times W$, only continuous at $x = 0$ and $\delta(x, \lambda) = 0 \ \forall \lambda$.

($\mathcal{H}2$) It has asymptotic developments that are unfoldings of Dulac's formal expansions. Let $\omega_i(x, \lambda)$ be the unfolding of the logarithm function

$$\omega_i(x, \lambda) = \begin{cases} \frac{x^{-\mu_i} - 1}{\mu_i} & \text{if} \quad \mu_i \neq 0 \\ -\ln x & \text{if} \quad \mu_i = 0 \end{cases} \tag{8.2}$$

where μ_i is a coordinate of the multi-parameter λ. The asymptotic developments are defined as follows: for any $k \in \mathbb{N}^*$ there exists a neighbourhood $W_k \subset W$ of $0 \in \mathbb{R}^{\Lambda}$ and a set $\overline{I} = [0, \epsilon]$ for some $\epsilon > 0$ such that

$$\begin{aligned} &\delta(x, \lambda) = \delta_k(x, \lambda) + x^k R_k(x, \lambda) \\ &\text{with} \\ &\delta_k(x, \lambda) = \sum_{0 \leq i + \sum j_l \leq k} \gamma_{i j_1 \ldots j_m} x_i Z_1^{j_1} \cdots Z_m^{j_m} \end{aligned} \tag{8.3}$$

where

$$Z_i(x, \mu_i) = x\omega_i$$

and R_k is analytic on $(0, \epsilon]$ and satisfies the following property (I_0^∞) which was introduced in [11]:

$$(I_0^\infty) \qquad \forall \nu \in \mathbb{N} \quad \lim_{x \to 0} x^\nu \frac{\partial^\nu R_k}{\partial x^\nu}(x, \lambda) = 0 \quad \text{uniformly in} \quad \lambda \in W_k.$$

$(\mathcal{H}3)$ There exists an integer $n = n_0 + n_1 + \cdots + n_m$ (the nontriviality order) such that

$$\delta(x, \lambda_0) = cx^{(n_0+n_1+\cdots+n_m)}(\ln x)^{(n_1+\cdots+n_m)} + \text{h.o.t} \qquad c \neq 0 \tag{8.4}$$

this means that for $\lambda = \lambda_0 = 0$ the function δ is not formally flat.

We need to introduce some notations before stating the main theorem for this class of non-differentiable functions. For $\epsilon > 0$ the set S_ϵ is defined by

$$S_\epsilon = \{u + \sqrt{-1}v;\ 0 < u < \epsilon,\ |v| < \epsilon|u|\}.$$

A map $f : X \longrightarrow Y$ is said to be of finite degree if there exists $d \in \mathbb{N}$ such that for any $y \in Y$ we have $\text{Card}\{f^{-1}(y)\} \leq d$.

Theorem 1.2 (Main Theorem). *There exist an integer N depending on $n = n_0 + n_1 + \cdots + n_m$, a neighbourhood $W_N \subset W$ of $0 \in \mathbb{R}^\Lambda$, a set $\overline{I} = [0, \epsilon]$ for some $\epsilon > 0$ and a map, which we call a "pseudo-isomorphism", $\Phi_N : I \times W_N \longrightarrow S_\epsilon \times W_N$ of finite degree such that*

$$\delta = \delta_N \circ \Phi_N, \tag{8.5}$$

where

$$\Phi_N(x, \lambda) = (\varphi_N(x, \lambda), \lambda) = (x(1 + \rho(x, \lambda)), \lambda) \tag{8.6}$$

and

$$\lim_{x \to 0} \rho(x, \lambda) = 0 \quad \textit{uniformly in } \lambda$$

Corollary 1.3 (Main Corollary). *If δ satisfies $(\mathcal{H}1)$, $(\mathcal{H}2)$ and $(\mathcal{H}3)$, then there exists a uniform bound in the parameter λ for the number of isolated zeros of the equation $\delta(x, \lambda) = 0$.*

Remark. Joyal has published a preparation theorem for non-differentiable functions having a certain type of expansion called Chebychev expansions [9]. The class of functions we consider in this work is more general.

This work has its first motivations in Hilbert's 16th problem for polynomial vector fields [5]:

Find the maximum number $H(n)$ and relative positions of limit cycles of a system

$$\begin{cases} \dot{x} &= P(x, y) \\ \dot{y} &= Q(x, y) \end{cases}$$

of degree $n = \sup(\deg P, \deg Q)$. The existential part of the problem is to prove that $H(n)$ is finite.

We begin by giving some definitions:

– The Hausdorff distance between compact sets of the sphere $\mathbb{S}^2$ is defined by

$$d_H(A, B) = \sup_{(x,y)\in A\times B} \{\text{dist}(x, B), \text{dist}(A, y)\}$$

– A compact set $\Gamma \subset \mathbb{S}^2$ is a limit periodic set (l.p.s) of a vector field X_{λ_0} if there is a sequence $(\lambda_i)_{i\geq 1} \to \lambda_0$ in the parameter space P, with a limit cycle γ_i for each X_{λ_i} such that: $(\gamma_i) \to \Gamma$ for d_H.

– A limit periodic set Γ of X_{λ_0} has finite cyclicity if there exist $N \in \mathbb{N}$, $\epsilon, \delta > 0$ such that any X_λ with $|\lambda - \lambda_0| < \delta$ has at most N limit cycles γ_i satisfying $d_H(\gamma_i, \Gamma) < \epsilon$. The minimum of such N when $\epsilon, \delta \to 0$ is called the cyclicity of Γ in the family X_λ.

– A singular point of a vector field is said *elementary* if it has at least one non-zero eigenvalue. It is hyperbolic (respectively *semi-hyperbolic*) if the two eigenvalues are non-zero (respectively one eigenvalue is zero).

– The *hyperbolicity ratio* of a hyperbolic saddle is the ratio $r = -\frac{\alpha_1}{\alpha_2}$, where $\alpha_1 < 0 < \alpha_2$ are the two eigenvalues. The hyperbolic saddle is *attracting* (respectively *repelling*, *neutral*) if $r > 1$ (respectively $r < 1$, $r = 1$).

– A graphic [4] is formed by singular points $p_1, \cdots, p_m, p_{m+1} = p_1$ and oriented regular orbits $s_1, \cdots, s_m$ connecting them so that s_j is an unstable characteristic orbit of p_j and a stable characteristic orbit of p_{j+1}, and normal orientations n_j of the regular orbits are coherent in the sense that if s_{j-1} has left-hand orientation then so does s_j. Graphics may or may not have a return map. Polycycles are graphics with a return map.

– A graphic is called *elementary* if all its singular points are elementary.

In studying this problem it is natural to compactify the phase space to the Poincaré sphere. The parameter space can be compactified as well. We obtain a family X_λ, $\lambda \in \Lambda$ of analytic vector fields defined on a compact phase space and depending on parameters λ varying in a compact set Λ. Using a compacity argument Roussarie [14] showed that to prove the existential part of Hilbert's 16th problem for polynomial vector fields it is sufficient to prove that any limit periodic set in the family X_λ has finite cyclicity, i.e. can give rise to uniformly bounded number of limit cycles in any perturbation inside X_λ [1]. In this context, we give a first application of the theory developed in the first part of this paper: uniform finiteness of the number of limit cycles which bifurcate from a degenerate hyperbolic 2-polycycle. The main theorem in the application is

Theorem 1.4. *Let X_0 be an analytic vector field on the plane which has a hyperbolic polycycle Γ with two vertices such that their ratios of hyperbolicity are rational different from* 1 *and their product is* 1. *If the polycycle is non-identical, then there exists an integer N depending only on the germ of X_0 along Γ such that the number of limit cycles bifurcating from Γ in any analytic deformation is bounded by N.*

Another problem is treated in our work, namely the degeneracy of Khovansii procedure as used by Ilyashenko and Yakovenko to prove the following theorem:

Theorem 1.5 ([8]). *For any $n \in \mathbb{N}$ there exists a number $E(n)$ such that an elementary polycycle appearing in a generic n-parameter family of smooth vector fields generate no more than $E(n)$ limit cycles. The number $E(n)$ may be effectively estimated by some primitive recursive function of n.*

One of the important steps to prove this theorem is the called Khovanskii procedure. One result of our paper is to show how this Khovanskii procedure used by Ilyashenko and Yakovenko degenerates when we study perturbations of a "generic" hyperbolic 2-polycycle. It follows that one has to impose strong conditions to get the finite cyclicity. To overcome such difficulties, we think that these new ideas which were elaborated in the preparation theorem for unfoldings of quasi-regular functions will be of great help.

Many questions are still open: what about the properties of the "pseudo-isomorphism". We could not answer this question yet. But an answer will be of great progress towards a best knowledge of the bifurcation diagram of $\delta(x, \lambda) = 0$.

To make our proofs understandable we show our main theorem for the case m=3 meaning we work with three $Z_i = x\omega_i$. The proofs are technical but constructive. The paper is organized as follows: in the second section we show that the equation $\delta(x, \lambda) = \delta_N(\varphi_N(x, \lambda), \lambda)$ is equivalent to an analytic functional equation in the unknown function ρ

$$h_N(x, \lambda, \rho) = \sum_{l=1}^{+\infty} \hat{a}_l(x, \lambda)\rho^l - x^N R_N(x, \lambda) = 0 \qquad (*)$$

where

$$\hat{a}_l(x, \lambda) = \frac{1}{l!} \frac{\partial \delta_N}{\partial x^l}(x, \lambda).$$

To show that the functional equation (*) has a solution, we will study the "ideal" $\mathcal{I}_0$ generated by the germs of the functions $\hat{a}_n$ on $(\mathbb{R}^+, 0) \times (\mathbb{R}^{\Lambda+3}, 0)$. This "ideal" will be the set obtained as a restriction of a certain ideal, in the ring of analytic germs $(\mathbb{R}^+, 0) \times (\mathbb{R}^{\Lambda+3}, 0)$, on the graphs given by $z_i = Z_i(x, \mu_i)$. The key theorem will be

Theorem 1.6. *There exists an integer $m \in \mathbb{N}$ such that $x^{m(1+r_1+r_2+r_3)}$ is in $\mathcal{I}_0$ where $r_i = 1 - \mu_i$ and μ_i is the parameter in (8.2) and the integer m can be explicitly computed.*

The next section will be devoted to prove this theorem. This is done in progressive steps. We will study the properties of monomials appearing in the asymptotic developments in ($\mathcal{H}3$). These monomials are $\psi_{ij_1j_2j_3} = x^i Z_1^{j_1} Z_2^{j_2} Z_3^{j_3}$. We introduce a derivation operator $\mathcal{L} = x\frac{\partial}{\partial x}$ which will be applied in an elimination algorithm to show that there exists a function h in this "ideal" such that

$$h(x,\lambda) = \sum_{i+j_1+j_2+j_3=n} \gamma_{ij_1j_2j_3}(\lambda)\psi_{ij_1j_2j_3}$$

and $h(x,0)$ is not equivalent to zero. Thanks to some linear algebra, differential calculus and analytic geometry like the idea of using the coefficient ideal associated with a perturbation of an analytic function on an interval, we succeed to prove that a function $b(\mu)x^{\nu(\mu)}$ is in this "ideal" but we could not say anything about the germ of $b(\mu)$ in $\mu = (\mu_1,\mu_2,\mu_3) = 0$. After some preparatory work we show the last theorem and later on we construct the function ρ. The key idea is to notice that when one chooses an integer N sufficiently large then $x^N R_N(x,\lambda)$ in expression (*) is in this "ideal" which is finitely generated. Tools from complex analysis like the fact that the ring of analytic germs at a certain point is noetherin, some compacity arguments as theorem of Rouché, formula of Jensen will be used. Moreover the theory of Khovanskii as in [13] will be applied to prove the main corollary. The last part will be devoted to the application. In Sect. 8.5, we will recall some known results about transition maps near a hyperbolic saddle point of finite order, i.e. it is not formally linearisable. In the Sect. 8.6, we show the theorem:

Theorem 1.7. *Let X_0 be an analytic vector field on the plane which has a hyperbolic polycycle Γ with two vertices such that their ratios of hyperbolicity are rational and their product is 1. If the polycycle is non-identical satisfying some conditions which will be exhibited later on and if one of its vertices has finite order then there exists an integer N depending only on the germ of X_0 along Γ such that the cyclicity of Γ in any C^∞-deformation is bounded by N.*

The proof uses an adequate definition of the displacement map associated with the perturbed polycycle. For this displacement map, one can apply the pfaffian equation satisfied by the Dulac map of one of the saddle vertices. This yields a simplified equation which can be completely studied. The last section contains the proof of the main theorem in the application mentioned previously. There we use the preparation theorem after reducing the displacement map by a new one by means of generalized Rolle's lemma and some differential analysis: change of variables, introduction of new compensators.

8.2 Reduction of the Equation $\delta = \delta_N \circ \Phi_N$ to a Functional Equation

For any sufficiently small $x_0 \in (0,\epsilon)$ the functions δ and δ_N are analytic on $]0,2x_0] \times W_N$. Furthermore these functions can be holomorphically continued on sectors $S(\theta) = \{z \in \tilde{\mathbb{C}};\ 0 < |z| < \epsilon(\theta),\ |\mathrm{Arg}(z)| < \theta\}$ where $\epsilon(\theta) \to 0$ if $\theta \to \infty$ and $\tilde{\mathbb{C}}$ is the universal covering of $\mathbb{C}^*$. This property is satisfied by Dulac maps of hyperbolic saddle points and their unfoldings [3]. It follows that for any $\lambda \in W_N$ the radius of convergence of their power series expansion in $(x - x_0)$ is $|x_0|$. Let

$$\varphi_N(x,\lambda) = x(1+\rho(x,\lambda)) \tag{8.7}$$

$$\begin{aligned}\delta_N(\varphi_N(x,\lambda),\lambda) &= \sum_{l=0}^{+\infty} \frac{1}{l!} x^l \frac{\partial^l \delta_N}{\partial x^l}(x,\lambda)\rho^l \\ &= \sum_{l=0}^{+\infty} \hat{a}_l(x,\lambda)\rho^l \end{aligned} \tag{8.8}$$

where we defined

$$\hat{a}_l(x,\lambda) = \frac{1}{l!} x^l \frac{\partial^l \delta_N}{\partial x^l}(x,\lambda). \tag{8.9}$$

The previous series converges for $|\rho| < 1$ and for any $(x,\lambda) \in I \times W_N$.

Using the expressions (8.7) and (8.8), the equation $\delta(x,\lambda) = \delta_N(\varphi_N(x,\lambda),\lambda)$ is equivalent to

$$\delta_N(x,\lambda) + x^N R_N(x,\lambda) = \sum_{l=0}^{+\infty} \hat{a}_l(x,\lambda)\rho^l \tag{8.10}$$

as $\hat{a}_0(x,\lambda) = \delta_N(x,\lambda)$ we obtain the following lemma

Lemma 2.1. *The functional equation $\delta(x,\lambda) = \delta_N(\Phi_N(x,\lambda),\lambda)$ is equivalent to the following functional equation which is analytic in ρ and is given by*

$$H_N(x,\lambda,\rho) = \sum_{l=1}^{+\infty} \hat{a}_l(x,\lambda)\rho^l - x^N R_N(x,\lambda) = 0 \tag{*}$$

To show that the functional equation (*) has a solution, we will study the "ideal" generated by the germs of the functions $\hat{a}_l$ in $(\mathbb{R}^+,0)\times(\mathbb{R}^{\Lambda+3},0)$ as it will be explained later on.

From now on we will work with an expansion of δ written with three compensators $\omega_1,\omega_2,\omega_3$. The general case is proved in the same way. The function

$$\begin{aligned}\hat{a}_0(x,\lambda) = \delta_N(x,\lambda) &= \hat{P}_0(x,Z_1,Z_2,Z_3,\lambda) \\ &= \sum_{0\le i+j_1+j_2+j_3\le N} \gamma_{ij_1j_2j_3}(\lambda) x^i Z_1^{j_1} Z_2^{j_2} Z_3^{j_3}\end{aligned}$$

where $\hat{P}_0$ is obviously a polynomial in the variables (x,z_1,z_2,z_3) of degree N with analytic coefficients in λ.

Recall that the parameter μ_i was introduced in (8.2) and define $r_i = 1-\mu_i$ then $\frac{\partial Z_i}{\partial x} = \frac{r_i Z_i - x}{x}$. As the function $\hat{a}_1(x,\lambda) = x\frac{\partial \delta_N}{\partial x}(x,\lambda)$ we obtain

$$\hat{a}_1(x,\lambda) = x\frac{\partial \hat{P}_0}{\partial x}(x,Z_1,Z_2,Z_3,\lambda) + \sum_{i=1}^{3}(r_i Z_i - x)\frac{\partial \hat{P}_0}{\partial z_i}(x,Z_1,Z_2,Z_3,\lambda) \tag{8.11}$$

It is easy to see that we can find a polynomial $\hat{P}_1$ in the variables (x, z_1, z_2, z_3) of degree N and analytic in λ such that $\hat{a}_1(x,\lambda) = \hat{P}_1(x, Z_1, Z_2, Z_3, \lambda)$. This fact can be generalized easily for any $n \in \mathbb{N}$

$$\hat{a}_n(x,\lambda) = \hat{P}_n(x, Z_1, Z_2, Z_3, \lambda) \tag{8.12}$$

Define the following sequence of functions $a_0(x,\lambda) = \delta_N(x,\lambda)$ and

$$a_n(x,\lambda) = \mathcal{L}^{(n)} a_0(x,\lambda), \tag{8.13}$$

for the derivation operator

$$\mathcal{L} = x\frac{\partial}{\partial x}.$$

It follows that the same conclusions for $\hat{a}_n$ are true for a_n, i.e. there exists a polynomial P_n in the variables (x, z_1, z_2, z_3) of degree N and analytic in λ such that

$$a_n(x,\lambda) = P_n(x, Z_1, Z_2, Z_3, \lambda). \tag{8.14}$$

Let $\mathcal{I}$ (respectively $\hat{\mathcal{I}}$ be the ideal generated by the germs of the polynomials $(P_n)_{n\geq 1}$ (respectively by $(\hat{P}_n)_{n\geq 1}$) in $(x, z_1, z_2, z_3, \lambda) = (0,0,0,0,0) \in \mathbb{R}^{\Lambda+4}$ in the ring $\mathcal{O}$ of the analytic germs. Let $\mathcal{I}_0$ (respectively $\hat{\mathcal{I}}_0$ be the set ("ideal" in this sense) obtained as a restriction of $\mathcal{I}$ (respectively $\hat{\mathcal{I}}$) on the graphs $z_i = Z_i(x, \mu_i)$.

Lemma 2.2. *The "ideals" $\mathcal{I}_0$ and $\hat{\mathcal{I}}_0$ satisfy $\mathcal{I}_0 = \hat{\mathcal{I}}_0$.*

Easy algebraic and derivation computations give the result above. Indeed for the first three functions we have

$$\begin{aligned} a_1(x,\lambda) &= \mathcal{L}a_0(x,\lambda) = x\frac{\partial a_0}{\partial x}(x,\lambda) = \hat{a}_1(x,\lambda) \\ a_2(x,\lambda) &= \mathcal{L}a_1(x,\lambda) = x\frac{\partial x a_0'}{\partial x}(x,\lambda) = \hat{a}_1(x,\lambda) + \hat{a}_2(x,\lambda) \end{aligned}$$

where f' stand for $\frac{\partial f}{\partial x}$. Define the following "zero sets"

$$\begin{aligned} \mathcal{Z}(\mathcal{I}_0) &= \{(x,\lambda);\ \forall n \geq 1\ a_n(x,\lambda) = 0\} \\ \mathcal{Z}(\mathcal{I}) &= \{(x, z_1, z_2, z_3, \lambda);\ \forall n \geq 1\ P_n(x, z_1, z_2, z_3, \lambda) = 0\} \end{aligned} \tag{8.15}$$

Theorem 2.3. *There exists an integer $m \in \mathbb{N}$ such that $x^{m(1+r_1+r_2+r_3)}$ is in $\mathcal{I}_0$.*

The next section will be devoted to show this assertion which is important to solve Eq. (*).

8.3 A Constructive Proof of Theorem 1.6

This proof is constructive because we can give the value of the integer m defined in Theorem 1.6 as a function of the nontriviality order $n = n_0 + n_1 + n_2 + n_3$ defined in $(\mathcal{H}3)$. The proof is very technical and could not be avoided.

8.3.1 Existence of a Function $b(\mu)x^{\nu(\mu)}$ in the "Ideal" $\mathcal{I}_0$

Recall that

$$\delta_N(x, 0) \equiv x^n (\ln x)^{n_1+n_2+n_3} \tag{8.16}$$

and we write

$$\delta_N(x, \lambda) = \sum_{i=0}^{N} p_i(x, Z_1, Z_2, Z_3, \lambda) \tag{8.17}$$

where the p_i, for $1 \leq i \leq N$, is a homogeneous polynomial in the variables (x, z_1, z_2, z_3) of degree i and analytic in λ.

Lemma 3.1. *The ideal $\mathcal{I}$ contains a function h which is homogeneous polynomial in the variables (x, z_1, z_2, z_3) and analytic in λ such that*

$$h(x, z_1, z_2, z_3, \lambda) = \sum_{i+j_1+j_2+j_3=n} \gamma_{ij_1j_2j_3}(\lambda) M_{ij_1j_2j_3} \tag{8.18}$$

where

$$M_{ij_1j_2j_3} = x^i z_1^{j_1} z_2^{j_2} z_3^{j_3} \tag{8.19}$$

and furthermore we have

$$\gamma_{n000}(\lambda) \equiv 1 \tag{8.20}$$

This means that the restriction of the function h on the graphs $z_i = Z_i(x, \mu_i)$ is in the "ideal" $\mathcal{I}_0$, restriction of the ideal $\mathcal{I}$ on the graphs $z_i = Z_i(x, \mu_i)$.

Proof. First of all one has the following property

$$\begin{aligned} \hat{\mathcal{L}} M_{ij_1j_2j_3} = {} & (i + j_1 r_1 + j_2 r_2 + j_3 r_3) M_{ij_1j_2j_3} \\ & - j_1 M_{i+1j_1-1j_2j_3} - j_2 M_{i+1j_1j_2} - j_3 M_{i+1j_1j_2j_3-1} \end{aligned} \tag{8.21}$$

where for g, a polynomial in the variables (x, z_1, z_2, z_3) and analytic in λ, we defined

$$\hat{\mathcal{L}}(g) = x \frac{\partial g}{\partial x} + \sum_{i=1}^{3} (r_i z_i - x) \frac{\partial g}{\partial z_i} \tag{8.22}$$

By construction the ideal I is stable under the operator $\hat{\mathcal{L}}$ and this operator satisfies

$$\mathcal{L}(g(x, Z_1, Z_2, Z_3, \lambda)) = \hat{\mathcal{L}}(g)(x, Z_1, Z_2, Z_3, \lambda) \tag{8.23}$$

So it is not ambiguous when we apply the operator $\mathcal{L}$ directly on $\mathcal{L}_0$. All the properties of the ideal $\mathcal{L}$ are inherited automatically by the "ideal" $\mathcal{L}_0$, i.e. stability with respect to addition, multiplication, etc.

The proof is based on "elimination algorithm". Indeed we will eliminate all the monomials $\psi_{ij_1j_2j_3}(x, \lambda) = M_{ij_1j_2j_3}(x, Z_1, Z_2, Z_3, \lambda)$ such that $i + j_1 + j_2 + j_3 \neq n$ in $\delta_N(x, \lambda)$.

Take such a monomial then one can write:

$$\delta_N(x, \lambda) = Q(x, Z_1, Z_2, Z_3, \lambda) + \gamma_{ij_1j_2j_3}\psi_{ij_1j_2j_3}(x, \lambda) \tag{8.24}$$

We apply the derivation operator $\mathcal{L}$ and the result of (8.21) to obtain

$$\begin{aligned} \mathcal{L}\delta_N(x, \lambda) = \mathcal{L}Q(x, Z_1, Z_2, Z_3, \lambda) + \gamma_{ij_1j_2j_3}[\\ (i + j_1r_1 + j_2r_2 + j_3r_3)\psi_{ij_1j_2j_3} \\ - j_1\psi_{i+1j_1-1j_2j_3} - j_2\psi_{i+1j_1j_2} - j_3\psi_{i+1j_1j_2j_3-1}]. \end{aligned} \tag{8.25}$$

Subtracting

$$(i + j_1r_1 + j_2r_2 + j_3r_3)\delta_N(x, \lambda) - \mathcal{L}\delta_N(x, \lambda) = Q_1(x, Z_1, Z_2, Z_3, \lambda)$$

where on the one hand Q_1 does not contain any more the monomial $\psi_{ij_1j_2j_3}$ and on the other hand the function $(i + j_1r_1 + j_2r_2 + j_3r_3)\delta_N(x, \lambda) - \mathcal{L}\delta_N(x, \lambda)$ is still in the ideal $\mathcal{I}_0$. This procedure permits to eliminate all the monomials $\psi_{ij_1j_2j_3}$ with $i + j_1 + j_2 + j_3 \neq n$. At each step we get functions in the "ideal" $\mathcal{I}_0$. At the end of this procedure, we obtain a function $h(x, Z_1, Z_2, Z_3, \lambda)$ such that for $\lambda = 0$ we have

$$h(x, -\ln x, -\ln x, -\ln x, 0) \not\equiv 0.$$

Using the properties of the $\mathcal{L}$, there exists $p \in \mathbb{N}$ such that $\mathcal{L}^p h(1, 0, 0, 0, 0) \neq 0$. Up to a division by a non-zero coefficient, one can replace h by $\mathcal{L}^p h$ to get a new function denoted again by h as described in (8.18) with new functions $\gamma_{ij_1j_2j_3}(\lambda)$ which we still denote as those of the beginning and satisfying (8.20). □

We cannot apply the same procedure above to the function h, otherwise we lose the control on our nontriviality order, i.e. it will be multiplied by a function which vanishes for $\lambda = 0$ (the function h and all of its monomials satisfy a differential equation which has for $\lambda = 0$ a unique eigenvalue $= n$). So we apply some techniques from the linear algebra and differential calculus.

Let N_0 be the cardinal of the set containing all the monomials $\psi_{ij_1j_2j_3}$ such that $i + j_1 + j_2 + j_3 = n$; the value of N_0 is

$$N_0 = \frac{(N + 3)}{3!n!}$$

and let $\mathcal{J}$ be the set $\{(i,j_1,j_2,j_3);\ i+j_1+j_2+j_3=n,(j_1,j_2,j_3)\neq 0\}$. We apply the derivation operator (N_0-1) times to the function h of (8.18). We obtain then

$$\begin{cases} h & = \sum_{\mathcal{J}} \gamma_{ij_1j_2j_3}\psi_{ij_1j_2j_3} + x^n \\ \mathcal{L}(h) & = \sum_{\mathcal{J}} \gamma_{ij_1j_2j_3}\mathcal{L}(\psi_{ij_1j_2j_3}) + \mathcal{L}(x^n) \\ \vdots & \quad \vdots \\ \mathcal{L}^{N_0-1}(h) & = \sum_{\mathcal{J}} \gamma_{ij_1j_2j_3}\mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3}) + \mathcal{L}^{N_0-1}(x^n) \end{cases} \tag{8.26}$$

This system can be transformed in the following one

$$\begin{cases} \sum_{\mathcal{J}} \gamma_{ij_1j_2j_3}\psi_{ij_1j_2j_3} + (x^n - h) & = 0 \\ \sum_{\mathcal{J}} \gamma_{ij_1j_2j_3}\mathcal{L}(\psi_{ij_1j_2j_3}) + (\mathcal{L}(x^n) - \mathcal{L}(h)) & = 0 \\ \vdots & \vdots \\ \sum_{\mathcal{J}} \gamma_{ij_1j_2j_3}\mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3}) + (\mathcal{L}^{N_0-1}(x^n) - \mathcal{L}^{N_0-1}(h)) & = 0 \end{cases} \tag{8.27}$$

The system (8.27) can be seen as a linear system of N_0 equations with coefficients $\mathcal{L}^l(\psi_{ij_1j_2j_3})$ and $(\mathcal{L}^l(x^n)-\mathcal{L}^l(h))$; $(i,j_1,j_2,j_3)\in\mathcal{J}$ and $0\leq l\leq N_0-1$. It has a non-trivial solution

$$((\gamma_{ij_1j_2j_3})_{\mathcal{J}},1)$$

where the monomials $\psi_{ij_1j_2j_3}$ are ordered by the lexicographical order denoted $\prec$. So the determinant of the system must be zero.

$$0=\det\begin{pmatrix} \cdots & \psi_{ij_1j_2j_3} & \cdots & x^n-h \\ \cdots & \mathcal{L}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}(x^n)-\mathcal{L}(h) \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}^{N_0-1}(x^n)-\mathcal{L}^{N_0-1}(h) \end{pmatrix} \tag{8.28}$$

with $(i,j_1,j_2,j_3)\in\mathcal{J}$. Using the multi-linearity of the determinant, we obtain

$$\begin{aligned} \Delta(x,\mu) &= \det\begin{pmatrix} \cdots & \psi_{ij_1j_2j_3} & \cdots & x^n \\ \cdots & \mathcal{L}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}(x^n) \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}^{N_0-1}(x^n) \end{pmatrix} \\ &= \det\begin{pmatrix} \cdots & \psi_{ij_1j_2j_3} & \cdots & h \\ \cdots & \mathcal{L}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}(h) \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}^{N_0-1}(h) \end{pmatrix} \end{aligned} \tag{8.29}$$

where $\mu = (\mu_1, \mu_2, \mu_3)$. The "ideal" $\mathcal{I}_0$ is stable under the action of $\mathcal{L}$ and h is in this "ideal" then the second determinant in expression (8.29) is also an element of the "ideal" $\mathcal{I}_0$. Consequently the determinant Δ is in $\mathcal{I}_0$. The next step is to compute this determinant.

Lemma 3.2. *There exist two analytic functions b and ν of μ such that*

$$\begin{aligned} \Delta(x,\mu) &= b(\mu)x^{\nu(\mu)} \\ b(\mu) &= \Delta(1,\mu) \\ \nu(\mu) &= \left(\sum_{i=0}^{(n-1)}(n-i)\frac{(i+1)(i+2)}{2}\right)(1+r_1+r_2+r_3) \end{aligned} \tag{8.30}$$

Proof. Apply the operator derivation $\mathcal{L}$ on Δ which is a determinant of functions. For any (i, j_1, j_2, j_3) with $i+j_1+j_2+j_3 = n$, call $C_{ij_1j_2j_3}$ the $(ij_1j_2j_3)$th column vector in Δ. It is under the form

$$C_{ij_1j_2j_3} = \begin{pmatrix} \psi_{ij_1j_2j_3} \\ \mathcal{L}(\psi_{ij_1j_2j_3}) \\ \vdots \\ \mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3}) \end{pmatrix} \tag{8.31}$$

Now we have $\Delta = \det(C_{ij_1j_2j_3})_{i+j_1+j_2+j_3=n}$ and it follows

$$\mathcal{L}(\Delta(x,\mu)) = \sum_{i+j_1+j_2+j_3=n} \Delta_{ij_1j_2j_3} \tag{8.32}$$

where $\Delta_{ij_1j_2j_3}$ stands for the determinant obtained from the matrix $C_{lk_1k_2k_3}$, $(l+k_1+k_2+k_3 = n)$, where we applied the derivation operator to and only to the elements of the $(ij_1j_2j_3)$th column vector. We set

$$\nu_{ij_1j_2j_3} = i + j_1r_1 + j_2r_2 + j_3r_3 \tag{8.33}$$

Using the result about the derivation operator $\mathcal{L}$ in expression (8.21) we obtain

$$\begin{aligned} \mathcal{L}(C_{ij_1j_2j_3}) &= \begin{pmatrix} \mathcal{L}(\psi_{ij_1j_2j_3}) \\ \mathcal{L}(\mathcal{L}(\psi_{ij_1j_2j_3})) \\ \vdots \\ \mathcal{L}^{N_0-1}(\mathcal{L}(\psi_{ij_1j_2j_3})) \end{pmatrix} \\ &= \nu_{ij_1j_2j_3}C_{ij_1j_2j_3} - j_1C_{i+1j_1-1j_2j_3} - j_2C_{i+1j_1j_2-1j_3} \\ &\quad - j_3C_{i+1j_1j_2j_3-1} \end{aligned} \tag{8.34}$$

Thanks to the multilinearity and the properties of the determinant, we find

$$\Delta_{ij_1j_2j_3}(x,\mu) = (i + j_1r_1 + j_2r_2 + j_3r_3)\Delta(x,\mu) \tag{8.35}$$

this gives

$$\begin{aligned} \mathcal{L}(\Delta) &= \sum_{i+j_1+j_2+j_3=n} (i + j_1r_1 + j_2r_2 + j_3r_3)\Delta \\ &= \nu(\mu)\Delta \\ \nu(\mu) &= \left(\sum_{i=0}^{(n-1)} (n-i)\frac{(i+1)(i+2)}{2} \right)(1 + r_1 + r_2 + r_3) \end{aligned} \tag{8.36}$$

The definition of $\mathcal{L}$ shows that Δ satisfies the differential equation

$$x\frac{\partial \Delta}{\partial x}(x,\mu) = \nu(\mu)\Delta(x,\mu).$$

One solves it and we obtain the result. □

8.3.2 *Proof of Theorem 1.6*

We cannot yet assert that $x^{\nu(\mu)}$ is in the "ideal" $\mathcal{I}_0$ because we know only that $b(\mu)x^{\nu(\mu)}$ is in $\mathcal{I}_0$ and we have no information about the germ of $b(\mu)$ in $\mu = 0$. So we need some extra preparatory work to show the assertion of Theorem (1.6).

Lemma 3.3. *The monomials $\psi_{ij_1j_2j_3} = x^i Z_1^{j_1} Z_2^{j_2} Z_3^{j_3}$, $i + j_1 + j_2 + j_3 = n$, satisfy the following differential equation*

$$\mathcal{L}^{N_0}(\psi) = \sum_{i=1}^{(N_0-1)} c_i(\mu)\mathcal{L}^i(\psi) \tag{8.37}$$

where c_i are analytic functions of μ.

Proof. We show the lemma in four steps. The first three steps will be devoted to prove the assertion for $x^{\nu_{ij_1j_2j_3}}$, $\nu_{ij_1j_2j_3} = i + j_1r_1 + j_2r_2 + j_3r_3$. In the fourth step we show the assertion for any monomial $\psi_{ij_1j_2j_3}$ using the fact that it can be written as a linear combination of $x^{\nu_{ij_1j_2j_3}}$.

1. Let ψ be any $x^{\nu_{ij_1j_2j_3}}$ for $i + j_1 + j_2 + j_3 = n$. Let us define a $(N_0 + 1) \times (N_0 + 1)$ matrix $(C, \tilde{C}_{ij_1j_2j_3})_{(i+j_1+j_2+j_3=n)}$ defined as follows:

$$\tilde{C}_{ij_1j_2j_3} = \begin{pmatrix} x^{\nu_{ij_1j_2j_3}} \\ \mathcal{L}(x^{\nu_{ij_1j_2j_3}}) \\ \vdots \\ \mathcal{L}^{N_0-1}(x^{\nu_{ij_1j_2j_3}}) \\ \mathcal{L}^{N_0}(x^{\nu_{ij_1j_2j_3}}) \end{pmatrix} \text{ and } C = \begin{pmatrix} \psi \\ \mathcal{L}(\psi) \\ \vdots \\ \mathcal{L}^{N_0-1}(\psi) \\ \mathcal{L}^{N_0}(\psi) \end{pmatrix} \tag{8.38}$$

The properties of the determinant yield

$$E(\psi) = \det(C, \tilde{C}_{ij_1j_2j_3})_{(i+j_1+j_2+j_3=n)} \equiv 0$$

So we can write

$$E(\psi) = \det \begin{pmatrix} \psi & \cdots & x^{\nu_{ij_1j_2j_3}} & \cdots & x^n \\ \mathcal{L}(\psi) & \cdots & \mathcal{L}(x^{\nu_{ij_1j_2j_3}}) & \cdots & \mathcal{L}(x^n) \\ \vdots & \vdots & \vdots & \vdots & \\ \mathcal{L}^{N_0-1}(\psi) & \cdots & \mathcal{L}^{N_0-1}(x^{\nu_{ij_1j_2j_3}}) & \cdots & \mathcal{L}^{N_0-1}(x^n) \\ \mathcal{L}^{N_0}(\psi) & \cdots & \mathcal{L}^{N_0}(x^{\nu_{ij_1j_2j_3}}) & \cdots & \mathcal{L}^{N_0}(x^n) \end{pmatrix} \tag{8.39}$$

Develop this determinant with respect to the first column. Then we obtain the following differential equation satisfied by ψ which is written

$$\Delta_{N_0}\mathcal{L}^{N_0}(\psi) + \cdots + \Delta_i\mathcal{L}^i(\psi) + \cdots + \Delta_0\mathcal{L}^0(\psi) = 0 \tag{E}$$

where Δ_i stands for the determinant obtained from the matrix $(\tilde{C}_{ij_1j_2j_3})$ where $(i + j_1 + j_2 + j_3 = n)$, without the ith row. The 0th row is the first row of the matrix.

2. Let us calculate the determinant Δ_{N_0}. An easy computation shows that for any integer p

$$\mathcal{L}^p(x^{\nu_{ij_1j_2j_3}}) = \nu^p_{ij_1j_2j_3}x^{\nu_{ij_1j_2j_3}} \tag{8.40}$$

then the determinant Δ_{N_0} becomes

$$\Delta_{N_0} = \prod_{i+j_1+j_2+j_3=n} x^{\nu_{ij_1j_2j_3}} \det \begin{pmatrix} \cdots & 1 & \cdots & 1 \\ \cdots & \nu_{ij_1j_2j_3} & \cdots & n \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \nu^{N_0-1}_{ij_1j_2j_3} & \cdots & n^{N_0-1} \end{pmatrix} \tag{8.41}$$

Due to the fact that the determinant in expression (8.41) is a Vandermond determinant

$$\Delta_{N_0}(x, \lambda) = b_{N_0}(\mu)x^{\nu_\mu} \tag{8.42}$$

where

$$b_{N_0}(\mu) = \prod_{\substack{i+j_1+j_2+j_3 = k+l_1+l_2+l_3 = n \\ (i,j_1,j_2,j_3) \prec (k,l_1,l_2,l_3)}} (\nu_{ij_1j_2j_3} - \nu_{kl_1l_2l_3})$$

3. for $i \neq 0$, we can apply the same computations to obtain the determinant Δ_i. We find

$$\Delta_i(x,\lambda) = b_i(\mu)x^{\nu_\mu} \tag{8.43}$$

where

$$b_i(\mu) = \det \begin{pmatrix} \cdots & 1 & \cdots & 1 \\ \cdots & \nu_{ij_1j_2j_3} & \cdots & n \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \nu_{ij_1j_2j_3}^{i-1} & \cdots & n^{i-1} \\ \cdots & \nu_{ij_1j_2j_3}^{i+1} & \cdots & n^{i+1} \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \nu_{ij_1j_2j_3}^{N_0-1} & \cdots & n^{N_0-1} \\ \cdots & \nu_{ij_1j_2j_3}^{N_0} & \cdots & n^{N_0} \end{pmatrix}$$

If we let two columns be the same which means to take $\nu_{ij_1j_2j_3} = \nu_{kl_1l_2l_3}$, then $b_i(\mu)$ will be zero. This fact implies that $\nu_{ij_1j_2j_3} - \nu_{kl_1l_2l_3}$ divides this determinant in the ring of polynomials $\mathbb{R}[\nu_{ij_1j_2j_3}]$ for all (i,j_1,j_2,j_3) satisfying $i+j_1+j_2+j_3 = n$. Consequently there exist polynomials c_i in the ring $\mathbb{R}[\mu]$ such that

$$b_i(\mu) = c_i(\mu)b_{N_0}(\mu). \tag{8.44}$$

Replace each Δ_i by its value in equation (E). Then this equation can be desingularized

$$\mathcal{L}^{N_0}(\psi) + \cdots + c_i\mathcal{L}^i(\psi) + \cdots + c_0\mathcal{L}^0(\psi) = 0 \tag{8.45}$$

4. Easy computations show that the monomials $\psi_{ij_1j_2j_3}$ can be written as a linear combination of the functions $x^{\nu_{ij_1j_2j_3}}$. Indeed

$$\mu_i Z_i = x(x^{-\mu_i} - 1) = x^{r_i} - x \tag{8.46}$$

which yields

$$\mu_1^{j_1}\mu_2^{j_2}\mu_3^{j_3}\psi_{ij_1j_2j_3} = x^i(x^{r_1} - x)^{j_1}(x^{r_2} - x)^{j_2}(x^{r_3} - x)^{j_3} \tag{8.47}$$

So for any $\mu \notin F = \{\mu;\ \prod_{i=1}^{3} \mu_i = 0\}$, the monomials satisfy the differential equation (8.45). For $\mu_0 \in F$ then Z_i and $\mathcal{L}^p(Z_i)$ converge to $-x\ln x$ and $\mathcal{L}^p(-x\ln x)$ when $\mu \to \mu_0$ and $\mu \notin F$ uniformly for $x \in \overline{I}$. Then for such μ_0 the monomials $\psi_{ij_1j_2j_3}$ satisfy also the differential equation.

□

Recall that from expression (8.29) we have got

$$\Delta(x,\mu) = \det \begin{pmatrix} \cdots & \psi_{ij_1j_2j_3} & \cdots & h \\ \cdots & \mathcal{L}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}(h) \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3}) & \cdots & \mathcal{L}^{N_0-1}(h) \end{pmatrix} \tag{8.48}$$

We compute the determinant of the matrix by developing with respect to the last right column containing the coefficients $\mathcal{L}^i(h)$. Then we obtain

$$\begin{aligned} \Delta(x,\mu) &= b(\mu)x^{\nu(\mu)} \\ &= \sum_{i=0}^{N_0-1} D_i(x,\mu)\mathcal{L}^i(h)(x,\lambda) \end{aligned} \tag{8.49}$$

where the D_i are the determinants of $(N_0-1)\times(N_0-1)$ matrices obtained from

$$(C_{ij_1j_2j_3})_{(i+j_1+j_2+j_3=n)}$$

by eliminating the vector column C_{n000} and the ith row.

The last step to prove the Theorem (1.6) consists in showing that $D_i(.,\mu) = d_i(\mu)b(\mu)$ where d_i will be described later. Besides, the Lemma (8.18) the key idea is to use the coefficient ideal as in [15]. For instance, take a function $\delta(.,\lambda)$ analytic in (x,λ). For $x_0 \in (0,\epsilon)$ the function δ is analytic at the variables (x,λ). Take its power expansion series at $(x-x_0)$

$$\delta(x,\lambda) = \sum_{i=0}^{\infty} g_i(\lambda, x_0)(x-x_0)^i$$

for x close to x_0. Consider the ideal J_{x_0} generated by the germs of functions g_i at $\lambda = 0$. The ideal J_{x_0} does not depend on $x_0 \neq 0$. This ideal will be the coefficient ideal associated with δ in the ring of analytic germs at $\lambda = 0$.

Let us associate with D_i, for a fixed $i:\ 0 \le i \le N_0-1$, its coefficient ideal. Then there exist some generators of this ideal and quasi-regular functions such that

$$D_i(x,\mu) = \sum_{\text{finite}} \phi_{ji}(\mu)g_{ij}(x,\mu).$$

The coefficient ideal does not depend on x then we can calculate its generators for $x = 1$. Take u close to 0 then we can write

$$D_i(1+u,\mu) = \sum_{l=0}^{\infty} \frac{1}{l!} u^l \frac{\partial^l D_i}{\partial x^l}(1,\mu) \tag{8.50}$$

As a consequence of Lemma (2.2), it is sufficient to compute $\mathcal{L}^l D_i(1,\mu)$ because $\frac{\partial^l D_i}{\partial x^l}(1,\mu)$ depends linearly, with coefficients in $\mathbb{R}$, on $\mathcal{L}^j D_i(1,\mu)$ for $1 \leq j \leq l$. Let us proceed by steps again.

– It is easy to see that

$$\psi_{ij_1j_2j_3}(1,\mu) = 0 \ \forall (j_1,j_2,j_3) \ \neq 0 \tag{8.51}$$

– For any integer k define the following row vectors:

$$L_k = (\cdots, \mathcal{L}^k(\psi_{ij_1j_2j_3}), \cdots) \ (i,j_1,j_2,j_3) \in \mathcal{J} \tag{8.52}$$

then

$$D_j = \det(L_0, L_1, \cdots, L_{j-1}, L_{j+1}, \cdots, L_{N_0-1}).$$

– For any k and j integers such that $0 \leq j \leq N_0 - 1$,

$$\mathcal{L}^k(D_j) = \sum_{i=0}^{N_0-1} f_{jki}(\mu) D_i \tag{8.53}$$

Indeed let us make a recurrence on k. For $k = 1$ we have

$$\begin{aligned} \mathcal{L}(D_j) &= \det(L_1, L_1, \cdots, L_{j-1}, L_{j+1}, \cdots, L_{N_0-1}) \\ &+ \det(L_0, L_2, \cdots, L_{j-1}, L_{j+1}, \cdots, L_{N_0-1}) \\ &+ \cdots + \det(L_0, L_1, \cdots, L_{j-1}, L_{j+1}, \cdots, L_{N_0-2}, L_{N_0}) \end{aligned} \tag{8.54}$$

Thanks to the Lemma (3.3) we know that

$$\begin{aligned} L_{N_0} &= (\cdots, \mathcal{L}^{N_0}(\psi_{ij_1j_2j_3}), \cdots) \\ &= (\cdots, \sum_{l=1}^{N_0-1} c_l(\mu) \mathcal{L}^l(\psi_{ij_1j_2j_3}), \cdots) \\ &= \sum_{l=1}^{N_0-1} c_l(\mu) L_l \end{aligned} \tag{8.55}$$

The previous expression and the multilinearity of the determinant give

$$\mathcal{L}^1(D_j) = \sum_{i=0}^{N_0-1} f_{j1i}(\mu) D_i$$

Suppose that the recurrence is true for any integer $\leq k$. Then we obtain

$$\begin{aligned}\mathcal{L}^{k+1}(D_j) &= \sum_{i=0}^{N_0-1} f_{jki}(\mu)\mathcal{L}(D_i) \\ &= \sum_{i=0}^{N_0-1} f_{jki}(\mu) \sum_{l=0}^{N_0-1} f_{i1l}(\mu)(D_l)\end{aligned}$$

An easy computation gives that

$$\mathcal{L}^{k+1}(D_j) = \sum_{i=0}^{N_0-1} f_{jk+1i}(\mu) D_i$$

– An easy remark is the following:

$$D_i(1,\mu) = 0 \ \forall\, i\, 1 \leq i \leq N_0 - 1 \tag{8.56}$$

Indeed

$$L_0 = (\cdots, \psi_{ij_1j_2j_3}, \cdots) \quad i + j_1 + j_2 + j_3 = n;\ (j_1, j_2, j_3) \neq 0$$

as it was noticed before $\psi_{ij_1j_2j_3}(1,\mu) = 0$ when $(j_1, j_2, j_3) \neq 0$ so $D_i(1,\mu) = \det(L_0, \cdots) = 0$.

– It remains to compute D_0. Take the determinant $\Delta(1,\mu)$ which has the following expression

$$\Delta(1,\mu) = \det \begin{pmatrix} \cdots & 0 & \cdots & 1 \\ \cdots & \mathcal{L}(\psi_{ij_1j_2j_3})(1,0) & \cdots & n \\ \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathcal{L}^{N_0-1}(\psi_{ij_1j_2j_3})(1,0) & \cdots & n^{N_0-1} \end{pmatrix}$$

Develop $\Delta(1,\mu)$ with respect to the first row where all the coefficients are zero except the last which is 1. It follows that

$$\Delta(1,\mu) = D_0(1,\mu) = b(\mu) \tag{8.57}$$

As $D_j = \det(L_0, L_1, \cdots, L_{j-1}, L_{j+1}, \cdots, L_{N_0-1})$ and using (8.53) we can infer that $b(\mu)$ divides all the $\mathcal{L}^k(D_j)(1,\mu)$ which is also true for $\frac{\partial^n D_i}{\partial x^n}(1,\mu)$. Then

we apply this result to the generators of the coefficients ideal associated with each D_i to obtain

$$D_i(x,\mu) = b(\mu)d_i(x,\mu) \tag{8.58}$$

where $d_i(x,\mu)$ is a quasi-analytic function which is bounded and converges to 0 when x tends to 0 uniformly in the parameter.

– The last step to show the Theorem (1.6) goes as the following. In expression (8.49) we can simplify by $b(\mu)$ and we get

$$x^{\nu(\mu)} = \sum_{i=0}^{N_0-1} d_i(x,\mu)\mathcal{L}^i(h)(x,\lambda)$$

Consequently

$$x^{\nu(\mu)} \in I_0 \quad \square$$

8.4 Construction of the "Pseudo-Isomorphism" Φ

8.4.1 The Existence of a Map ρ Solving (*)

In this section we will solve the functional equation

$$H_N(x,\lambda,\rho) = \sum_{l=1}^{+\infty} \hat{a}_l(x,\lambda)\rho^l - x^N R_N(x,\lambda) = 0 \tag{*}$$

We have shown in the two last sections that there exists an integer N_1 such that

$$x^{N_1(1+r_1+r_2+r_3)} \in \mathcal{I}_0 \tag{8.59}$$

As said along this article the "ideal" $\mathcal{I}_0$ is defined as restriction of the ideal $\mathcal{I}$ on the graphs $z_i = Z_i(x,\mu_i)$. Let $\{P_1,\cdots,P_L\}$ be a system of analytic generators of $\mathcal{I}$ such that their restrictions on the graphs $z_i = Z_i(x,\mu_i)$ are $\hat{a}_j$ defined in (8.9). Then there exist some functions $\hat{h}_i$ which are quasi-regular in x and analytic in λ satisfying

$$x^{N_1(1+r_1+r_2+r_3)} = \sum_{i=1}^{L} \hat{h}_i(x,\lambda)\hat{a}_i(x,\lambda) \tag{**}$$

Remark. A first consequence of the equality (**) is that the multiplicity of an isolated zero of δ_λ (or of $\delta_{N,\lambda}$) which is different from 0 is bounded by L uniformly in $\lambda \in W_N$. In the case of analytic deformations of regular germs in 0, this property

implies the local uniform finiteness of the number of the isolated zeros. We will show that it is the case also here.

The equality (**) implies that, for $N > 5N_1$, the function $x^N R_N(x,\lambda)$ can be written

$$-x^N R_N(x,\lambda) = \sum_{i=1}^{L} g_i(x,\lambda)\hat{a}_i(x,\lambda) \tag{8.60}$$

where the functions g_i satisfy the property I_0^∞ and precisely

$$g_i = R_N O(x^{N-N_1(1+r_1+r_2+r_3)}) \tag{8.61}$$

This property follows after multiplying (**) by $x^{N-N_1(1+r_1+r_2+r_3)}R_N$. Let us consider, for $j = 1,\cdots,l$ the sets

$$V_j = \{(x,\lambda) \in I \times W_N;\ |\hat{a}_j(x,\lambda)| > \frac{1}{2}|\hat{a}_l(x,\lambda)| \text{ for } l = 1,\cdots,L\} \tag{8.62}$$

It is easy to see that $\cup_{j=1}^{L} V_j = I \times W_N$. Let us make the following remark: if $\hat{a}_i(x,\lambda) = 0$ for $l = 1,\cdots,L$, then according to (**), we get that $x = 0$. Let us set

$$\rho(0,\lambda) \equiv 0 \tag{8.63}$$

Let us define $\rho(x,\lambda)$ for $(x,\lambda) \in I \times W_N$ where $I =]0,\epsilon]$. Take (x,λ) in V_L for instance, then the functions $e_l = \frac{\hat{a}_l}{\hat{a}_L}$ are analytic and bounded: $|e_l| < 2$. Let (x,λ) be fixed then using the same arguments as in [15] and a theorem of Hervé applied to the extensions of functions $\hat{a}_l$ in the complex plane [15], we can write the series $\sum_{l=1}^{\infty} \hat{a}_l\rho^l$ as follows:

$$\sum_{l=1}^{\infty} \hat{a}_l\rho^l = \sum_{l=1}^{L} \hat{a}_l\rho^l + \rho^{L+1}\sum_{l=1}^{L} e_l f_l \tag{8.64}$$

where the functions f_l are analytic functions in $(x,\lambda)) \in I \times W_N$ and holomorphic in ρ for $|\rho| \le \frac{1}{2}$ and there exist constants c_l such that

$$\|f_l\|_{I\times W_N \times B(O,\frac{1}{2})} \le c_l.$$

Let $e = (e_1,\cdots,e_{L-1})$ then for any (x,λ) in V_L, Eq. (*) is equivalent to the following one:

$$\tilde{H}(x,\lambda,e,\rho) = \sum_{l=1}^{L} e_l g_l + \sum_{l=1}^{L} e_l\rho^l + \rho^{L+1}\sum_{l=1}^{L} e_l f_l = 0 \tag{***}$$

with $e_L = 1$ and e is a supplementary parameter which varies in the compact $[-2,2]^{L-1}$. We will solve Eq. (***) in the neighbourhood of each point in the

compact $[-2,2]^{L-1}$ and after we conclude by a compacity argument. Let us begin by the point $e = (0, \cdots, 0)$ to show how the resolution is done. Denote

$$G(x, \lambda, e, \rho) = \rho^L + \rho^{L+1} \sum_{l=1}^{L} e_l f_l = 0 \tag{8.65}$$

then for each $|\rho| \leq \eta$, where $\eta > 0$ and close to zero, we have

$$\frac{1}{2}|\rho|^L < |G| < \frac{3}{2}|\rho|^L \tag{8.66}$$

Up to a reduction of the interval I (take a smaller ϵ) and for small $\|e\|$, we have

$$|\tilde{H} - G|_{D(O,\eta)} < |G|_{D(O,\eta_0)} \tag{8.67}$$

where $D(O, \eta)$ denotes a polydisk. The expression (8.67) implies two things:

1. Using the theorem of Rouché, Eq. (***) has exactly L roots (ρ_i) and at least one on the ball $D(O, \eta)$.
2. Using the formula of Jensen, there exists $c > 0$ such that

$$\prod_{i=1}^{L} |\rho_i| < c \left| \sum_{l=1}^{L} e_l g_l \right| \tag{8.68}$$

Take for the value of $\rho(x, \lambda)$ the root of Eq. (***) which has the smaller module, a positive imaginary part (this exists as Eq. (***) has real coefficients) and the smaller real part. Then the formula (8.68) implies that $\rho = O(x)$.

For any other point $e \in [-2,2]^{L-1}$, we consider the integer l_0 such that $e_{l_0} \neq 0$. On V_L we repeat the same procedure to define the function $\rho(x, \lambda)$. The properties announced above for the roots ρ_i for $(x, \lambda) \in I \times W_N$ assure that $\lim_{x \to 0} \rho(x, \lambda) = 0$ uniformly in λ. Now it is clear that the procedure works on any V_l.

8.4.2 *The Uniform Finiteness of the Degree of the Map φ*

In this section we will show the second part of our main theorem, namely the existence of a uniform bound for the degree of $\varphi = x(1 + \rho)$.

Lemma 4.1. *Denote* $\varphi_\lambda = \varphi(., \lambda)$ *then for any* $w \in$

$$S_\epsilon = \{w = u + iv \in \mathbb{C};\ 0 < |v| < u < \epsilon\}$$

and for any λ *in* W_N

$$Card\{\phi_\lambda^{-1}(w)\} \leq L \tag{8.69}$$

Proof. Let x_1 and x_2 be in I such that $\varphi_\lambda(x) = \varphi_\lambda(x_1)$ then we write

$$x_1 = x(1 + \eta(x, \lambda)) \tag{8.70}$$

with $\eta(x, \lambda) = O(|x|)$ and η is a real function. Furthermore according to the definition of ρ, we obtain that $\delta(x, \lambda) = \delta(x_1, \lambda)$. Then the same arguments as for ρ in the previous sections yield that η satisfies a differential equation

$$G(x, \lambda, \eta) = \sum_{l=1}^{\infty} \frac{1}{l!} x^l \frac{\partial^l \delta}{\partial x^l}(x, \lambda)) = 0 \tag{8.71}$$

which we will solve using the same arguments as those used to solve (*). Let us denote $A_l(x, \lambda) = \frac{1}{l!} x^l \frac{\partial^l \delta}{\partial x^l}(x, \lambda))$ then we have for any l integer

$$A_l = \hat{a}_l + x^N R_{Nl} \tag{8.72}$$

where R_{Nl} satisfies the property I_0^∞. Recall that there exists $c_1 > 0$ such that for any j the function $\hat{a}_j$ is written

$$\hat{a}_j(x, \lambda) = \sum_{l=1}^{L} f_{jl}(x, \lambda) \hat{a}_l(x, \lambda) \tag{8.73}$$

and

$$\|f_{jl}\| \leq c_1 \|\hat{a}_{jl}\|.$$

Furthermore if we put $g(x, \lambda) = x^N R_N(x, \lambda)$ we get

$$g(x(1 + \eta), \lambda) = \sum_{l=0}^{\infty} t_l(x, \lambda)) x^l \eta^l \tag{8.74}$$

where the series converges for $|\eta| < 1$. Let $\eta_0 \in]0, 1]$, the theory of holomorphic functions with one variable gives the following estimate

$$|t_l(x, \lambda))| \leq \frac{c_0}{x^l \eta_0^l} \sup_{|\eta = \eta_0|} |g(x(1 + \eta), \lambda)|. \tag{8.75}$$

Now $\sup_{|\eta=\eta_0|} |g(x(1 + \eta), \lambda)| \leq c_1 x^N$ where c_0 and c_1 are independent of l and η_0. Let η_0 tend to 1 then we obtain

$$|x^l t_l(x, \lambda))| \leq c_2 x^N \tag{8.76}$$

so we can write

$$g(x(1+\eta),\lambda) = \sum_{l=0}^{\infty} x^N R_{Nl}(x,\lambda))\eta^l \tag{8.77}$$

with

$$\|R_{Nl}\| \leq c_2.$$

The same arguments applied to the function δ_N show that there exists $c_3 > 0$ such that for any j, $\|\hat{a}_j\| \leq c_3$. Furthermore the relation (**) shows that for each j we can write

$$x^N f_{jN} = x^{n_1} \sum_{l=1}^{L} g_{jl}\hat{a}_l \tag{8.78}$$

where $\|g_{jl}\| \leq c_2$ and $n_1 > 0$. Let A be the $L \times L$ matrix such that

$$\begin{pmatrix} A_1 \\ \vdots \\ A_L \end{pmatrix} = \begin{pmatrix} \hat{a}_1 \\ \vdots \\ \hat{a}_L \end{pmatrix} \tag{8.79}$$

We have that $\det A(0,\lambda) = 1$ while $A(0,\lambda) \equiv Id$, so for ϵ sufficiently small, we get $\|B_{lj}\| \leq c_4$ where $(B_{lj})_{1\leq j,l\leq L}$ is the inverse matrix of A. So for any l we can write

$$A_l = \sum_{j=1}^{L} \hat{g}_{jl} Aj$$

with $\|\hat{g}_{jl}\| \leq c_5$. This shows that we can sum up the series as the following:

$$G(x,\lambda,\eta) = \sum_{l=1}^{L} A_l\eta^l + \eta^{L+1} \sum_{j=1}^{L} A_l h_l \tag{8.80}$$

where the functions are analytic in η on $B(O,1)$ and bounded on $I \times W_N \times B(O,1)$. The previous expression looks like (***) and we can solve $G(x,\lambda,\eta) = 0$ in the same fashion. This ensures our assertion for small η, i.e. the equation $G(x,\lambda,\eta) = 0$ has at most L nonzero roots $\eta(x)$ for fixed λ. As a consequence we get the local and uniform finiteness of Card$\{\delta_\lambda = 0\}$. □

8.4.3 *Proof of the Main Corollary*

Let us show now the result of uniform finiteness of the number of the zeros for the equation $\delta_N(w,\lambda) = 0$ where w is an element of the following sector

$$S_{\epsilon_1\epsilon_2} = \{w = u + iv \in \mathbb{C};\ 0 < u < \epsilon_1,\ 0 < |v| < \epsilon_2 u\}$$

and λ is in W_N. Put $x_1 = u$ and $x_2 = \frac{v}{u}$, then for $w \in S_{\epsilon_1\epsilon_2}$ we have that $0 < x_1 < \epsilon_1$, $|x_2| < \epsilon_2$ and $w = x_1(1 + ix_2)$ where $i^2 = -1$. Let us evaluate the functions $Z_l(w, \lambda)$ for $l = 1, 2, 3$. Recall that

$$\omega_l(w, \lambda) = \frac{1}{\mu_l}[x_1^{-\mu_l}(1 + ix_2)^{-\mu_l} - 1] \tag{8.81}$$

$$X_{2l} = \frac{(1 + ix_2)^{-\mu_l} - 1}{\mu_l} \tag{8.82}$$

$$(1 + ix_2)^{-\mu_l} = 1 + \mu_l X_{2l} \tag{8.83}$$

and X_{2l} is holomorphic in x_2 for $|x_2| < \epsilon_2 < 1$. Denote $\omega_{1l} = \frac{x_1^{-\mu_l} - 1}{\mu_l}$ then one gets

$$\omega_l(w, \lambda) = \frac{1}{\mu_l}[(1 + \mu_l\omega_{1l})(1 + \mu_l X_{2l}) - 1] = \omega_{1l} + X_{2l} + \omega_{1l}X_{2l} \tag{8.84}$$

If we set $\hat{Z}_{1l} = x_1\omega_{1l}$, then

$$Z_l(\omega, x) = \hat{Z}_{1l}(1 + \mu_l X_{2l})(1 + ix_2) + x_1(1 + ix_2)X_{2l}. \tag{8.85}$$

The graph V_l for $l = 1, 2, 3$ of $\hat{Z}_{ll}(x_1, \lambda)$ is tangent to the kernel of the real 1-form

$$\hat{\Omega}_l = xdz_l - (r_l z_l - x_1)dx_1$$

furthermore X_{2l} satisfies

$$(1 + ix_2)\frac{\partial X_{2l}}{\partial x_2} = -i(\mu_l X_{2l} + 1)$$

We have seen that $\delta_N(\omega, \lambda) = P_0(\omega, Z_1(\omega, \lambda), Z_2(\omega, \lambda), Z_3(\omega, \lambda), \lambda)$ where P_0 is a polynomial in (ω, z_1, z_2, z_3) of degree n and analytic in λ. Then

$$\delta_N(\omega, \lambda) = Q(x_1, x_2, \hat{Z}_1(x_1, \lambda), \hat{Z}_2(x_1, \lambda), Z_3(x_1, \lambda), \lambda) \tag{8.86}$$

where Q is polynomial in (x_1, z_1, z_2, z_3) and analytic in (x_2, λ). So we obtain that

$$\text{Card}\{\omega \in S_{\epsilon_1\epsilon_2};\ \delta_N(x, \lambda) = 0\} = \text{Card}\{Q(x_1, x_2, z_1, z_2, z_3, \lambda) = 0\} \cap \cap_{l=1}^{3} V_l \tag{8.87}$$

Applying the theorem 2 of [13]

Theorem. *Let X be a semi-analytic set of $\mathbb{R}^n$ and (V_k, Ω_k, M) for $k = 1, 2, \cdots, q$ be separating hypersurfaces. If M is relatively compact, then the number of the connected components of $X \cap V_1 \cap V_2 \cdots \cap V_q$ is finite.*

shows that there exists an integer N_1 such that

$$\text{Card}\{Q(x_1, x_2, z_1, z_2, z_3, \lambda) = 0\} \cap \cap_{l=1}^{3} V_l \leq N_1$$

for any $\lambda \in W_N$. This proves the uniform finiteness for $\{\delta_\lambda = 0\}$.

8.5 Transition Maps Near a Hyperbolic Saddle Point

We call hyperbolic polycycles any graphic whose singular points are saddle points. We want to prove that some hyperbolic polycycles have finite cyclicity. To establish these results, we will use some basic properties of the transition map near a hyperbolic saddle point and we will exhibit the pfaffian differential equation verified by a such map.

8.5.1 Transition Map Near a Hyperbolic Saddle

Let X_λ be a C^∞ Λ-parameter family of vector fields defined in a neighbourhood of hyperbolic saddle. Since we are interested in the germ of the family at the point (P_0, λ_0), without loss of generality, we can suppose that the vector field X_λ is defined in a neighbourhood V_0 of $P_0 = (0, 0) \in \mathbb{R}^2$, for parameter values λ in a neighbourhood W of $0 \in \mathbb{R}^\Lambda$. We can suppose also that the coordinate axes are the invariant manifolds near the saddle point P_0. Finally we suppose that P_0 is the only singular point of X_λ.

Using the theory of normal forms, we can write some explicit expressions of the vector field X_λ up to C^k-equivalence, for $k \in \mathbb{N}$. These results have been known for vector fields, without parameters, since Poincaré. We will refer more to the work of Ilyashenko and Yakovenko [7]. Let $r_0 = r(0)$ be the ratio of hyperbolicity of X_0 in P_0.

- If $r(0)$ is irrational, then for any k a fixed integer, the vector field X_λ, for λ in some neighbourhood W_k, is C^k-equivalent to

$$\begin{cases} \dot{x} = x \\ \dot{y} = r(\lambda) y \end{cases} \tag{8.88}$$

  where $r(\lambda)$ denotes the ratio of hyperbolicity of X_λ in P_0.
- If $r(0) = \frac{p}{q}$ is rational, then for any k a fixed integer, the vector field X_λ, for λ in some neighbourhood W_k, is C^k-equivalent to

$$\begin{cases} \dot{x} = x \\ \dot{y} = y(-r_0 + \sum_{i=0}^{N(k)} \alpha_{i+1}(\lambda)(x^p y^q)^i). \end{cases} \tag{8.89}$$

where $N(k)$ is an integer depending on k.

In the following, we suppose that the vector field X_λ is given by one of the expressions (8.88) or (8.89). Let σ and τ be segments transverse to the vector field X_λ. They are defined by

$$\sigma = \{(x, y);\ |x| \leq 2, y = 1\} \text{ and } \tau = \{(x, y);\ |y| \leq 2, x = 1\}$$

The flow of X_λ induces a transition map $D_\lambda = D(., \lambda)$, also called the Dulac map:

$$D(., \lambda) :]0, x_0] \to]0, y_0] \quad \lambda \in W_k.$$

which can be extended continuously by $D(0, \lambda) \equiv 0$ for all $\lambda \in W_k$.

The Dulac map D_λ is C^∞ for $x \neq 0$. To know much more about its behaviour near $x = 0$, we have the following theorem due to Mourtada [11]:

Theorem 5.1. *The Dulac map D_λ associated with systems (8.88) and (8.89) can be written as follows:*

$$D_\lambda(x) = D(x, \lambda) = x^{r(\lambda)}[1 + \phi(x, \lambda)]\ \forall (x, \lambda) \in [0, x_0[\times W, \tag{8.90}$$

where ϕ is C^∞ for $(x, \lambda) \in [0, x_0[\times W$; furthermore, ϕ has the following property (I_0^∞):

$$(I_0^\infty) :\ \forall n \in \mathbb{N}\ \lim_{x \to 0} x^n \frac{\partial^n \phi}{\partial x^n}(x, \lambda) = 0 \text{ uniformly in } \lambda \in W, \tag{8.91}$$

- *If r_0 is irrational, then ϕ is identically equal to zero.*
- *If $r_0 = \frac{p}{q}$, $p \wedge q = 1$, then the expression (8.90) is not sufficient to overcome some of the problems we will study after.*

The following theorem has been proved by El Morsalani in [2], it will permit us to define for the Dulac map well ordered expansions:

For a family X_λ as in (8.89), we define the following unfolding of the logarithm:

$$\omega(x, \lambda) = \begin{cases} \frac{x^{-\alpha_1(\lambda)} - 1}{\alpha_1(\lambda)} & \text{if } \alpha_1(\lambda) \neq 0 \\ -\ln x & \text{if } \alpha_1(\lambda) = 0 \end{cases} \tag{8.92}$$

with

$$\alpha_1(\lambda) = r(0) - r(\lambda)$$

Theorem 5.2 ([2]). *For any $k \in \mathbb{N}$, there exists a neighbourhood W_k of $\lambda = 0$ in $\mathbb{R}^\Lambda$, some transverse segments σ and τ to the vector field X_λ parametrized in class C^k, respectively, by x and y in $[0, \epsilon[$ and C^∞ functions $\alpha_{ij} : W_k \to \mathbb{R}$ such that the Dulac map for the vector field given by the expression (8.89) has the equivalent forms:*

$$\begin{cases} y = D_\lambda(x) & = x^{r_0} + \alpha_1 x^{r_0}\omega + \sum_{1\le j\le i\le K(k)} \alpha_{ij} x^{(iq+1)r_0}\omega^j + \psi_k(x,\lambda) \\ & = x^{r(\lambda)}\left(1 + \sum_{1\le j\le i\le K(k)-1} \overline{\alpha}_{ij} x^{(iq)r_0}\omega^j + \overline{\psi}_k(x,\lambda)\right) \end{cases} \tag{8.93}$$

where α_{ij} and $\overline{\alpha}_{ij}$ are polynomials in $\alpha_1, \alpha_2, \cdots, \alpha_i$ of expression (8.89). The functions ψ_k and $\overline{\psi}_k$ are C^k functions k-flat with respect to $x = 0$ and they can be written:

$$\psi_k(x,\lambda) = x^k R_k(x,\lambda)$$
$$\overline{\psi}_k(x,\lambda) = x^k \overline{R}_k(x,\lambda)$$

where R_k and $\overline{R}_k$ satisfy the property I_0^∞.

We introduce a partial order between the monomials which corresponds to the flatness order in $x = 0$ and $\lambda = 0$

$$x^{l+n\alpha_1}\omega^m \prec (\text{less flat})\, x^{l'+n'\alpha_1}\omega^{m'} \Leftrightarrow \begin{cases} l < l' \\ \text{or} \\ l = l', \quad n = n' \text{ and } m > m' \end{cases}$$

8.5.2 *Pfaffian Equation Near a Hyperbolic Saddle Point*

Let P_0 be a hyperbolic saddle of a vector field X_λ as in the previous paragraph. Let k be an integer which will be fixed in this paragraph. So, up to C^k-equivalence, the family X_λ can have one of the expressions given by (8.88) and (8.89). Let $y = D_\lambda(x)$ be the graph of the Dulac map near the hyperbolic saddle point P_0.

Lemma 5.3. *The graph $y = D_\lambda(x)$ is an orbit of one of the following differential equations*

$$xdy - r(\lambda)ydx = 0 \quad \text{if} \quad r_0 \notin \mathbb{Q} \tag{8.94}$$

or

$$qxF(x^p,\lambda)dy + yF(y^q,\lambda)[-r_0 + F(x^p,\lambda)]dx = 0 \quad \text{if} \quad r_0 = \frac{p}{q} \in \mathbb{Q} \tag{8.95}$$

where

$$F(u,\lambda)=\sum_{i=0}^{N(k)}\alpha_{i+1}(\lambda)u^i. \tag{8.96}$$

Proof. The first assertion of the lemma is easy to prove. In this case, the Dulac map has a simple expression:

$$y=D_\lambda(x)=x^{r(\lambda)}.$$

To prove the second assertion, consider the family X_λ given by the expression (8.89):

$$X_\lambda\begin{cases}\dot{x}=x\\ \dot{y}=y(-r_0+\sum_{i=0}^{N(k)}\alpha_{i+1}(\lambda)(x^py^q)^i).\end{cases} \tag{8.97}$$

Let us perform the change of variables:

$$\begin{cases}x &= x\\ y &= x^{r_0}u\end{cases}$$

The system (8.97) is then transformed to the following polynomial system with separated variables

$$X_\lambda\begin{cases}\dot{x}=x\\ \dot{u}=\sum_{i=0}^{N(k)}\alpha_{i+1}(\lambda)u^{1+iq}=uF(u^q,\lambda).\end{cases} \tag{8.98}$$

The second equation of the system (8.98) can be written under the following form:

$$\frac{du}{uF(u^q,\lambda)}=dt \tag{8.99}$$

where t is the time variable. One has to remark that $u_{|\sigma}=x^{r_0}$ and $u_{|\tau}=y$, moreover $-\ln x$ is the necessary time to go from $x\in\sigma$ to τ. Integrating Eq. (8.99), we obtain

$$\int_{x^{r_0}}^{y}\frac{du}{uF(u^q,\lambda)}=\int_0^{-\ln x}dt \tag{8.100}$$

$G(x,\lambda)$ be a primitive function of let $G(x,\lambda)$ be a primitive function of $\frac{du}{uF(u^q,\lambda)}$ then the expression (8.100) can be written:

$$G(y,\lambda)-G(x^{r_0},\lambda)=-\ln x \tag{8.101}$$

After, a differentiation with respect to the variables x and y, we get

$$\frac{y}{yF(y^q,\lambda)}dy-\frac{r_0}{xF(x^{qr_0},\lambda)}dx+\frac{1}{x}dx=0$$

After some calculations, we obtain the result of the lemma. □

Definition 5.4. If r_0 is rational, then we say that the saddle point P_0 is of finite order (or non-formally linearizable) if there exists an integer k such that for $\lambda = 0$ the family X_λ given by the expression (8.89) is not reduced to the system:

$$\begin{cases} \dot{x} &= x \\ \dot{y} &= -r_0 y \end{cases}$$

It follows then that there exists an integer m such that the polynomial $F(.,\lambda)$ has the following form

$$F(u,0) = \alpha u^m + o(u^m) \text{ and } \alpha \neq 0,\ m \geq 1.$$

We call the integer m the order of the resonant saddle point P_0.

In [7], the authors have proved that if the hyperbolic saddle point P_0 has a finite order m then the perturbation X_λ, for any integer $k \geq 2m$ is C^k-equivalent to a polynomial family of vector fields with a degree independent of k. This degree is $2m$. So, in this case, the family can be written:

$$\begin{cases} \dot{x} &= x \\ \dot{y} &= y(-r_0 + F(x^p y^q, \lambda)) \end{cases} \tag{8.102}$$

where

$$F(u,\lambda) = \sum_{i=0}^{m-1} \alpha_{i+1}(\lambda) u^i + \alpha(\lambda) u^m (1 + \alpha_{2m+1}(\lambda) u^m)$$

with $\alpha(0) \neq 0$ and $\alpha_i(0) = 0$. This special expression will be used to prove one of our theorems. In the other cases, we will use the well-ordered expansions of the Dulac map near a hyperbolic saddle point.

8.6 Degeneracy of Khovanskii Procedure

In this paragraph we want to prove the following theorem:

Theorem 6.1. *Let X_0 be an analytic vector field on the plane which has a hyperbolic polycycle Γ with two vertices such that the ratios of hyperbolicity are rational and their product is 1. If the polycycle is non-identical (i.e. its return map is non-identically the identity map), satisfying some genericity conditions which will be exhibited later on and if one of its vertices has finite order, then there exists an integer N depending only on the germ of X_0 along Γ such that the cyclicity of Γ in any C^∞-deformation is bounded by N.*

8.6.1 Displacement Map

Let X_λ be a family of sufficiently differentiable vector fields on the plane, $\lambda \in R^\Lambda$. For $\lambda = 0$, X_0 has a polycycle Γ_0 as described above with two vertices P_1 and P_2 satisfying $r_1(0).r_2(0) = 1$ and $r_1(0) \in \mathbb{Q} :\ r_1(0) = \frac{p}{q}$ with $p \wedge q = 1$.

Therefore for any integer K sufficiently large, there exists a neighbourhood U_K of 0 in $\mathbb{R}^\Lambda$ and local charts A_1^K, A_2^K around, respectively, P_1 and P_2 such that X_λ is defined by two families X_λ^1 and X_λ^2 in the following way:

(1) In A_1^K of local coordinates (x_1, y_1), P_1 is the only singular point of X_λ^1 and the vector field is written:

$$X_\lambda^1 : \begin{cases} \dot{x}_1 = x_1 \\ \dot{y}_1 = y_1 \left(-r_1(\lambda) + \sum_{i=0}^{N_1(k)} \alpha_{i+1}(\lambda)(x_1^p y_1^q)^i\right) \end{cases} \tag{8.103}$$

where the α_i are sufficiently differentiable functions on $U_K \to \mathbb{R}$ that depend on K but their values for $\lambda = 0$ depend only on X_0.

(2) In A_2^K of local coordinates (x_2, y_2), P_2 is the only singular point of X_λ^2 and the vector field $(-X_\lambda^2)$ is written:

$$(-X_\lambda^2) : \begin{cases} \dot{x}_2 = -x_2 \left(s_2(\lambda) + \sum_{i=0}^{N_2(k)} \beta_{i+1}(\lambda)(x_2^q y_2^p)^i\right) \\ \dot{y}_2 = y_2 \end{cases} \tag{8.104}$$

where $s_2(\lambda)$ is the ratio of hyperbolicity 0f $(-X_\lambda^2)$ in P_2 and it fulfils $s_2(\lambda) = 1/r_2(\lambda)$, the functions $\beta_i(\lambda)$ are sufficiently differentiable and their values in $\lambda = 0$ depend only on X_0.

Remark. $s_2(0) = r_1(0) = \frac{p}{q}$. From now on, we note $s_2(\lambda) = s(\lambda)$, $r_1(\lambda) = r(\lambda)$ and $r_0 = r(0)$. Up a scale, we can suppose that the two local charts contain the balls

$$\{(x_i, y_i);\ \|(x_i, y_i)\| \leq 2\}$$

In the local charts A_i^K, the flow of X_λ^i defines the Dulac maps between the segments $\sigma_i = \{y_i = 1\}$ and $\tau_i = \{x_i = 1\}$. Now, let us define the "displacement map" $\delta(., \lambda) : \tau_2^+ \to \sigma_2$ by

$$\delta(y_2, \lambda) = -h(y_2, \lambda) + D_{2,\lambda} \circ S_{2,\lambda}(y_2) \tag{8.105}$$

where y_2 is the parameter of τ_2^+ and

$$\begin{aligned} h(y_2, \lambda) &= R_{1,\lambda} \circ D_{1,\lambda}(y_2) \\ S_{2,\lambda} &= R_{2,\lambda}^{-1} \end{aligned} \tag{8.106}$$

where the functions $R_{i,\lambda}$ are the regular transition maps defined on the segments τ_i:

$$R_{i,\lambda}(y_i) = \sum_{j=0}^{K} c_{ij}(\lambda) y_i^j + o(y_i^K) \tag{8.107}$$

The following change of coordinates on τ_2

$$y_2 = R(x,\lambda) = c_1(\lambda) + o(x), \quad c_1(\lambda) \neq 0 \tag{8.108}$$

transforms the regular map $S_{2,\lambda}$ into a genuine translation:

$$S_{2,\lambda}(y_2) = y_2 + b(\lambda)$$

where $b(\lambda) = S_{2,\lambda}(0)$. So the map δ has a new expression in the new parameter x as follows:

$$\delta(x,\lambda) = H(x,\lambda) - D_{2,\lambda}(x + b(\lambda)) \tag{8.109}$$

Lemma 6.2. *For any $K \in \mathbb{N}^*$, the map $H(x,\lambda)$ admits well-ordered expansions of order K which are written:*

$$H(x,\lambda) = \sum_{\substack{0 \le j \le i \le Kq+1 \\ m \ge 0;\ ir_0 + m \le Kq+1}} \gamma_{ijm}(\lambda) x^{ir_0+m} \omega^j + x^K \phi_{1,K}(x,\lambda) \tag{8.110}$$

where γ_{ijm} are functions of c_{ij} and α_i while ω is the compensator introduced above for the Dulac map $D_{2,\lambda}$ and $\phi_{1,K}$ is a function C^K, K-flat with respect to $x = 0$ and satisfies the property I_0^∞. In the previous development, the monomials $x^{ir_0+m}\omega^j$ are ordered by their order of flatness with respect to $x = 0$ when $\lambda = 0$.

The lemma is proved as theorem 1.1 [2].

Remark. To simplify the proof of Theorem 6.1, we will show it for the case $r_1(0) = r_2(0) = 1$.

As discussed in Sect. 8.2, the graph $y = D_{2,\lambda}(x + b)$ is an orbit of the differential 1-form:

$$\Omega = (x+b)F(x+b,\lambda)dy + yF(y,\lambda)(-1 + F(x+b,\lambda))dx = 0, \tag{8.111}$$

where F is the polynomial function:

$$F(u,\lambda) = \sum_{i=0}^{m-1} \beta_{i+1}(\lambda) u^i + \beta(\lambda) u^m (1 + \beta_{2m+1}(\lambda) u^m) \tag{8.112}$$

with $\beta_i(0) = 0$ for $1 \le i \le m$ and $\beta(0) \neq 0$. The integer m is the order of the saddle point P_2. Without loss of generality, we suppose that $\beta(0) = 1$.

Remark. We will work in some large class of differentiability.

The equation $\delta(x,\lambda)=0$ is equivalent to the following system:

$$\begin{cases} H(x,\lambda)-y & =0 \\ D_{2,\lambda}(x+b)-y & =0 \end{cases} \tag{8.113}$$

As $y=H(x,\lambda)$ is a connected graph the generalized Rolle's lemma (Khovanskii procedure) allows to assert that the number of solutions of the system (8.113) is at most 1 plus the number of the solution of the following system:

$$\begin{cases} H(x,\lambda)-y & =0 \\ \Omega\wedge D(H(x,\lambda)-y) & =0 \end{cases} \tag{8.114}$$

where $D(H)$ represents the total differential of H. The last system is equivalent to

$$\begin{cases} 0 & =\det\begin{pmatrix} D_x(H(x,\lambda)-y) & D_y(H(x,\lambda)-y) \\ yF(y,\lambda)(-1+F(x+b,\lambda)) & (x+b)F(x+b,\lambda) \end{pmatrix} \\ y & =H(x,\lambda) \end{cases} \tag{8.115}$$

An easy computation yields the following system:

$$\begin{cases} 0 & =(x+b)F(x+b,\lambda)H'(x,\lambda)+yF(y,\lambda)(-1+F(x+b,\lambda)) \\ y & =H(x,\lambda) \end{cases} \tag{8.116}$$

Substituting $y=H(x,\lambda)$ yields a new equivalent equation:

$$\begin{aligned} \delta_1(x,\lambda) &= (x+b)F(x+b,\lambda)H'(x,\lambda) \\ &\quad +H(x,\lambda)F(H(x,\lambda),\lambda)(-1+F(x+b,\lambda)) \\ &=0 \end{aligned} \tag{8.117}$$

The number of isolated roots of $\delta_1(x,\lambda)=0$ plus one bounds the number of isolated roots of $\delta(x,\lambda)=0$.

8.6.2 Nontriviality Order

To show Theorem 6.1 and to see how degenerate is the Khovanskii procedure, we will study the behaviour of the nontriviality order of δ_0 under this procedure. We suppose that $\delta(x,0)\not\equiv 0$ and this corresponds to Γ_0 is a non-identical polycycle. A result of Ilyashenko [6] ensures that in such case $\delta(x,0)$ is not infinitely flat and consequently $\delta(x,0)\sim x^k\ln^m x$ for some $k\in\mathbb{Q}^+$ and $m\in\mathbb{N}$ [2]. The degeneracy of Khovanskii procedure happens in case 2.c below.

For $\lambda = 0$, the map $\delta_1(x, 0)$ is reduced to the following expression:

$$\delta_1(x,0) = (x)F(x,0)H'(x,0) + H(x,0)F(H(x,0),0)(-1 + F(x,0)) \tag{8.118}$$

where

$$\begin{cases} F(x,0) & = x^m + \beta_{2m+1}x^{2m} \\ H(x,0) & = \gamma_{10}(0)x + \sum_{2\le j\le i} \gamma_{ji}(0)x^i(-\ln x)^j \end{cases} \tag{8.119}$$

All the hypotheses introduced in this part depend only on the germ of X_0 along the polycycle Γ_0. We will denote, in what follows, the coefficients $\gamma_{ij}(0)$ simply by γ_{ij}.

1. (Hyperbolic case) If $\delta(x, 0)$ is equivalent to x which corresponds to $\gamma_{10} \neq 1$, then $\delta_1(x, 0)$ is equivalent to x^{m+1}. Indeed put $H(x, 0) = \gamma_{10}x + h(x)$ where $h(x)$ contains all the terms of $H(x)$ which are flatter than x. Then the function $\delta_1(x, 0)$ is written:

$$\begin{aligned} \delta_1(x,0) = {} & (\gamma_{10}x + xh'(x))(x^m + \beta_{2m+1}x^{2m}) \\ & + (\gamma_{10}x + h(x))\left((\gamma_{10}x + h(x))^m + \beta_{2m+1}(\gamma_{10}x + h(x))^{2m}\right) \\ & \times (-1 + x^m + \beta_{2m+1}x^{2m}) \end{aligned} \tag{8.120}$$

   As $h(x)$ is at least equivalent to $x^2 \ln x$ we find that $\delta_1(x, 0)$ is

$$\delta_1(x,0) = \gamma_{10}(1 - \gamma_{10}^m)x^{m+1} + o(x^{m+1}) \tag{8.121}$$

   where $o(x^{m+1})$ is a function satisfying $\lim_{x\to 0} \frac{o(x^{m+1})}{x^{m+1}} = 0$.
2. When $\gamma_{10} = 1$ then the function h has the following form:

$$h(x) = ax^n \ln^p x + o(x^n \ln^p x) \text{ with } p = 0 \text{ or } 1 \text{ and } a \neq 0 \tag{8.122}$$

   Let us recall that for $\lambda = 0$ the displacement map is written:

$$\delta(x,0) = H(x,0) - D_{2,0}(x) \tag{8.123}$$

   So we have the following subcases:

2.a. If $2 \le n < m + 1$ which corresponds to

$$\delta(x,0) = ax^n \ln^p x + o(x^n \ln^p x) \tag{8.124}$$

   then we obtain

$$\delta_1(x,0) = (n - m - 1)x^{n+m} \ln^p x + o(x^{n+m} \ln^p x) \tag{8.125}$$

   that shows that the function $\delta_1(x, 0)$ is not infinitely flat in 0.

2.b If $m+1 < n \leq \infty$ which corresponds to

$$\delta(x,0) = x^{m+1}\ln x + o(x^{m+1}\ln x) \tag{8.126}$$

then the function $\delta_1(x,0)$ has the form:

$$\delta_1(x,0) = (x^{2m+1}\ln x + o(x^{2m+1}\ln x) \tag{8.127}$$

in this case $h(x)$ can be eventually infinitely flat when $n = \infty$.

2.c If $n = m+1$ (i.e. the saddle points have the same order), then we have to discuss two subcases. Let us begin by developing more the functions $h(x)$ and $D_{2,0}$. Indeed they can have the following form:

$$\begin{aligned} h(x) &= ax^{m+1}\ln x + cx^{m+1} + o(x^{m+1}) \\ D_{2,0}(x) &= x - x^{m+1}\ln x + o(x^{m+1}\ln x) \end{aligned} \tag{8.128}$$

If

$$a+1 \neq 0 \text{ which is equivalent to say that } \delta(x,0) \sim x^{m+1}\ln x \tag{8.129}$$

then

$$\delta_1(x,0) \sim x^{2m+1}\ln x \tag{8.130}$$

it follows that δ_1 is not infinitely flat.

If $a+1 = 0$ and $c \neq 0$, then one has to impose some new "generic conditions" to ensure that $\delta_1(x,0)$ is not infinitely flat. It is so surprising that such degeneracy occurs when the two saddle points have the same finite order and the polycycle is not identical. That shows that this case can become completely "degenerate" under the Khovanskii procedure even if it is "strongly generic".

Our theorem is then

Theorem 6.3. *Let X_0 be any analytic vector field on the plane which has a hyperbolic polycycle Γ_0 with two vertices such that their ratios of hyperbolicity are rational and their product is 1. If the polycycle is non-identical and satisfies one of the conditions 1., 2.a., 2.b., and 2.c. (the subcase $a+1=0$ in 2.c. is not included) and if one of its vertices has finite order, then there exists an integer N depending only on the germ of X_0 along Γ_0 such that the number of limit cycles bifurcating from Γ_0 in any analytic deformation is bounded by N.*

Proof. After multiplication by a monomial $x^I\omega^J$, for some fixed integers I and J, the functions $\delta_1(x,\lambda)$ have the following form:

$$\delta_1(x,\lambda) = \sum_{(i,j)\in S} \mu_{ij}(\lambda)x^i\omega^j + \rho(\lambda)x^N\omega^p(1+\phi(x,\lambda) \tag{8.131}$$

where

$$S = \{(i,j);\ j \leq I\ \mu_{ij}(0) = 0\}$$

is a finite set and $\rho(0) \neq 0$ and the function ϕ satisfies the property I_0^∞.

The equation $\delta_1(x,\lambda) = 0$ is studied in El Morsalani [2] where the author showed that there exists a uniform bound for the number of its isolated zeros. □

8.7 Finiteness Cyclicity of Hyperbolic 2-Polycycle When $r_1(0).r_2(0) = 1, r_1(0) \neq 1$ and $r_1(0) \in \mathbb{Q}$

In this paragraph we will announce our general theorem in the case: the two ratios of hyperbolicity are rational different from 1 and the polycycle is not identical.

8.7.1 Nice Asymptotic Developments

Let X_λ be an analytic unfolding of the polycycle Γ_0. The limit cycles of X_λ close to the polycycle Γ_0 correspond to the isolated solutions of the following system

$$\begin{cases} S_{2,\lambda}(y) = D_{1,\lambda}(x) & (1) \\ S_{1,\lambda}(x) = D_{2,\lambda}(y) & (2) \end{cases} \tag{8.132}$$

In the plane (x,y), Eqs. (1) and (2) represent curves C_1 and C_2 which are graphs in the variables (x,y). So between two intersections of C_1 and C_2 there exists a point $q = (x,y) \in]q_1, q_2[\subset C_1$ where a vertical translation of the curve C_2 cut tangentially the curve C_1 and this contact is isolated because the two curves are analytic for $x > 0$ and $y > 0$. Hence the number of isolated solutions of the system (8.132) is bounded by the number of solutions of the following system plus one:

$$\begin{cases} S_{2,\lambda}(y) = D_{1,\lambda}(x) & (1) \\ \frac{\partial D_{1,\lambda}}{\partial x}(x)\frac{\partial D_{2,\lambda}}{\partial y}(y) - \frac{\partial S_{1,\lambda}}{\partial x}(x)\frac{\partial S_{2,\lambda}}{\partial y}(y) = 0 & (3) \end{cases} \tag{8.133}$$

For a fixed $k \in \mathbb{N}$ we write the Dulac maps as in expression (8.12) in which appear the compensators ω_i, $i = 1,2$ associated with each saddle point. As the ratios of hyperbolicity are different from 1 for the parameter $\lambda = 0$ then Eq. (3) in the previous system is equivalent to the following equation:

$$Y = y(1 + \psi_{2,k}(y,\lambda)) = A(\lambda)x^{r(\lambda)}(1 + \psi_{1,k}(x,\lambda)) \tag{8.134}$$

with $r(\lambda) = \frac{1-r_1(\lambda)}{r_2(\lambda)-1}$, $r(0) = \frac{p}{q}$ and $A(0) > 0$. As proved in [11] the equation

$$Y = y(1 + \psi_{2,k}(y, \lambda)) \tag{8.135}$$

is equivalent to

$$y = Y(1 + \overline{\psi}_{2,k}(Y, \lambda)) \tag{8.136}$$

The function $\overline{\psi}_{2,k}$ has the same expression as $\psi_{2,k}$.

The system (8.133) is equivalent to the equation:

$$\delta(x, \lambda) = S_{2,\lambda}(y) - D_{1,\lambda}(x) = 0 \tag{8.137}$$

where $y = Y(1 + \overline{\psi}_{2,k}(Y, \lambda))$ and $Y = A(\lambda)x^{r(\lambda)}(1 + \psi_{1,k}(x, \lambda))$. Moreover the map $\delta(x, 0)$ is not identically zero. Using formula (8.93) it can be seen that

$$\psi_{1,k}(x, \lambda) = \sum_{1 \le i+j \le K(k)} \alpha_{ij}(\lambda) x^i Z_1^j + x^k R_{1,k}(x, \lambda) \tag{8.138}$$

with $Z_1 = x\omega_1(x, \lambda)$ and

$$\overline{\psi}_{2,k}(Y, \lambda) = \sum_{1 \le i+j \le K(k)} \beta_{ij}(\lambda) Y^i Z_2^j + Y^k R_{2,k}(Y, \lambda) \tag{8.139}$$

with $Z_2 = Y\omega_2(Y, \lambda)$. Recall that ω_i, $i = 1, 2$ are independent compensators as the one defined in expression (8.11) and functions $R_{i,k}$ satisfy the property I_0^∞.

Let us introduce $\mu(\lambda) = r(0) - r(\lambda)$ and the compensator

$$\omega_3(x, \lambda) = \begin{cases} \frac{x^{-\mu(\lambda)}-1}{\mu(\lambda)} & \text{if} \quad \mu(\lambda) \neq 0 \\ -\ln x & \text{if} \quad \mu(\lambda) = 0. \end{cases} \tag{8.140}$$

We introduce $Z_3 = x\omega_3(x, \lambda)$. We can always suppose that $r(0) = r_1(0) = \frac{p}{q} > 1$. This allows to write

$$Y = A(\lambda)x^{\frac{p-q}{q}}(x + \mu(\lambda)Z_3(x, \lambda))(1 + \psi_{1,k}(x, \lambda)) \tag{8.141}$$

Indeed

$$\begin{aligned} Y &= A(\lambda)x^{r(\lambda)}(1 + \psi_{1,k}(x, \lambda)) \\ &= A(\lambda)x^{r_0-\mu(\lambda)}(1 + \psi_{1,k}(x, \lambda)) \\ &= A(\lambda)x^{r_0}(1 + \mu(\lambda)\omega_3)(1 + \psi_{1,k}(x, \lambda)) \\ &= A(\lambda)x^{r_0-1}(x + \mu(\lambda)Z_3(x, \lambda))(1 + \psi_{1,k}(x, \lambda)) \end{aligned} \tag{8.142}$$

The computation of $Z_2(Y,\lambda)$ as a function of x requires the introduction of a new compensator

$$\omega_4(x,\lambda) = \begin{cases} \frac{x^{-\beta(\lambda)}-1}{\beta(\lambda)} & \text{if} \quad \beta(\lambda) \neq 0 \\ -\ln x & \text{if} \quad \beta(\lambda) = 0. \end{cases} \tag{8.143}$$

where $\beta(\lambda) = (r_2(0) - r_2(\lambda))r(\lambda)$. Introducing $Z_4 = x\omega_4$, we then have

$$Z_2(Y,\lambda) = \sum_{0\leq i+j+m+n\leq K(k)} \rho_{ijmn}(\lambda)x^{\frac{i}{q}}Z_1^j Z_3^m Z_4^n + x^k \overline{R}_k(x,\lambda) \tag{8.144}$$

where $Z_2 = Y\omega_2(Y,\lambda)$ with

$$\omega_2(x,\lambda) = \begin{cases} \frac{x^{-\beta_1(\lambda)}-1}{\beta(\lambda)} & \text{if} \quad \beta_1(\lambda) \neq 0 \\ -\ln x & \text{if} \quad \beta_1(\lambda) = 0. \end{cases} \tag{8.145}$$

where $\beta_1(\lambda) = r_2(0) - r_2(\lambda)$. To prove the nice expansion in (8.144) we replaced Y by its first expression in (8.142) and we put it in ω_2. We obtained then

$$\omega_2(Y,\lambda) = \frac{(A(\lambda))^{-\beta_1(\lambda)}x^{-\beta_1 r(\lambda)}(1+\psi_{1,k}(x,\lambda))^{-\beta_1} - 1}{\beta_1} \tag{8.146}$$

Now as in [2, Lemma 6.2], there exist an analytic function $C(\lambda)$ and a function $\overline{\psi}_1(x,\lambda)$ with the same properties as $\psi_1(x,\lambda)$ such that $(A(\lambda))^{-\beta_1} = 1 - \beta_1 C(\lambda)$, $x^{-\beta_1 r(\lambda)} = 1 + \beta(\lambda)\omega_4(x,\lambda)$ and $(1+\psi_{1,k}(x,\lambda))^{-\beta_1} = 1 - \beta_1(\lambda)\overline{\psi}_1(x,\lambda)$. Straightforward calculations give the result.

From above we obtain the following proposition:

Proposition 7.1. *For any $k \in \mathbb{N}$ there exist an integer $K(k)$, a neighbourhood $W_k \subset W$ of $0 \in \mathbb{R}^\lambda$ and a set $\overline{I} = [0,\epsilon]$ for some $\epsilon > 0$ such that*

$$\begin{aligned} \delta(x,\lambda) &= \delta_k(x,\lambda) + x^k R_k(x,\lambda) \\ \delta_k(x,\lambda) &= \sum_{0\leq i+j+m+n\leq K(k)} \gamma_{ijmn}(\lambda)x^{\frac{i}{q}}Z_1^j Z_3^m Z_4^n \end{aligned} \tag{8.147}$$

In this part we want to study the equation $\delta(x,\lambda) = 0$ in (8.147). Up a coordinate change

$$X = x^{\frac{1}{q}}$$

and new notations we can suppose that under the hypothesis of Proposition 7.1.

$$\delta_k(x,\lambda) = \sum_{0\leq i+j+m+n\leq K(k)} \gamma_{ijmn}(\lambda)x^i Z_1^j Z_2^m Z_3^n \tag{8.148}$$

where $Z_i = x\omega_i$, $I = 1, 2, 3$ with ω_i are independent compensators:

$$\omega_i(x, \lambda) = \begin{cases} \frac{x^{-\mu_i(\lambda)}-1}{\mu_i(\lambda)} & \text{if} \quad \mu_i(\lambda) \neq 0 \\ -\ln x & \text{if} \quad \mu_i(\lambda) = 0. \end{cases}$$

Moreover it is easy to see that $\delta(, .0)$ is not formally flat, i.e. its formal expansion in x and $\ln x$ is not identically zero. This fact follows from the condition that the polycycle Γ_0 is not identical and all the operations which have been accomplished do not destroy this property. Consequently, the map $\delta(x, \lambda)$ satisfies the properties required to apply the preparation theorem and using the main corollary one obtains the following theorem:

Theorem 7.2. *Let X_0 be any analytic vector field on the plane which has a hyperbolic polycycle Γ_0 with two vertices such that their ratios of hyperbolicity are rational different from* 1 *and their product is* 1. *If the polycycle is non-identical, then there exists an integer N depending only on the germ of X_0 along Γ_0 such that the number of limit cycles bifurcating from Γ_0 in any analytic deformation is bounded by N.*

8.8 Mourtada New Results

In the last years Mourtada has extended the results we exposed in this article [12]. In the following we will give a summary of these results.

Let Γ_p be a hyperbolic polycycle with p singularities and tangent to a real analytic vector field X_0 defined in a neighbourhood U_0 of Γ_p. The fundamental theorem proved in this work is the following:

Theorem 8.1. *Let X_ν be an analytic unfolding of X_0 with q parameters. Then there exist some integers N and L and some neighbourhoods $\Gamma_p \subset U \subset U_0$ and V neighbourhood of $0 \in R^q$ such that*

(i) for all $\nu \in V$, the number of limit cycles of X_ν in U is bounded above by N.
(ii) The multiplicity of each limit cycle is bounded above by L.

The result is very strong as it handles all the different cases of hyperbolic polycycles: identical or non-identical. The major idea is like the one we demonstrated here: the return map of the deformed polycycle will be "conjugated" via preparation procedure to a some finite jet which is a fewnomial. Khovanskii theory for fewnomials [10] can be applied in this case to these finite jets. In the next part we will give a summary of the basic ideas of Mourtada proof.

8.8.1 Local Algebras and Derivations

Let $(x, \alpha) = (x_1, \ldots, x_p, \alpha_1, \ldots, \alpha_q)$ be some local analytic coordinates on $((\mathbb{R}^{+*})^p \times \mathbb{R}^q, 0)$ and let $B_p = \{x_1 \times \ldots \times x_p = 0\}$. Let $\mathcal{B} \supset \mathbb{R}\{x, \alpha\}$ be a local ring of germs of analytic functions on $((\mathbb{R}^{+*})^p \times \mathbb{R}^q, 0)$ and continuous on $(B_p, 0)$. Let χ be a germ of vector fields at 0 with its components in the local algebra $\mathcal{B}$. The vector field χ acts infinitesimally as a derivation on $\mathcal{B}$. The derivations of interest fulfil the following properties:

(1) $\chi(\mathcal{B}) \subset \mathcal{B}$
(2) The set of singularities of χ satisfies $\mathrm{Sing}(\chi) \subset (B_p, 0)$
(3) $(B_p, 0)$ is invariant by the flow φ_χ.

A big part of the work of Mourtada is dedicated to show topological and algebraic finiteness properties of the elements of the local algebra $\mathcal{B}$ relatively to the action of the derivation χ on $((\mathbb{R}^{+*})^p \times \mathbb{R}^q, 0)$.

Consider $U \subset ((\mathbb{R}^{+*})^p \times \mathbb{R}^q, 0)$ an open set on which acts the derivation χ. Using the flow $\varphi_{\chi,U}$ associated with χ on U we can define the integral projection map along the orbits of χ:

$$\pi_{\chi,U} : U \to \tilde{U} = U/_{\varphi_{\chi,U}} \tag{8.149}$$

In the following we introduce some definitions that play a major role:

- A germ $f \in \mathcal{B}$ at $0 \in ((\mathbb{R}^{+*})^p \times \mathbb{R}^q$ is said to be χ-regular if there exists an open set U, like the one defined above, such that the degree of $\pi_{\chi,U}$ restricted to the zero set of f in U is finite.
- For $f \in \mathcal{B}$ the differential ideal $I_{\chi,f}$ is the ideal generated by $\langle \chi^n f;\ n \in \mathbb{N} \rangle$ in the ring $\mathcal{B}$.
- f is said to be χ-finite if it is χ-regular and its differential ideal $I_{\chi,f}$ is noetherian in a star extension of $\mathcal{B}$.
- f is said to be locally χ-finite if it is χ-regular and there exists a finite subdivision (U_i) of U invariant by χ and such that each restriction ideal $I_{\chi,f|_{U_i}}$ is noetherian in a star extension of the restriction ring $\mathcal{B}_{|U_i}$.
- A sub-algebra or a sub-class of $\mathcal{B}$ is said to be χ-finite (respectively locally χ-finite) if each of its elements is χ-finite (respectively, locally χ-finite).

8.8.2 Main Base Lemmas

The first lemma is fundamental as it describes the finiteness properties of "well prepared" germs.

Lemma 8.2 (Lemma of χ-Finiteness). *Let $\mathcal{B}_0 \subset \mathcal{B}$ be a χ-finite algebra (respectively, locally χ-finite) on $U_0 \subset ((\mathbb{R}^{+*})^p \times \mathbb{R}^q, 0)$ where U_0 is invariant by χ. We suppose that $\mathcal{B}_0$ is χ-stable ($\chi(\mathcal{B}_0) \subset \mathcal{B}_0$). Let $\mathcal{M}$ be the maximal ideal of $\mathcal{B}$*

and $\mathcal{M}_0 \subset \mathcal{M}$ a stable ideal. Let $\mathcal{N}_{U_0} \subset \mathcal{B}$ be the ideal of germs that are zero on $(U_0, 0)$. Then the class

$$\mathcal{C}_{\mathcal{B}_0,\mathcal{M}_0} = \{f \in g + \mathcal{M}_0 I_{\chi,g} + \mathcal{N}_{U_0};\ g \in \mathcal{B}_0\} \subset \mathcal{B}$$

is χ-finite on U_0 (respectively, locally χ-finite).

In this case the functions g and f are said to be χ-equivalent and one can speak of χ-equivalent algebras or classes.

We need the following definition before announcing the second lemma. An open $U \in ((\mathbb{R}^{+*})^p \times \mathbb{R}^q, 0)$ is said to be admissible if it satisfies:

- χ can be extended continuously to $\overline{U}$ and the singular set of χ in $\overline{U}$ is contained in $\partial_0 U = B_p \cap \overline{U}$.
- $\partial_0 U$ is a union of χ-orbits in $\overline{U}$.

U_0 is an elementary semi-analytic subset of $\mathcal{B}$ if there exist $V \in (U, 0)$ containing U_0 and some germs $f_1, \cdots, f_n, g_1, \cdots, g_m \in \mathcal{B}$ defined on V such that

$$U_0 = \{y \in V : f_1(y) > 0, \cdots, f_n(y) > 0, g_1(y) = 0, \cdots, g_m(y) = 0\}$$

U_0 is said to be semi-analytic of $\mathcal{B}$ if it is a finite union of elementary semi-analytic sets of $\mathcal{B}$.

The second lemma gives a way to construct χ-finite algebras.

Lemma 8.3 (Lemma of Tougeron Extension). *Let $\mathcal{B}_0$ be a χ-stable sub-algebra of $\mathcal{B}$. We suppose that*

(i) The semi-analytic sets described by $\mathcal{B}_0$ have a finite number of connected components
(ii) The ring $\mathcal{B}_0$ is noetherian in $\mathcal{B}$.
(iii) There exist $(p + q - 1)$ 1-forms $\Omega_j = \sum_{i=1}^{p+q} a_{ij} dy_i$ with $a_{ij} \in \mathcal{B}_0$ and an admissible open set U such that every orbit of χ in U is a transverse intersection of separating solutions of Ω_j.

Then the algebra $\mathcal{B}_0$ is χ-finite.

For the next lemma we need to introduce $\mathcal{I}_{\chi,f}[U]$ the differential sheaf of $f \in \mathcal{B}$ defined on U, an admissible open set. Its fibre $I_{\chi,f,m}$ (or equivalently $I_{\chi,f}(m)$) at a point $m \in U$ is the differential ideal $I_{\chi_m,f_m} \subset \mathcal{B}_m = \mathbb{R}\{y - y_m\}$ where χ_m and f_m are the germs of χ and f at m. Moreover $y - y_m$ are local coordinates at m.

Lemma 8.4 (Coherence Lemma). *The sheaf $\mathcal{I}_{\chi,f}[U]$ is coherent: precisely, for all $m \in U$, there exist an open set V_m and an integer l_m such that at every point $m' \in V_m$ the fibre $I_{\chi,f}(m')$ is generated by the germs at m' of $\langle f, \chi f, \cdots, \chi^{l_m} f\rangle$. Furthermore if the fibre at 0 $I_{\chi,f}(0) = I_{\chi,f}$ is noetherian in $\mathcal{B}$, then there exists an open set $V \subset (U, 0)$ subset of U such that the sheaf $\mathcal{I}_{\chi,f}[U \cup \partial_0 V]$ is coherent.*

The next lemma connects the previous defined sheaf and the projection $\pi_{\chi,U}$.

Lemma 8.5 (Roussarie Isomorphy Lemma). *The differential sheaf $\mathcal{I}_{\chi,f}$ is compatible with the projection $\pi_{\chi,U}$ in the following way: if γ is an orbit of χ in U and if $m_1, m_2 \in \gamma$, then the fibres $I_{\chi,f}(m_1)$ and $I_{\chi,f}(m_2)$ are isomorphic, with the isomorphism being the germ of the flow $\varphi_{\chi,U}$ in the neighbourhood of m_1 and m_2.*

Let γ be a regular orbit of χ in U and $\sigma_m \subset U$ is a transversal segment to γ at m. Let $i_{\sigma_m} : \sigma_m \to U$ be the canonical injection then we can define a star morphism by $i^*_{\sigma_m} : f \in \mathcal{B} \to i^*_{\sigma_m}(f) = f \circ i_{\sigma_m}$. The χ-transverse ideal along γ is defined as:

$$J_{\chi,f,\gamma} = i^*_{\sigma_m}(I_{\chi,f}(m)) = I_{\chi,f}(m)_{|\sigma_m} \subset \mathbb{R}\{\beta\} \ \forall m \in \gamma$$

where the analytic coordinates β on σ_m are first integrals of χ along the orbit γ.

Lemma 8.6 (Saturation Lemma). *For all $m \in \gamma$ we have*

$$I_{\chi,f}(m) = \pi^*_{\chi_m}(J_{\chi,f,\gamma})$$

The transverse lemma is linked to the coefficient ideal we introduced in the proof of the Theorem (1.6).

A natural question that arises: when γ adheres to 0, what is the link between the differential fibre and the saturation of the transverse ideal? It turns out that this link will be the strongest when the orbit is principal in U.

Definition 8.7. Let γ be an orbit of χ in U. γ is said to be principal if

(i) It adheres to 0
(ii) It has a an analytic transversal segment $\sigma_0 \subset U$ that intersects in at most one point each orbit of χ in U.
(iii) for every analytic transversal segment $\sigma \subset \sigma_0$ the saturation

$$\varphi_{\chi,U}(.,\sigma) = \pi^{-1}_{\chi,U}(\pi_{\chi,U}(\sigma))$$

is a neighbourhood of 0 in U.

The ideal $I((B_p, 0))$ is principal generated by $\theta = \prod_{j=1}^{p} x_j$. Then the link between the differential fibre and the saturation of the transverse ideal will be expressed in general as a double inclusion relaxing the equality along γ. Indeed we have a sort of Nullstellensatz of Hilbert

$$(\theta^n)\pi^*_{\chi_m}(J_{\chi,f,\gamma}) \subset I_{\chi,f} \subset \pi^*_{\chi_m}(J_{\chi,f,\gamma}) \qquad (*)$$

The smallest n of these integers satisfying the inclusion above is the multiplicity $m_\chi(f)$ of f relatively to χ. If the ring $\mathcal{B}$ has an asymptotic structure, then the defined multiplicity is closely linked to the algebraic multiplicity $ma_\chi(f)$ which is defined as the stationarity index of an ascending chain of transverse ideals that converge to $J_{\chi,f,\gamma}$. Therefore studying the action of χ on the finite jets of f and using the double inclusion (*) Mourtada were able to show that f is χ-equivalent to its jet

of order $ma_\chi(f)$. Applying the finiteness lemma above yields the desired finiteness properties.

8.8.3 A Small Sketch of the Tools Used in the Proof

The germ of the Dulac map of a hyperbolic saddle point and its unfoldings belong to some algebras $QR\mathcal{H}^{1,\cdot}$ defined as follows.

Definition 8.8. The Hilbert quasi-regular algebra $QR\mathcal{H}^{p,q}$ is defined in the following way. Let $q = (q_1, q_2) \in \mathbb{N}^2$ and $\alpha = (\mu, \nu)$ be some coordinates on $\mathbb{R}^{q_1} \times \mathbb{R}^{q_2}$. Define the elementary functions as $z_{i,0}(x_i) = x_i \log(x_i)$ and their unfoldings $z_{i,j}(x_i, \mu_j) = x_i Ld(x_i, \mu_j)$. Recall that $Ld(y, \beta) = (y^\beta - 1)/\beta$ for $\beta \neq 0$ and $Ld(y, 0) = \log(y)$ is the Ecalle–Roussarie compensator. We will use the following notations: $X_i = (x_i, z_{i,0}, z_{i,1}, \cdots, z_{i,q_1})$, $X = (X_1, \cdots, X_p)$ and $\hat{x}^i = (x_1, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)$. We will apply the following immersions too $c_i(x, \alpha) = (X_i, \hat{x}^i, \alpha)$ and $c(x, \alpha) = (X, \alpha)$. Let $QR\mathcal{H}^{0,q} = \mathbb{R}\{\alpha\}$. Now we can define the Hilbert quasi-regular algebra $QR\mathcal{H}^{p,q}(x, \alpha) \subset AQ^{p,|q|}(x, \alpha)$ as the set of germs f with an asymptotic development of "Hilbert type": for each $i = 1, \cdots, p$, there exist a sequence $(G_{i,m})_m$ in $QR\mathcal{H}^{p-1,q}(\hat{x}^i, \alpha)[X_i]$ of homogeneous polynomials in the variable X_i of degree m such that for any $n \in \mathbb{N}, f(x, \alpha) = \sum_{m=0}^{n} G_{i,m} \circ c_i(x, \alpha) + x^n h_n$ with $h_n \in SB^{p,|q|}$ and converging to 0 in each θ-sector when $w \to \infty$.

Let $\Xi[QR\mathcal{H}^{p,q}]$ be the $QR\mathcal{H}^{p,q}$-module of germs at 0 of vector fields with components in $QR\mathcal{H}^{p,q}$ and that keep the algebra $QR\mathcal{H}^{p,q}$ invariant. This module contains the sub-module generated by the elementary derivations $x_j \frac{\partial}{\partial x_j}$ for $j = 1, \cdots, p$. We will be interested in a particular sub-class of $\Xi[QR\mathcal{H}^{p,q}]$ that appears under a certain desingularization that is linked with the geometry of unfolded polycycle. This sub-class is denoted $\Xi\mathcal{H}[QR\mathcal{H}^{p,q}]$ and is defined as follows: let $k \leq p$ and $r_j = 1 + \mu_j$ for $j = 1, \cdots, k-1$, $x = (x_1, \cdots, x_k)$ and $x' = (x_{k+1}, \cdots, x_p)$. The elements of $\Xi\mathcal{H}_k[QR\mathcal{H}^{p,q}]$ are the germs at 0 of the vector fields χ satisfying the following conditions:

(a) if $k = 1$ then $\Xi\mathcal{H}_1[QR\mathcal{H}^{p,q}] = \{x_1 \frac{\partial}{\partial x_1}, \cdots, x_p \frac{\partial}{\partial x_p}\}$.

(b) if $k > 1$ then $\chi x_1 = \prod_{j=1}^{k} x_j$ and χ has as first integrals the coordinate functions $\alpha' = (x', \alpha)$ and $(k-1)$ germs $g_j(x_j, x_{j+1}, \alpha') = d_j(x_j, , \alpha') - x_{j+1}$ with $d_j = x_j^{r_j}(1 + D_j)$ and $D_j = O(x_j) \in QR\mathcal{H}^{p-k+1,q}(x_j, x', \alpha)$.

The integer $k - 1$ is called the non-triviality dimension of χ and the germs g_j are said to be the non-trivial first integrals of χ. This permits to define

$$\Xi\mathcal{H}[QR\mathcal{H}^{p,q}] = \cup_{k=1}^{p} \Xi\mathcal{H}_k[QR\mathcal{H}^{p,q}]$$

Let χ be a Hilbert derivation in $\Xi\mathcal{H}_k$ defined on an open set $U_k \in ((\mathbb{R}^{+*})^k \times \mathbb{R}^q, 0)$ and its non-triviality dimension is $k-1$. It acts on the algebra $QR\mathcal{H}^{k,\cdot}$ and it has a principal orbit in U_k. Then the main theorem is a consequence of the following one:

Theorem 8.9. *The algebra $QR\mathcal{H}^{k,\cdot}$ is locally χ-finite and satisfies the double inclusion (*) locally.*

To obtain the previous result above one has to desingularize a given χ. Indeed there exists a desingularization $(\pi_k, \mathcal{N}_k)$ described entirely by the algebras $QR\mathcal{H}^{k,\cdot}$ and such that the reduced singularities of χ are under the following form

$$\chi_l = \rho\frac{\partial}{\partial\rho} - \sum_{j=1}^{l} s_j u_j \frac{\partial}{\partial u_j}$$

for $l = 0, \cdots, k-1$. Theorem 8.9 is then a consequence of the study of the action of the reduced derivations χ_l on the algebras $QR\mathcal{H}^{p,\cdot}(\rho, \rho', .)$ with $p \leq q$. Now, a derivation χ_l on an open U_p admits a principal orbit if and only if $p = 1$. This yields to the following principal results.

Theorem 8.10. *The algebra $QR\mathcal{H}^{1,\cdot}(\rho, .)$ is χ_0-finite and satisfies the double inclusion (*).*

Recall that $\chi_0 = \rho\frac{\partial}{\partial\rho}$. This theorem leads to the other theorems. If we denote $QR\mathcal{H}^{1,\cdot}_{cvg}$ as the restriction of the analytic ring $\mathbb{R}\{.\}$ to the graph of elementary functions of the corresponding algebra $QR\mathcal{H}^{1,\cdot}$. Its χ-finiteness is a consequence of classical analytic geometry and the theory of Khovanskii–Tougeron [10, 16].

Theorem 8.11. *For every l, the algebra $QR\mathcal{H}^{1,\cdot}_{cvg}$ satisfies the double inclusion (*) relatively to χ_l.*

Theorem 8.12. *For every l, $QR\mathcal{H}^{1,\cdot}(\rho, .)$ is locally χ_l-finite and satisfies locally the double inclusion (*).*

The proof of the Theorem 8.9 is based on the three theorems above and the principal lemmas. Let $\overline{\mathcal{D}}_k$ be the exceptional divisor of the morphism $(\pi_k, \mathcal{N}_k)$ of the desingularization of χ. Let $f \in QR\mathcal{H}^{k,\cdot}$ and let $\tilde{f}$ and $\tilde{\chi}$ be the preimages of f and χ by π_k. The goal is to prove that the sheaf $\mathcal{I}_{\tilde{\chi},\tilde{f}}[\overline{\mathcal{D}}_k]$ is locally $\tilde{\chi}$-finite. The derivation $\tilde{\chi}$ has a unique singularity a_0 on $\mathcal{D}_k$. Let $\gamma_1 \subset \mathcal{D}_k$ be an orbit of $\tilde{\chi}$ and $a_1 = \overline{\gamma}_1 \cap \partial\mathcal{D}$. Using the compacity of $\overline{\mathcal{D}}_k$, it is sufficient to show that the sheaf $\mathcal{I}_{\tilde{\chi},\tilde{f}}[a_0\gamma_1 a_1]$ is locally $\tilde{\chi}$-finite.

The orbit $\gamma_0 = \pi_k^{-1}(\gamma)$ is principal in a neighbourhood U_{1,a_0} of a_0. Therefore the result at a_0 is a consequence of the fundamental Theorem 8.12. At every point $a \in \gamma_1$, a representative of the germ (γ_1, a) is principal in a neighbourhood $U_{1,a}$ of a; however, it is included in the boundary of $U_{1,a}$. Thanks to the coherence lemma, the results of Theorem 8.12 at a_0 can be germified at any point $a \in \gamma_1$

sufficiently close to a_0: the germ at a of $\tilde{f}$ is $\tilde{\chi}$ -equivalent to an element g_a of a convergent algebra $QR\mathcal{H}^{1,\cdot}_{cvg}$ which satisfies the fundamental Theorem 8.11. As $\tilde{f}$ can be extended above γ_1 to a function g whose germs belong to a convergent algebra. The isomorphy lemma applies to the sheaves of this algebra along the orbits included in the boundary. A gluing of the ideals of g and $\tilde{f}$ gives the results above γ_1. At a_1 Mourtada uses a recurrence argument on the non-triviality dimension of the Hilbert derivation and applies again the fundamental theorems and the principal lemmas.

The Dulac map of each singularity of X_v is induced by an element of an algebra $QR\mathcal{H}^{1,\cdot}$. The cycle limits of X_v correspond to the isolated intersections of orbits of Hilbert derivation $\chi \in \Xi\mathcal{H}_k$ and the fibres of a germ $f \in QR\mathcal{H}^{k,\cdot}$. Then Theorem 8.1 is a simple consequence of Theorem 8.9. Indeed, the property (1) is equivalent to the χ-regularity of f and the property (2) is a consequence of the fact that the differential ideal $I_{\chi,f}$ is noetherian or local noetherian.

References

1. Dumortier, F., El Morsalani, M., Rousseau, C.: Hilbert's 16th problem of quadratic systems and cyclicity of elementary graphics. Nonlinearity **9**, 1209–1261 (1996)
2. El Morsalani, M.: Bifurcations de polycycles infinis pour les champs de vecteurs polynomiaux. Ann. Fac. Sci. Toulouse **3**, 387–410 (1994)
3. El Morsalani, M., Mourtada, A., Roussarie, R.: Quasi-regularity property for unfolding of hyperbolic polycycles. Astérisque **220**, 303–326 (1994)
4. Françoise, J.P., Pugh, C.C.: Keeping track of limit cycles. J. Differ. Equ. **65**, 139–157 (1986)
5. Hilbert, D.: Mathematische probleme (lecture). In: The Second International Congress of Mathematicians Paris 1900, Nach. Ges. Wiiss. Gottingen, Math.-Phys. KI, 1900, pp. 253–297; Mathematical developments arising from Hilbert's problems. In: F. Browder (ed.) Proceeding of Symposium in Pure Mathematics, vol. 28, pp. 50–51. AMS, Providence, RI (1976)
6. Ilyashenko, Yu.: Limit cycles of polynomial vector fields with non-degenerate singular points in the real plane. Funct. Anal. Appl. **18**, 199–209 (1985)
7. Ilyashenko, Yu., Yakovenko, Yu.: Finitely smooth normal forms for local diffeomorphisms and vector fields. Russ. Math. Surv. **46**, 1–43 (1991)
8. Ilyashenko, Yu., Yakovenko, Yu.: Finite cyclicity of elementary polycycles in generic families. Am. Math. Soc. Transl. **165**, 21–95 (1995)
9. Joyal, P.: Un théorème de préparation pour les fonctions à développement tchébychévien. Ergod. Theory Dyn. Syst. **14**, 305–329 (1994)
10. Khovanskii, A.: Fewnomials. AMS, Providence, RI (1991)
11. Mourtada, A.: Cyclicité finie des polycycles hyperboliques des champs de vecteurs du plan: mise sous forme normale. In: Lecture Notes in Mathematics, vol. 1455, pp. 272–314. Springer, New York (1990)
12. Mourtada, A.: Action de dérivations irreductibles sur les algèbres quasi-régulières d'Hilbert. Preprint. arXiv: 0912.1560 v.1, 81 pp. (2009)
13. Moussu, R., Roche, C.: Théorème de Khovanskii et problémes de Dulac. Invent. Math. **105**, 431–441 (1991)

14. Roussarie, R.: A note on finite cyclicity and Hilbert's 16th problem. In: Lecture Notes in Mathematics, vol. 1331, pp. 161–168. Springer, New York (1988)

15. Roussarie, R.: Cyclicité finie des lacets et des points cuspidaux. Nonlinearity **2**, 73–117 (1989)

16. Tougeron, J.-C.: Algébres analytiques topologiquement noethériennes et théorie de Khovanski. Ann. Inst. Fourier tome **41**, fasc. 4, 823–840 (1991)

Chapter 9
Self-Inversive Cubic Curves

Raymond R. Fletcher

Abstract Let γ denote an irreducible nonsingular cubic curve which inverts onto itself with respect to a circle ω with center X. Depending on the type of γ, we show that γ inverts onto itself via a second circle orthogonal to ω or that γ inverts onto itself via two additional circles υ, η with ω,υ,η mutually orthogonal. To accomplish this an algebra (γ,δ) with a ternary operation δ is defined on the points of γ by setting $\delta(a,b,c)$ equal to the fourth point, counting multiplicities, on circle (a,b,c) and on γ. If * is the binary operation defined on the points of γ by setting a*b equal to the third point, counting multiplicities, on the line [a,b] and on γ, we show that $\delta(a,b,c) = X*((X*a)*(b*c))$. This equation is used extensively to determine automorphisms of (γ,δ) and to discuss subalgebras.

Keywords Cubic curve • Inversion • Ternary operation • Circle chain • Triple system • Automorphism • Subalgebra • Perfect polygon

9.1 Introduction

We present here a study of cubic curves which invert onto themselves. This topic arose naturally from an investigation of *group circle systems.* If G is an abelian group and $g \in G$, then a *(G,g) circle system* is defined as follows. Let $\phi: G \to \Pi$ be an injective mapping from G into the projective plane Π such that no five points in $\phi(G)$ are cocyclic. If for every four element subset {a,b,c,d} of G with $a + b + c + d = g$, the corresponding points $\{\phi(a), \phi(b), \phi(c), \phi(d)\}$ are cocyclic or collinear, we call the set of points $\phi(G)$ and the corresponding circles or lines, a *(G,g) circle system.* To avoid repeated use of the expression "cocyclic or collinear," we coin the terms *cirkle and cocyklic* to embrace both possibilities. We found that the points (or *vertices*) of a group circle system lie on a self-inversive algebraic curve, and that such systems can be constructed on any irreducible nonsingular self-inversive cubic curve. In particular, for a (Z,0) circle system, there exists a circle ω which inverts each pair

R.R. Fletcher (✉)
Department of Mathematics and Economics, Virginia State University, Petersburg, VA 23806, USA
e-mail: cielbleu66@gmail.com

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_9

of vertices of the form {q,−q}. We will show that the same circle ω also inverts the algebraic curve γ which serves as an envelope for the vertices of the system. In case γ is a cubic we obtain two orthogonal circles or even three mutually orthogonal circles which invert γ onto themselves. We begin with the development of a formula for an irreducible self-inversive cubic. The polar version of this formula is quadratic in r, and thus provides an efficient means for graphing and investigating properties of such curves.

9.2 The Equation of a Self-Inversive Cubic Curve

Let γ be a nonsingular irreducible cubic curve which inverts onto itself via a circle ω. By translation and dilation, we can assume that ω is the unit circle with center O at the origin and radius 1. If P is a point in the Euclidean plane different from O, then the inverse of P wrt ω is the point P′ on ray OP such that $(OP)(OP') = 1$. If $P = (a,b)$, then we find that $P' = (a/(a^2 + b^2), b/(a^2 + b^2))$. There is a point at infinity on γ whose inverse wrt to ω can be naturally defined to be the center O of ω. Thus the center (0,0) of ω must lie on γ, and so the constant term in the equation of γ is 0. The equation of a general cubic γ with constant term zero is given by

$$ax^3 + bx^2y + cxy^2 + dy^3 + ex^2 + fxy + gy^2 + hx + ky = 0. \tag{9.1}$$

We obtain the inverse γ′ of γ by substituting $x/(x^2 + y^2)$ for x and $y/(x^2 + y^2)$ for y in (9.1). If we then clear fractions by multiplying by $(x^2 + y^2)^3$ we obtain

$$\begin{aligned} ax^3 + bx^2y + cxy^2 + dy^3 + \left(x^2 + y^2\right)\left(ex^2 + fxy + gy^2\right) \\ + \left(x^2 + y^2\right)^2 (hx + ky) = 0. \end{aligned} \tag{9.2}$$

In order for γ′ to be identical with γ they must both have degree 3. This can be achieved if (i) $e = f = g = h = k = 0$, or (ii) $x^2 + y^2$ occurs as a factor of $ax^3 + bx^2y + cxy^2 + dy^3$. In case (ii) $x^2 + y^2$ can be factored out of (9.2), leaving a cubic. In case (i) the equation of γ reduces to $ax^3 + bx^2y + cxy^2 + dy^3 = 0$. Since we are assuming that γ is irreducible, it must be that a, d are nonzero, otherwise, either y or x would be a factor. Dividing by a and renaming the remaining coefficients, we obtain $x^3 + bx^2y + cxy^2 + dy^3 = 0$ as the equation of γ. Let q be a real number and divide $x^3 + bx^2y + cxy^2 + dy^3$ by $x + qy$ to obtain quotient: $x^2 + (b - q)xy + (c - bq + q^2)y^2$, and remainder: $(d - cq + bq^2 - q^3)y^3$. The cubic equation $d - cq + bq^2 - q^3 = 0$ has at least one real solution for q, and consequently $x^3 + bx^2y + cxy^2 + dy^3$ has a linear factor. Since this contradicts the irreducibility of γ, we conclude that case (i) is inadmissible. So we must have case (ii), i.e., $x^2 + y^2$ must be a factor of $ax^3 + bx^2y + cxy^2 + dy^3$. Dividing $ax^3 + bx^2y + cxy^2 + dy^3$ by $x^2 + y^2$ gives a quotient: $ax + by$ and a remainder: $(c - a)xy^2 + (d - b)y^3$.

The remainder must be identically 0, so we must have $c = a$ and $d = b$. Equation (9.2) reduces to

$$\left(x^2 + y^2\right)(ax + by) + \left(x^2 + y^2\right)\left(ex^2 + fxy + gy^2\right) + \left(x^2 + y^2\right)^2 (hx + ky) = 0. \tag{9.3}$$

Now factoring out $x^2 + y^2$ we obtain the cubic:

$$hx^3 + kx^2y + hxy^2 + ky^3 + ex^2 + fxy + gy^2 + ax + by = 0. \tag{9.4}$$

By comparing (9.1) and (9.4) we see that we must have $h = a$ and $k = b$. Then, after factoring out $x^2 + y^2$ (9.3) reduces to

$$(ax + by)\left(x^2 + y^2 + 1\right) + \left(ex^2 + fxy + gy^2\right) = 0. \tag{9.5}$$

This is our final version for the equation of an irreducible cubic curve which inverts onto itself via the unit circle. To insure that (9.5) represents a cubic, one of a,b must be nonzero, and we must also assume that the coefficients a,b,e,f,g are chosen so that (9.5) remains irreducible. For the purpose of efficient graphing, the polar version of (9.5) is very useful. Substituting $x = r\cos\theta$ and $y = r\sin\theta$ into (9.5) and factoring out r we obtain

$$(a\cos\theta + b\sin\theta)\left(r^2 + 1\right) + r\left(e\cos^2\theta + f\sin\theta\cos\theta + g\sin^2\theta\right) = 0. \tag{9.6}$$

Dividing (9.6) by $a\cos\theta + b\sin\theta$ we obtain

$$r^2 + rf(\theta) + 1 = 0 \quad \text{where}$$
$$f(\theta) = \left(e\cos^2\theta + f\sin\theta\cos\theta + g\sin^2\theta\right) / (a\cos\theta + b\sin\theta). \tag{9.7}$$

Solving (9.7) for r, we obtain $r = \left(-f(\theta) \pm \sqrt{(f(\theta))^2 - 4}\right)/2$, and it is not difficult to show that the full graph of γ is obtained by

$$r = \frac{-f(\theta) + \sqrt{(f(\theta))^2 - 4}}{2}. \tag{9.8}$$

The homogenization of (9.5) is

$$(ax + by)\left(x^2 + y^2 + z^2\right) + \left(ex^2 + fxy + gy^2\right)z = 0. \tag{9.9}$$

To find the points at infinity on γ, we set $z = 0$ in (9.9) to obtain $(ax + by)(x^2 + y^2) = 0$. Since (0,0,0) is not a point in the projective plane, we must have $ax + by = 0$. If b is nonzero, we obtain the point (1, −a/b, 0), and if $b = 0$, we obtain the point (0,1,0), i.e., the point at infinity on vertical lines. Thus there is a unique point at infinity on γ and we will designate it henceforth by the symbol ∞.

9.3 The (Z,0) Circle System

The purpose of this section is to show that a (Z,0) circle system exists and can be constructed on any irreducible nonsingular self-inversive cubic curve γ. Suppose Ω is a (Z,0) circle system and consider the cirkles (1,−1,2,−2), (1,−1,3,−3), (2,−2,3,−3) of Ω. (Note that for ease of notation we refer to the vertices of Ω by their integer labels.) These cirkles must occur in one of two possible orientations indicated in Figs. 9.1 and 9.2. In Fig. 9.1 the radical center R of the three circles

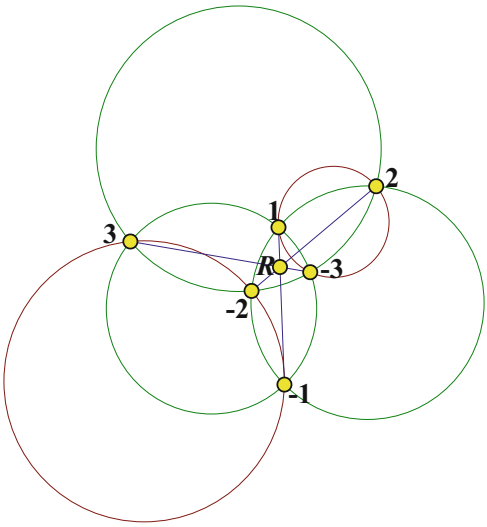


Fig. 9.1 Impossible orientation for three circles of a (Z,0) circle system

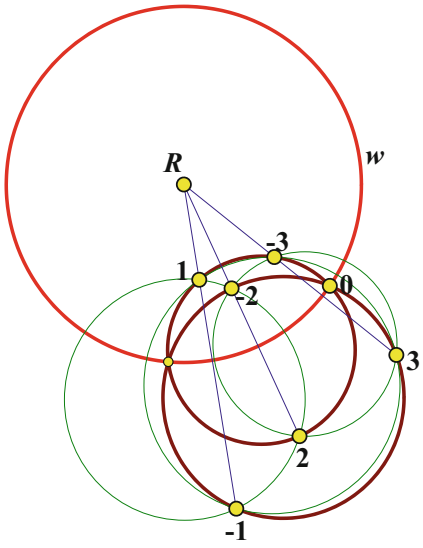


Fig. 9.2 For a (Z,0) circle system there exists a circle w which inverts every pair of points of the form {q,−q}

occurs in the interior of each circle, and in Fig. 9.2 the radical center occurs in the exterior of all three circles. Since 5 collinear points are not allowed, at most one of the three cirkles can be a line L. In this case the intersection of L with the line joining the contact points of the remaining two circles is R. No matter how the six labels ±1, ±2, ±3 are assigned to the points in Fig. 9.1, the cirkles (1,2,−3), (−1,−2,3) are disjoint. However, in Ω, the cirkles (1,2,−3,0), (−1,−2,3,0) are obviously not disjoint. We conclude that our three topical cirkles must occur in the orientation illustrated in Fig. 9.2. As in Fig. 9.2, let R denote the radical center and let w be a circle with center R which is orthogonal to circle (1,−1,2,−2). Then {1,−1}, {2,−2}, {3,−3} must all be inverse pairs wrt w. If $q \in Z$ and $q \neq 0$, ±1, ±2, ±3, then the similar set of cirkles: (1,−1,q,−q), (1,−1,2,−2), (2,−2,q,−q) must also have R as the radical center. As a consequence every pair of points {q,−q} is an inverse pair wrt w, and every circle of Ω with the form (s,−s,q,−q) is orthogonal to w. The cirkles (1,2,−3,0), (−1,−2,3,0) are inverse wrt w, so their common point 0 must lie on w. We have proved:

Theorem 1. *Let Ω be a (Z,0) circle system. Then there exists a circle ω which contains vertex 0 and which inverts every pair of points of the form {q,−q}.* □

For the construction of a (Z,0) circle system we need the following results, the proof of which can be found in [1].

Lemma 2. *Any five of the cirkles (A,B,C,D), (E,F,G,H), (A,B,F,E), (B,C,G,F), (C,D,H,G), and (D,A,E,H) imply the sixth. (See Fig. 9.3.)* □

We will refer to this configuration of six cirkles as a *circle chain*. The following is a related result involving tangent cirkles.

Lemma 3. *If (C,B,F,G), (B,A,D,F), (F,G,E,D), and (C,A,E,G) are cirkles, then cirkles (A,B,C), (A,D,E) are tangent at A. (See Fig. 9.4.)* □

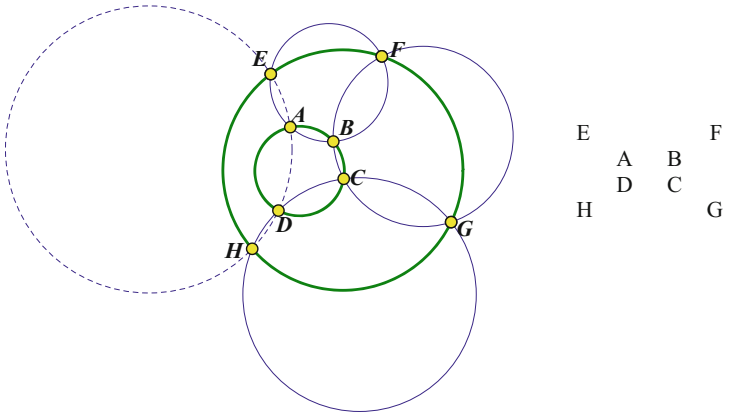


Fig. 9.3 Circle chain

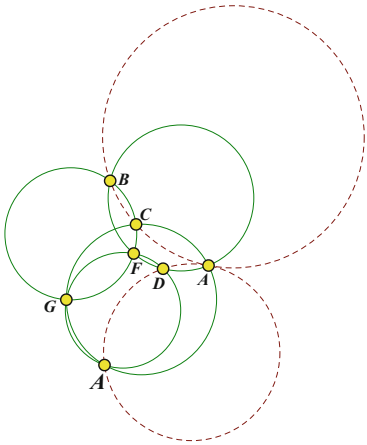


Fig. 9.4 Circle chain with tangent circles

Lemma 3 is a consequence of Theorem 4. Figures 9.3 and 9.4 also include a schematic representation of the circle chains which will be useful in applying these theorems. The inner and outer squares of points represent cirkles, and the four points on each of the adjacent parallel sides of these squares represent the remaining four cirkles.

Theorem 4. *Let ω be a circle and put vertex 0 on ω. Choose distinct vertices 1, 2 in the plane not on ω so that 1,2 are not inverses with respect to ω, and let −1, −2 be the inverses of 1, 2, respectively, wrt ω. Let vertex −3 lie on cirkle (0,1,2) and not on ω, and let vertex 3 be the inverse of −3 wrt ω. In general, if k is a positive integer and vertices* $\{-k+1, \ldots, k-1\}$ *have been defined, let* $-k = (0,1,k-1) \cap (-1,2,k-1)$*, and let vertex k be the inverse of −k wrt ω. Let Ω denote the set of integer labeled points defined in this way. We assume that the generators {1,2,3} of Ω have been chosen so that no two points in Ω have been labeled with the same integer. Then Ω is a (Z,0) circle system.*

Proof. It must be shown that every four element subset {a,b,c,d} of Z with $a+b+c+d=0$ corresponds to four cocyklic points. We find it useful to show simultaneously that if (a,b,c), (d,e,f) are two cirkles such that $a+b+c=d+e+f=m$, then these cirkles are tangent at vertex −m. We observe that if (a,b,c,d) is a circle, then so is (−a,−b,−c,−d) since inversion maps circles to circles. If m is any positive integer, then circle (0,m,−m) is inverted onto itself and so is orthogonal to ω. If (0,k,−k) is a second such circle, then it must be tangent to (0,m,−m) at vertex 0. We will proceed by induction on k ($k>0$) by showing that all four element subsets of the form {−k, a,b,c} such that $a+b+c=k$ and such that a,b,c are distinct integers which lie properly between −k and k correspond to four cocyklic points. We call these cirkles and their inverses the *circles of level k.* We will also show that all cirkles $\{(-k,s,k-2s): k \neq 3s, s \in \{1,2, \ldots, k-2\}\}$ are tangent at s to every cirkle of the form (−q, s, q − 2s) with $0 \leq q < k$. We call these tangencies and their inverses the *tangencies at level k.* To inaugurate the induction we find it

useful to prove by hand all cirkles and tangencies of levels 4,5. Since vertex −4 is defined to be the intersection point, besides 3, of cirkles (0,1,3), (−1,2,3) we have the cirkles (−4,0,1,3), (−4,2,3,−1) and their inverses (4,0,−1,−3), (4,−2,−3,1). This accounts for all cirkles of level 4. The circle chains:

$$\begin{array}{cccccccc} -2 & & & 4 & 1 & & & 3 \\ & 1 & -3 & & & -3 & -1 & \\ & 2 & 0 & & & 4 & 0 & \\ -1 & & & -1 & -2 & & & -2 \end{array}$$

account for all tangencies at level 4. We have defined vertex −5 to be the intersection point, besides 4, of cirkles (0,1,4), (−1,2,4), so we have the cirkles (−5,0,1,4), (−5,−1,2,4) and their inverses (5,0,−1,−4) and (5,1,−2,−4). The two remaining level 5 cirkles with −5 are (−5,0,2,3) and (−5,3,4,−2). The following two circle chains prove these cirkles:

$$\begin{array}{cccccccc} -1 & & & -4 & 0 & & & -1 \\ & 2 & 3 & & & 3 & -2 & \\ & -5 & 0 & & & -5 & 4 & \\ 4 & & & 1 & 2 & & & -1 \end{array}$$

These two cirkles and their inverses account for all level 5 cirkles. To prove the level 5 tangencies, it suffices to show that cirkles (−5,1,3), (−2,1,0) are tangent at vertex 1; cirkles (−5,2,1), (−3,2,−1) are tangent at vertex 2, and cirkles (−5,3,−1), (−4,3,−2) are tangent at vertex 3. These are proved by the following circle chains:

$$\begin{array}{cccccccccc} 0 & & & -2 & -1 & & -3 & -2 & & -4 \\ & 4 & -2 & & & 4 & 0 & & 4 & 2 \\ & -5 & 3 & & & -5 & 1 & & -5 & -1 \\ 1 & & & 1 & 2 & & & 2 & 3 & & & 3 \end{array}$$

Now suppose k ≥ 6 and assume inductively that all cirkles and tangencies for every level less than k have been proved. To complete the induction we must prove all cirkles and tangencies of level k. There are three phases in the induction; in the first phase, we will prove all level k cirkles which contain vertices {−k, 0}. We have by definition the cirkles (−k, 0, 1, k − 1), (−k, −1, 2, k − 1), and so we have their inverses (k,0,−1, −k + 1), (k,1,−2,−k + 1). The circle chains:

$$\begin{array}{cccccccc} 1 & & & -k+1 & 0 & & & -k+2 \\ & 0 & k-2 & & & 1 & k-3 & \\ & -k & 2 & & & -k & 2 & \\ k-1 & & & -1 & k-1 & & & -1 \end{array}$$

prove the cirkles (−k,0,2,k − 2), (−k,1,2,k − 3). Note that since k ≥ 6, these cirkles involve four distinct points. Also note that all entries in these circle chains, except for −k, lie properly between −k and k. Now assume inductively (this is an induction within an induction!) that we have proved the cirkles (−k,0, t, k − t), (−k,1, t, k − t − 1) for some t ∈ {2,3,4, ..., m} where we take m = −1 + ((k − 1)/2) if k is odd and m = −1 + ((k − 2)/2) if k is even. The circle chains:

$$\begin{array}{ccccccccc}
0 & & & & -t-1 & -k+2t+1 & & & -2t-1 \\
& k-t & & -k+2t+1 & & & k-t-1 & t+1 & \\
& -k & & k-t-1 & & & -k & 0 & \\
t & & & & 1 & & k-t & & t
\end{array}$$

prove cirkles (−k, k − t, k − t − 1, −k + 2t + 1) and (−k, 0, t + 1, k − t − 1), then the circle chain:

$$\begin{array}{ccccc}
t & & & & -k+1 \\
& & 1 & k-t-2 & \\
& & -k & t+1 & \\
k-t-1 & & & & 0
\end{array}$$

proves cirkle (−k, 1, t + 1, k − t − 2) and so completes the inner induction. We thus obtain all level k cirkles which contain vertex 0. In the second phase of the induction we prove the remaining level k cirkles. First consider cirkles of the form (k,−s,−r,t) where 0 < s < r < k; k + t = s + r, and t ∈ {1,2, ..., k − 3}. This accounts for all cirkles at level k which contain two positive values k, t. The circle chain:

$$\begin{array}{ccccc}
-k+s & & & & k-t \\
& & -s & t & \\
& & k & -r & \\
0 & & & & -k+r
\end{array}$$

proves the cirkle (k, −s, −r, t). Note that the cirkles (k, 0, −r, −k + r), (k, 0, −s, −k + s) exist by the first phase of the induction. It remains to prove cirkles of the form (k,−s,−t,−r) where 0 < t < s < r < k, and k = s + t + r. The circle chain:

$$\begin{array}{ccccc}
s-1 & & & & t+1 \\
& & -s & -t & \\
& & k & -r & \\
-k+1 & & & & r-1
\end{array}$$

contains cirkles (k, −r, r − 1, −k + 1), (k, −s, s − 1, −k + 1) each of which contains two positive entries, {k, r − 1} and {k, s − 1}. Since these have been proved by the previous circle chain, the present circle chain proves the cirkle (k, −s, −t, −r). The second phase of the induction is now complete; we have proved all level k cirkles.

Since tangencies at lower levels have been used to prove level k cirkles, we must, in phase 3, prove all level k tangencies in order to complete the induction. Let U_s denote the cirkle $(-k, s, k-2s)$ where $k \neq 3s$ and $s \in \{1,2,\ldots, k-2\}$. Then the circle chain:

$$\begin{array}{ccccc} 1-s & & & & -1-s \\ & k-1 & & 2s-k+1 & \\ & -k & & k-2s & \\ s & & & & s \end{array}$$

shows that U_s is tangent at vertex s to the cirkle $(s, 1-s, -1-s)$, and thus is tangent at s to all cirkles $(s, q, -q-2s)$ with vertices which lie properly between $-k$ and k. The induction is now complete, and the theorem is proved. □

The construction of Theorem 4 is illustrated in Fig. 9.5. It is very interesting that such a construction is possible, and perhaps even more striking that the vertices of a (Z,0) circle system appear to lie on a closed curve. We will, in fact, show that they lie on an algebraic curve with maximum degree 6.

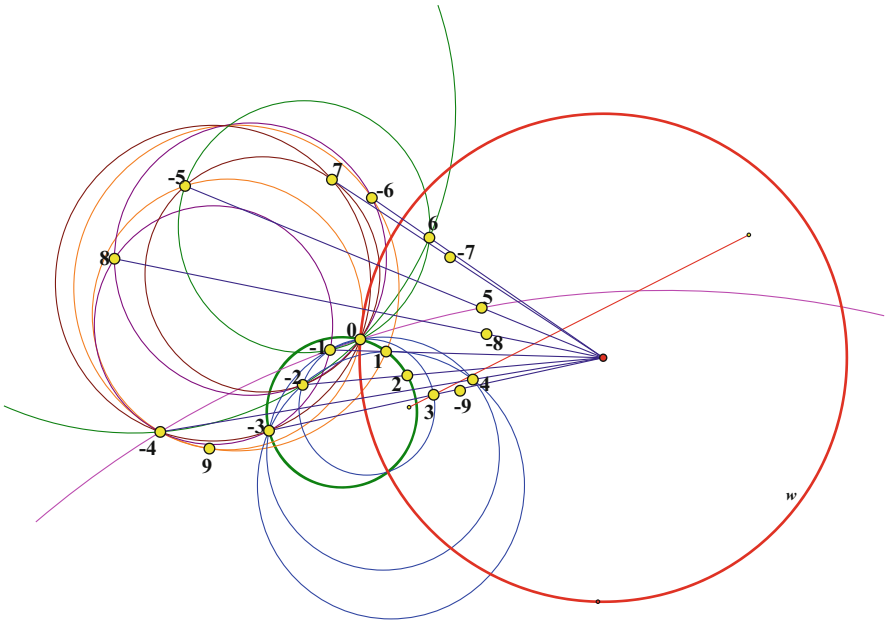


Fig. 9.5 Construction of (Z,0) circle system

9.4 The (Z,0) Geometric Triple System

Let (G,+) be an abelian group and $g \in G$. Let Ψ: $G \to \Pi$ be an injective mapping from G into the projective plane such that no four points in $\Psi(G)$ are collinear. If for each three element subset {a,b,c} of G with $a + b + c = g$, the corresponding points $\{\Psi(a), \Psi(b), \Psi(c)\}$ are collinear, then we call the set of points $\Psi(G)$ and the associated collinear triples, a *(G,g) geometric triple system*, or simply a *(G,g) triple system*. Suppose Ω is a (Z,0) circle system and we invert by a circle ω with center at vertex 0 to obtain a collection of points Ω'. All circles in Ω of the form (a,b,c,0) are inverted into lines [a,b,c] with $a + b + c = 0$ in Ω'. Pairs {q,−q} of points in Ω form mutually parallel lines in Ω'. We regard the point at infinity on this set of lines to be the inverse of 0, and thus we obtain all lines $\{[q, -q, 0]: q \in Z, q \neq 0\}$ in Ω'. In short, Ω' is a (Z,0) triple system. In fact, Ω' is also a (Z,0) circle system since inversion maps cirkles to cirkles. This connection between Ω and Ω' will explain the appearance of a curve containing the points of a (Z,0) circle system.

If γ is an irreducible cubic curve, then a binary operation * can be defined on the nonsingular points of γ by setting a*b equal to the third point on line [a,b] and on γ. Here we use the fact that every line which meets γ in two points, also meets γ in a third point, counting multiplicities. The product a*a refers to the third point besides a which lies on γ and on the tangent to γ at a. In case $a*a = a$, the point a is called a *flex*. Nonsingular irreducible cubic curves are known to have three collinear flexes in the projective plane. The following is a well known result which can be found in [2].

Theorem 5. *If γ is an irreducible cubic curve, then a binary operation * can be defined on the nonsingular points of γ by setting a*b equal to the third point on line [a,b] and on γ. The following identities are satisfied by *: (i) $a*b = b*a$; (ii) $a*(a*b) = b$; and (iii) $(a*b)*(c*d) = (a*c)*(b*d)$.* □

The commutative and absorptive identities (i) and (ii) are obvious, but identity (iii) is a remarkable property of irreducible cubic curves. We shall call (iii) the *hypercommutative property*, and the collection of algebras satisfying (i), (ii), and (iii), the *binary hypercommutative* variety. Besides the cubic curve model given in Theorem 5, we can obtain members of this variety by taking any abelian group (G,+) and any fixed element g of G, and defining $a*b = g - (a + b)$. We want to show that the vertices of a (Z,0) triple system lie on an irreducible cubic curve. To accomplish this, the following lemma is essential. If T is a (G,t) triple system, then a *t-square* for T is a 3×3 matrix with entries from G such that each row and each column represents a triple belonging to T.

Lemma 6. *Let T be a (G,t) triple system. If eight of the entries of a t-square for T lie on an irreducible cubic curve γ, then so does the ninth.*

Proof. Let $\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & k \end{vmatrix}$ be a t-square for T. Since the rows and columns of a t-square can be permuted independently without affecting the row or column sums, we can suppose that vertex a is the only element of G not known to lie on γ. Then the second two rows and columns represent triples of T. Thus the (*) product of any two of these elements is the third element of the triple. So we have b*c = (e*h)*(f*k) = (e*f)*(h*k) = d*g. Consequently b*c = d*g is a point q on γ which must lie on both lines [b,c], [d,g]. But then q = [b,c] ∪ [d,g] = a lies on γ. □

The following construction of a (Z,0) triple system can be found in Fletcher R (Geometric Triple Systems, unpublished).

Theorem 7. *A (Z,0) triple system exists, and its vertices must lie on an irreducible cubic curve.*

Proof. The starting configuration S for our construction involves the nine points {0,±1,±2,±3,4,−5} and is illustrated by the bold lines in Fig. 9.6. Any nine points in the projective plane lie on some cubic curve, so let γ denote a cubic which contains these nine points. No three lines cover all the points in S so γ cannot consist of three lines. The removal of any line from S, except for [3,0,−3], leaves a configuration which still contains three collinear points. This configuration cannot be covered by any conic since no line can meet a conic in more than two points. There is a unique conic which contains the five points {±1,±2,−5}. If we select vertex 4 on line [−1,−3] so that it does not lie on this conic, then also in case we cover {3,0,−3} with a line, the remaining six points of S cannot be covered

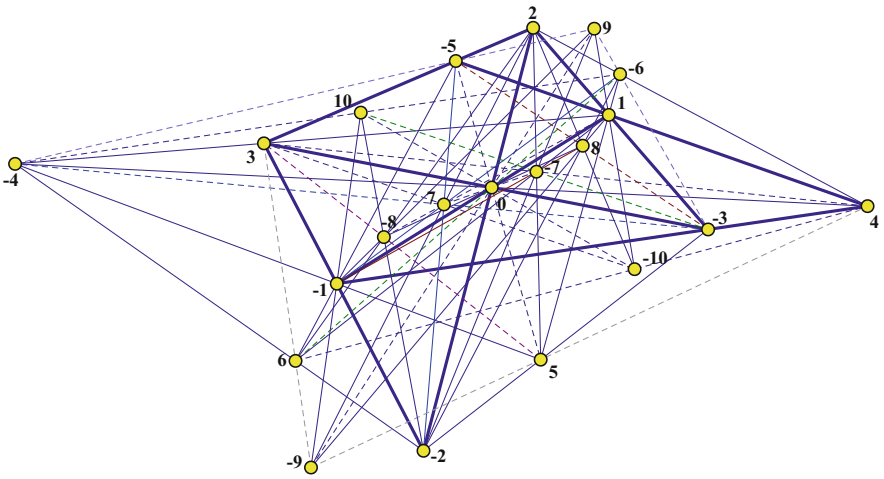


Fig. 9.6 (Z,0) triple system

by a conic. So we can suppose that γ is irreducible, and since every point in S lies on a line with two other points of S, we can also suppose that each of the nine points in S is a nonsingular point of γ. (A line through a singular point of a cubic cannot meet the cubic in two additional points.) Define $-4 = [1,3] \cup [0,4]$ and $5 = [0,-5] \cup [-2,-3]$. The 0-squares:

$$\begin{vmatrix} -4 & 1 & 3 \\ 4 & -3 & -1 \\ 0 & 2 & -2 \end{vmatrix}, \quad \begin{vmatrix} 5 & -2 & -3 \\ -5 & 3 & 2 \\ 0 & -1 & 1 \end{vmatrix}$$

imply by Lemma 6 that -4 and 5 lie on γ. Now suppose $k \geq 5$ and assume inductively that points $\{0, \pm 1, \pm 2, \pm 3, \ldots, \pm k\}$ have been defined and shown to lie on γ. Define vertices $\pm(k+1)$ by $-k-1 = [1,k] \cup [2,k-1]$ and $k+1 = [-1,-k] \cup [-2,-k+1]$. Then the 0-squares:

$$\begin{vmatrix} -k-1 & 1 & k \\ 2 & 0 & -2 \\ k-1 & -1 & -k+2 \end{vmatrix}, \quad \begin{vmatrix} k+1 & -1 & -k \\ -2 & 0 & 2 \\ -k+1 & 1 & k-2 \end{vmatrix}$$

imply, by Lemma 6, that the points $\pm(k+1)$ lie on γ. It now follows by induction that every integer has been assigned to a point in the plane which lies on γ. It remains to show that the resulting configuration is a (Z,0) triple system. Suppose $k \geq 5$, $0 < p < q < k$, and $p + q = k$. Assume inductively that if $|a|$, $|b|$, $|a+b| < k$, then we have $a*b = -a-b$. We have $k = [-1, -k+1] \cup [-2,-k+2]$ by definition and also $(0*p) = -p$ by the inductive hypothesis. Consequently:

$$\begin{aligned} k * -p &= (-1 * -k+1) * (0 * p) = (0 * -k+1) * (p * -1) \\ &= (k-1) * (-p+1) = p\text{–}k = -q. \end{aligned}$$

Note that $|-k+1|$, $|p-1|$, $|p-k|$, and $|k-p|$ are all less than k so this equation is legitimate due to the inductive hypothesis. So we have $k = -p * -q$. For negative values of k we can proceed as follows. Consider: $0*0 = (1 * -1)*(2*-2) = (1*2)*(-1 * -2) = (-3 * 3) = 0$. We will use this result to show that $-k = 0*k$ for every positive integer k. Assume inductively that this holds for all positive integers less than k. Consider: $-k = (1 * (k-1)) = (0*-1)*(0 * -k+1) = (0*0)*(-1 * (-k+1)) = 0*k$. Again suppose $0 < p < q < k$ and $p + q = k$. Now we want to show that $-k = p*q$. Consider: $-k = 0*k = (0*0)*(-p * -q) = (0*-p)*(0*-q) = p*q$. We can now conclude that $a * b = -a-b$ for all $a, b \in Z$ with a, b, $-a-b$ distinct, and thus our construction results in a (Z,0) triple system. □

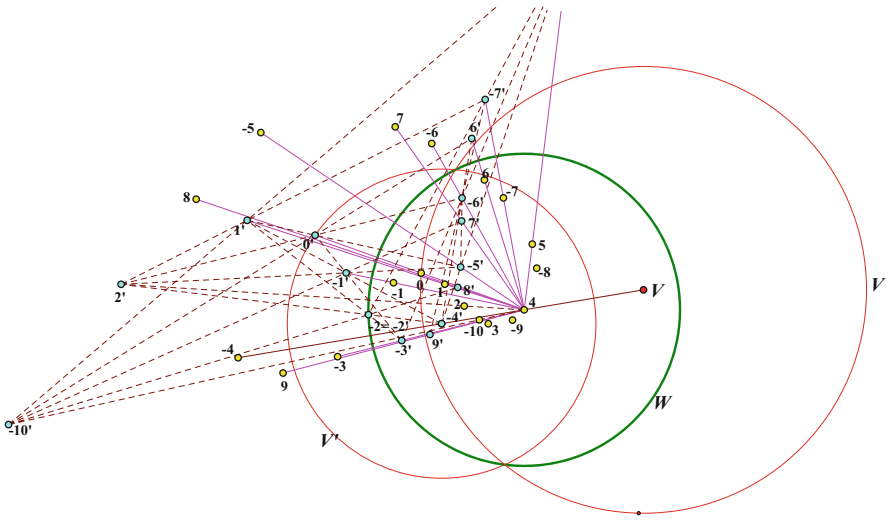


Fig. 9.7 Inverse of a (Z,0) circle system by circle w with center at vertex 4

In Fig. 9.6 we illustrate the construction of a (Z,0) triple system described in Theorem 7.

Theorem 8. *The vertices of a (Z,0) circle system lie on an algebraic curve with degree* ≤ 6 *which is the inverse of an irreducible self-inversive cubic curve.*

Proof. Let Ω be a (Z,0) circle system and let ω be a circle with center at vertex 4 of Ω. If p is a vertex of Ω, let p' denote the inverse of p wrt ω and let Ω' denote the inverse of Ω wrt ω as in Fig. 9.7. Then Ω' is a (Z,−4) triple system. By adding 1 to each point of Ω' we obtain a (Z,0) triple system. By Theorem 7, there exists an irreducible cubic curve γ which contains all the vertices of Ω'. By Theorem 1, there exists a circle υ which contains vertex 0 of Ω, and which inverts Ω onto itself. If υ' denotes the circle which is the inverse of υ wrt ω, then υ' inverts Ω' onto itself and it inverts the cubic envelope γ onto itself. Thus γ is an irreducible self-inversive cubic curve. If γ' is the inverse of γ wrt ω, then γ' contains all the vertices of Ω and, as it is the inverse of a cubic, its degree cannot exceed 6. Note that υ' inverts all pairs of points of the form $\{q', -q'\}$, and since Ω' is a (Z,−4) triple system, all the lines $[q', -q']$ contain the point $-4'$. Thus $-4'$ is the center of υ'. □

9.5 Ternary Hypercommutativity

We now describe an algebraic system which is an analog of the above mentioned binary hypercommutative variety, but which involves a ternary operation. Let S be a set and δ a ternary operation defined on S. If the axioms:

(i) δ(a,b,c) is invariant under any permutation of the three variables;
(ii) δ(a, b, δ(a,b,c)) = c; and
(iii) δ(δ(a,b,c), δ(d,e,f), δ(g,h,k)) is invariant under any permutation of the nine variables;

are satisfied for all elements a,b,c, ..., k of S, then (S,δ) is a *ternary hypercommutative algebra (THA).* Axiom (i) is *ternary commutativity;* axiom (ii) is *ternary absorption,* and axiom (iii) is *ternary hypercommutativity.* The collection of all algebras satisfying these three axioms comprise the *ternary hypercommutative variety*. Let G be an abelian group and g a fixed element of G. If we define δ: G → G by δ(a,b,c) = g − (a + b + c), then (G,δ) is a member of this variety. If γ is a nonsingular irreducible self-inversive cubic curve, then any circle υ which meets γ in three points must also meet γ in a fourth point. This is so since the equation for υ can be solved for $x^2 + y^2$ to obtain $x^2 + y^2 = sx + ty + r$ and this linear expression can be substituted for $x^2 + y^2$ in Eq. (9.5) to obtain

$$(ax + by)(sx + ty + r + 1) + \left(ex^2 + fxy + gy^2\right) = 0. \tag{9.10}$$

In (9.10) y^2 can be replaced by $-x^2 + sx + ty + r$ leaving an equation which we can solve for y to obtain

$$y = \frac{Ax^2 + Bx + C}{Dx + E}. \tag{9.11}$$

for some real numbers A,B,C,D,E. Now using (9.11) to substitute for y in the equation of the circle υ, and clearing fractions, we obtain P(x) = 0 where P is a degree 4 polynomial in x. Since we are assuming that υ meets γ three times, counting multiplicities, P(x) = 0 has three real solutions. Since complex solutions must occur in conjugate pairs, it must be that P(x) = 0 has a fourth real solution. So, on γ we can define a ternary operation by setting δ(a,b,c) equal to the unique fourth point on γ which also lies on circle (a,b,c). It is our intention to show that the resulting algebra (γ,δ) is a THA. Here δ(a,a,b) denotes the fourth point on the circle which is tangent to γ at a and contains the point b of γ, and δ(a,a,a) denotes the fourth point on γ and on the circle which meets γ at a with multiplicity 3. This is the *osculating circle at point a.* We will show that γ contains *δ-idempotent elements,* i.e., points x which satisfy δ(x,x,x) = x. The osculating circle at such a point x intersects γ with multiplicity 4 at x. A similar ternary operation δ can be defined on the points of a noncircular conic α by setting δ(a,b,c) equal to the fourth point on circle (a,b,c) and on α. The resulting algebra (α,δ) is shown to be a THA in [3].

Let γ be an irreducible, nonsingular, self-inversive cubic curve. We can suppose that γ inverts onto itself via the unit circle ω. In accordance with [2], γ may be one of two types: Type 1 consists of those irreducible nonsingular cubics which can be transformed into a cubic with equation: $y^2 = x(x^2 + kx + 1)$ where $-2 < k < 2$, and Type 2 cubics are those which can be transformed to $y^2 = x(x - 1)(x - w)$ where $w > 1$. Type 1 cubics consist of a single connected component, and Type 2 cubics

have two disjoint connected components, the *bell* and the *oval*, when regarded in the projective plane. The bell of a Type 2 cubic is a subalgebra of the binary hypercommutative algebra (γ,*), and the oval is an *anticlosed* subset of γ, i.e., if a,b $\in$ oval, then a*b $\in$ bell. If x $\in \gamma$, then $\sqrt{x}$ denotes the set $\{a \in \gamma: a*a = x\}$. If γ is Type 1, then $\sqrt{x}$ contains exactly two elements for each x $\in \gamma$. If γ is Type 2, then $\sqrt{x}$ is empty if x $\in$ oval, and $\sqrt{x}$ has exactly four elements if x $\in$ bell, two of these on the oval and two on the bell. If a*a = a, then a is called a *flex*; γ has three collinear flexes. A Type 1 cubic has exactly one point at infinity, and a Type 2 cubic can have as many as three points at infinity, but if γ is self-inversive, we have seen that γ contains a unique point (∞) at infinity. We start by considering a Type 1 cubic.

Theorem 9. *Let γ be a nonsingular irreducible cubic curve of Type 1 which inverts onto itself via a circle ω. Define a ternary operation δ on the points of γ, by setting δ(a,b,c) equal to the fourth point on γ and on the circle (a,b,c). In case a,b,c are collinear, set δ(a,b,c) = ∞. Then (γ,δ) is a THA.*

Proof. Let τ be a circle with center T not on γ, and which is orthogonal to ω. Let γ' denote the inverse of γ wrt τ, as illustrated in Fig. 9.8. The curve γ' is a bounded algebraic curve of degree 4 which inverts onto itself via ω. We will construct a (Z,0) circle system Ω on γ' by following the prescription given in Theorem 4. Let P be a point of contact of ω and γ' and let P be vertex 0 of Ω. Choose vertex 1 on γ' close to P, and let vertex -1 be the inverse of vertex 1 wrt ω. Choose vertex 2 on γ' close to vertex 1, and so vertices 0, 1, 2 occur in order along γ', and let vertex -2 be the inverse of vertex 2 wrt ω. Put vertex 3 at the intersection of circle (0,-1,-2) and γ', and let vertex -3 be the inverse of vertex 3 wrt ω. Let vertex 4 be the unlabeled intersection of circles (0,-1,-3), (1,-2,-3), and shift vertex 2 on γ' until vertex 4 lies on γ'. Let vertex -4 be the inverse of vertex 4 wrt ω, and follow the prescription given in Theorem 4 to complete the construction of Ω. By Theorem 8 we know that the vertices of Ω lie on an algebraic curve ξ which is the inverse of a self-inversive cubic curve ξ_0. We know that ξ has the nine points $\{0, \pm1, \pm2, \pm3, \pm4\}$ in common with γ', and ξ_0 has nine points in common with γ, namely the inverses of $\{0, \pm1, \pm2, \pm3, \pm4\}$ wrt τ. Since Ω inverts onto itself via ω, we know that ξ_0 also inverts onto itself wrt ω. We may suppose that ω is the unit circle and thus both ξ_0 and γ have equations of the form given in Eq. (9.5). Both curves contain the origin, and by the self-inversive nature of (9.5), if q is a point on either curve, then $-$q, the inverse of q wrt ω, will automatically satisfy (9.5). But we shall argue that the points {1,2,3,4}, or rather their inverses wrt τ, completely determine an irreducible cubic which self-inverts through the unit circle. From (9.5) we see that we cannot have both coefficients b, g equal to zero since otherwise x would be a factor and (9.5) would not represent an irreducible cubic. If $b \neq 0$, then we can divide (9.5) by b to obtain an equivalent equation with only four coefficients. Substituting for x and y the coordinates for points 1,2,3,4 we obtain four homogeneous linear equations in the four remaining coefficients, and thus the curve ξ_0 is completely determined, and must be identical with γ. It then follows that $\xi = \gamma'$, so all the points of Ω lie on γ'.

The vertices of Ω form a THA (Ω,δ) with δ(a,b,c) = $-$(a + b + c), which occurs as a subset of γ'. We want to show that this algebra can be extended to the

entire curve γ'. For this purpose we adopt the following analytical argument. For $a,b,c \in \gamma'$, we define $\delta(a,b,c)$ to be the fourth point on γ' and on circle (a,b,c). Since γ' is the inverse of the self-inversive cubic for which the analogously defined ternary operation is well defined, δ is well defined and identities (i) and (ii) of the ternary hypercommutative variety are satisfied. We need only to demonstrate hypercommutativity. For the sake of a little simplification, we observe that the hypercommutative axiom is a consequence (in the presence of axioms (i) and (ii)) of the seven variable identity:

$$\delta\left(\delta\left(a,b,c\right),\delta\left(d,e,f\right),q\right) = \delta\left(\delta\left(a,b,d\right),\delta\left(c,e,f\right),q\right). \tag{9.12}$$

Define a function f: $\gamma' \to \gamma'$ by $f(x,y,z,u,v,w,q) = \delta(\delta(x,y,z),\ \delta(u,v,w),\ q)$. This function is continuous in each of its seven variables, and since the points of γ' form a compact set, it is uniformly continuous. So, given $\varepsilon > 0$, there exists a positive real number η such that if points $\{x_0, y_0, z_0, u_0, v_0, w_0, q_0\}$ lie on γ', and $|x - x_0| < \eta$, $|y - y_0| < \eta$, $\ldots, |q - q_0| < \eta$, then $|f(x,y,z,u,v,w,q) - f(x_0,y_0,z_0,u_0,v_0,w_0,q_0)| < \varepsilon/2$, and $|f(x,y,u,z,v,w,q) - f(x_0,y_0,u_0,z_0,v_0,w_0,q_0)| < \varepsilon/2$. Now in our construction of Fig. 9.8, by shifting vertex 1 (and then adjusting vertex 2 suitably) we can make vertex 20 and vertex -20 coincide at the second contact point M of ω and γ'. When this occurs, we obtain a $(Z_{40}, 0)$ circle system on γ'. By moving vertex 1 closer and closer to P we can make vertex n and vertex $-n$ coincide at point M for arbitrarily large n, and thus construct a $(Z_{2n}, 0)$ circle system Ω_{2n} on γ'. Let Δ_{2n} denote the maximum distance (in the usual Euclidean metric) between consecutive points of Ω_{2n}. By choosing n sufficiently large, we can make $\Delta_{2n} < \eta$. Thus if a,b,c,d,e,f,q are any seven points on γ', we can find points $a_0,b_0,c_0,d_0,e_0,f_0,q_0$ of Ω_{2n} such that $|a - a_0| < \eta$, $|b - b_0| < \eta$, $\ldots$,,, $|q - q_0| < \eta$. Then we have $|f(a,b,c,d,e,f,q) - f(a_0,b_0,c_0,d_0,e_0,f_0,q_0)| < \varepsilon/2$, and

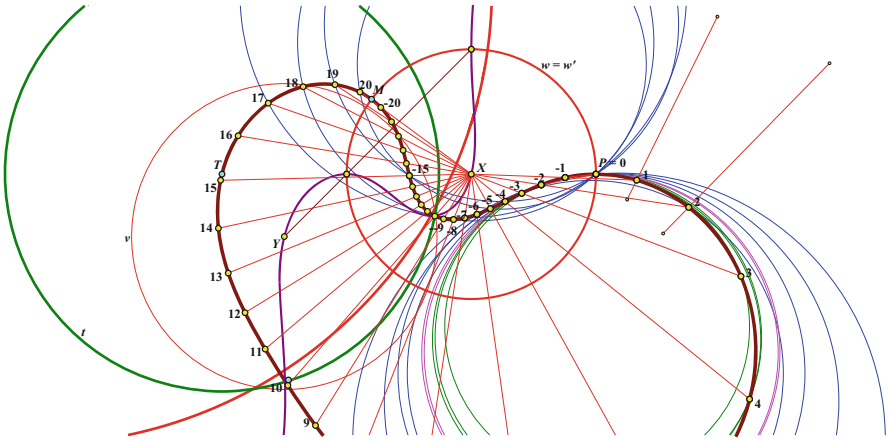


Fig. 9.8 Inverse of self-inversive cubic

$|f(a,b,d,c,e,f,q) - f(a_0,b_0,d_0,c_0,e_0,f_0,q_0)| < \varepsilon/2$. Since (Ω_{2n},δ) is a THA, we have $f(a_0,b_0,c_0,d_0,e_0,f_0,q_0) = f(a_0,b_0,d_0,c_0,e_0,f_0,q_0) = m_0$ for some point m_0 on γ'. Then

$$\begin{aligned}
&|f(a,b,c,d,e,f,q) - f(a,b,d,c,e,f,q)| \\
&= |f(a,b,c,d,e,f,q) - m_0 + m_0 - f(a,b,d,c,e,f,q)| \\
&\leq |f(a,b,c,d,e,f,q) - m_0| + |m_0 - f(a,b,d,c,e,f,q)| < \varepsilon/2 + \varepsilon/2 = \varepsilon.
\end{aligned}$$

Since $\varepsilon > 0$ is arbitrary, we obtain $f(a,b,c,d,e,f,q) = f(a,b,d,c,e,f,q)$, and thus identity (9.12) holds for all points a,b,c,d,e,f,q on γ'. Since inversion maps circles to circles, the same identity holds for points on γ, and thus (γ,δ) is a THA. □

In case γ is a Type 2 self-inversive cubic curve, we also obtain that (γ,δ) is a THA. There is only a slight difference in the way a (Z,0) circle system is constructed on γ.

Theorem 10. *Let γ be a Type 2 cubic curve which inverts onto itself via circle ω which has nonempty intersection with γ. Define a ternary operation δ on the points of γ, by setting $\delta(a,b,c)$ equal to the fourth point on γ and on the circle (a,b,c). In case a,b,c are collinear, set $\delta(a,b,c) = \infty$. Then (γ,δ) is a THA.*

Proof. Let τ be a circle with center T not on γ, and which is orthogonal to ω. Then the inverse γ' of γ wrt τ also inverts onto itself wrt ω. Let p be a point of contact of ω and γ and suppose p lies on the oval as in Fig. 9.9. Then ω must invert the oval onto itself and the bell onto itself, and so must also meet the bell at a point q. Similarly, if we assume that p lies on the bell, then ω must meet the bell at a point q. So we can assume, without loss of generality, that ω meets the oval at p and the bell at q. Let p' denote the inverse of p wrt τ, and let q' denote the inverse of q wrt τ. Note that γ' consists of two bounded components, the inverse $(\text{oval})'$ of the oval, and the inverse $(\text{bell})'$ of the bell wrt τ. The point p' lies on $(\text{oval})'$ and q' lies on $(\text{bell})'$. Now we construct a (Z,0) circle system Ω on γ' by first choosing q' as vertex 0, and then putting vertex 1 on $(\text{oval})'$ near p'. Let vertex -1 be the inverse of vertex 1 wrt ω, and put vertex 2 on $(\text{bell})'$ near vertex 0. Now the construction of Ω on γ' proceeds as in Theorem 9. All the even vertices of Ω will lie on $(\text{bell})'$ and all the odd vertices of Ω will lie on $(\text{oval})'$. The remainder of the proof is identical with Theorem 9. □

In the next section we will show that every Type 2 self-inversive cubic curve γ actually inverts onto itself through three mutually orthogonal circles, two of which have nonempty intersection with γ. So Theorem 10 actually applies to every Type 2 self-inversive cubic.

9.6 Self-Inversion Through Orthogonal Circles

This section contains our intended main geometric results concerning a self-inversive cubic curve γ. We start by developing a simple formula which relates the ternary operation δ and the binary operation *, defined in Theorems 5 and 9 on the points of γ.

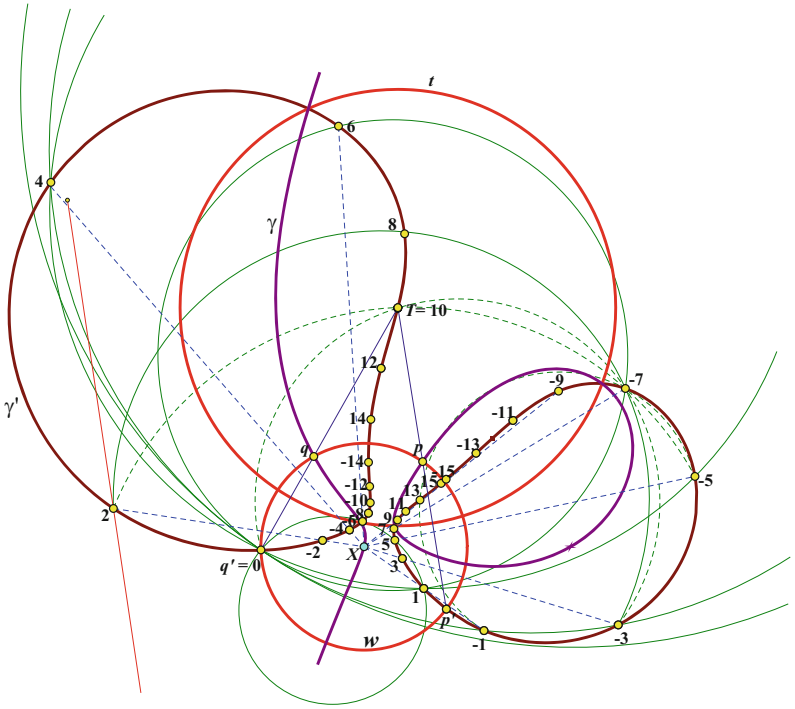


Fig. 9.9 (Z,0) circle system on Type 2 cubic

Theorem 11. *Let γ be a nonsingular irreducible cubic curve which inverts onto itself through a circle ω with center X. Then δ(a,b,c) = X*{(X*a)*(b*c)} holds for all points a,b,c on γ.*

Proof. We have X*X = ∞, and if a,b,c are three collinear points on γ, we define δ(a,b,c) = δ(a,b,a*b) = ∞. This is a natural definition, since if we move c along γ closer and closer to a*b, then the radius of the circle (a,b,c) gets large without bound and the point δ(a,b,c) moves towards the point (∞) at infinity on γ. We then obtain immediately: δ(a,b,∞) = a*b for all points a,b on γ. In particular: δ(a,X,∞) = a*X; δ(b,c,∞) = b*c, and

$$(\mathrm{X} * \mathrm{a}) * (\mathrm{b} * \mathrm{c}) = \delta(\mathrm{a}, \mathrm{X}, \infty) * \delta(\mathrm{b}, \mathrm{c}, \infty) = \delta(\delta(\mathrm{a}, \mathrm{X}, \infty), \delta(\mathrm{b}, \mathrm{c}, \infty), \infty). \tag{9.13}$$

Now by applying identity (9.12) we obtain

$$\begin{aligned}(\mathrm{X} * \mathrm{a}) * (\mathrm{b} * \mathrm{c}) &= \delta(\delta(\mathrm{a}, \mathrm{b}, \mathrm{c}), \delta(\mathrm{X}, \infty, \infty), \infty) \\ &= \delta(\delta(\mathrm{a}, \mathrm{b}, \mathrm{c}), \mathrm{X} * \infty, \infty) = \delta(\delta(\mathrm{a}, \mathrm{b}, \mathrm{c}), \mathrm{X}, \infty). \end{aligned} \tag{9.14}$$

But then we obtain $(X*a)*(b*c) = X*\delta(a,b,c)$, and thus $\delta(a,b,c) = X*\{(X*a)*(b*c)\}$ as desired. □

If γ is a nonsingular irreducible curve with flex f, then we can define an addition on the points of γ by $p + q = (p*q)*f$. It is well known (see [2]) that $(+)$ is an associative binary operation and that $(\gamma,+)$ is an abelian group G_γ with f as identity element and $-p = f*p$. In terms of the group operation $p*q = -p - q$. The representation in the following lemma of $\delta(a,b,c)$ in terms of the group operation will simplify some of the following results.

Lemma 12. *Let γ be a nonsingular irreducible cubic curve which inverts onto itself through a circle ω with center X. Then in terms of the group operation in G_γ we have $\delta(a,b,c) = \infty - a - b - c$ for any points a,b,c of γ, where ∞ denotes the point at infinity on γ.*

Proof. Let f be a flex of γ. Consider: $\infty + (-a) + (-b) + (-c) = \{(\infty * -a)*f\} + \{(-b * -c)*f\} = [\{(\infty * -a)*f\} * \{(-b * -c)*f\}]*f = [\{(\infty * -a)*(-b * -c)\}*(f*f)]*f = (\infty * -a)*(-b * -c) = \{(X*X)*(a*f)\} * \{(b*f)*(c*f)\} = \{(X*a)*(X*f)\} * \{(b*c)*f\} = \{(X*f)*f\}*\{(X*a)*(b*c)\} = X * \{(X*a)*(b*c)\} = \delta(a,b,c)$. □

Theorem 13. *Let γ be a Type 1 cubic curve which inverts onto itself through a circle ω with center X, and let (γ,δ) denote the THA defined on the points of γ. Then there exists a second circle υ, orthogonal to ω, which also inverts γ onto itself.*

Proof. Let p,q denote the two points of contact of γ and ω, and let $Y = p*q$ as in Fig. 9.10. Note that since $p*p = q*q = X$, the tangents to γ at p, q contain X. If Y were to lie on segment [p,q] in the interior of ω, then γ would have to cross this segment at a second point in the interior of γ. But then the line [p,q] would intersect the cubic γ in more than three points, which is impossible. So Y must lie on γ and in the exterior of circle ω. As a consequence we can find a circle υ with center Y which is orthogonal to ω. Note that $Y*Y = (p*q)*(p*q) = (p*p)*(q*q) = X*X = \infty$, so we could also define Y as the second element, besides X, in $\sqrt{\infty}$. Now let a be any point on γ different from ∞ and consider circle $\eta = (a,p,q)$. Since Y, p, q are collinear and p,q lie on ω, and υ is orthogonal to ω, it must be that p,q are inverses wrt υ, and thus η is orthogonal to υ. From Theorem 11 we have $\delta(a,p,q) = X*\{(X*a)*(p*q)\} = X*\{(X*a)*Y\} = X*\{(X*a)*(Y*\infty)\} = X*\{(X*\infty)*(Y*a)\} = X*\{X*(Y*a)\} = Y*a$. Since η is orthogonal to υ, and a, Y*a lie on η, it follows that Y*a is the inverse of a wrt υ. As a was chosen arbitrarily on γ, it follows that circle υ inverts γ onto itself. □

Proofs of the following three lemmas can be found in [4].

Lemma 14. *Let p,q,r,s be four points in the Euclidean plane such that {p,q} and {r,s} are inverse pairs wrt circle η, and {p,s}, {q,r} are inverse pairs wrt circle υ. Then $\omega = (p,q,r,s)$ is a circle. Circles η,υ,ω are mutually orthogonal, and $W = [p,r] \cap [q,s]$ is the radical center of η,υ,ω.* □

Lemma 15. *Let ω,υ be orthogonal circles with centers X, Y, respectively, and let ϕ_X, ϕ_Y, denote inversion mappings wrt ω,υ, respectively. Then $\phi_X\phi_Y = \phi_X\phi_Y$.* □

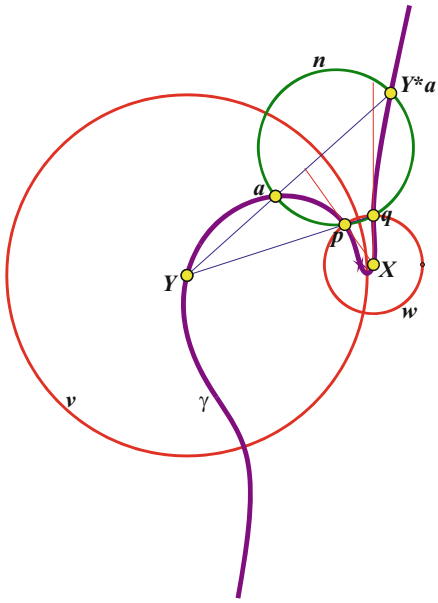


Fig. 9.10 Orthogonal circles which invert Type 1 cubic γ onto themselves

If ξ is a circle with center W, and p is a point in the plane different from W, then we define the *antiinverse* $p^{\#}$ *of p wrt* ξ to be the reflection of the inverse p′ of p across the center W of ξ. If r is the radius of ξ, then $[p,W][p^{\#},W] = r^2$ similar to the defining formula for the inverse p′ but now W lies in the interior of the segment $[p, p^{\#}]$.

Lemma 16. *Let ω,υ,η be three mutually orthogonal circles with centers X,Y,Z, respectively, and let* ϕ_X, ϕ_Y, ϕ_Z *denote the corresponding inversion mappings. Let W denote the radical center of ω,υ,η. Then there exists a circle ξ with center W which antiinverts each of the circles ω,υ,η onto themselves. If* $\phi_W(p)$ *denotes the antiinverse of a point p wrt ξ, then* $\phi_W = \phi_X\phi_Y\phi_Z$. □

Theorem 17. *Let γ be a Type 2 cubic curve which inverts onto itself via a circle ω. Then there exist circles υ,η which also invert γ onto themselves; the circles γ,υ,η are mutually orthogonal, and there exists a fourth circle ξ which antiinverts γ onto itself.*

Proof. First we assume that ω has nonempty intersection with γ. Then the center X of ω must lie on the bell, and we may suppose, as in Fig. 9.11, that $\omega \cap \gamma = \{p,q,r,s\}$ where p,s lie on the bell and q,r lie on the oval. Then we have $p^*p = q^*q = r^*r = s^*s = X$, and $X^*X = \infty$. The points p*s, q*r must both lie on the bell in the exterior of ω and both square to ∞. The bell contains exactly two points which square to ∞, one of which is X. Since X lies in the interior of ω, it must be that p*s and q*r both equal the remaining point on the bell which squares to ∞. Let Y represent this point, so we have $p^*s = q^*r = Y$, and $Y^*Y = \infty$. Let υ denote the circle with center Y which is orthogonal to ω. We can then argue, as

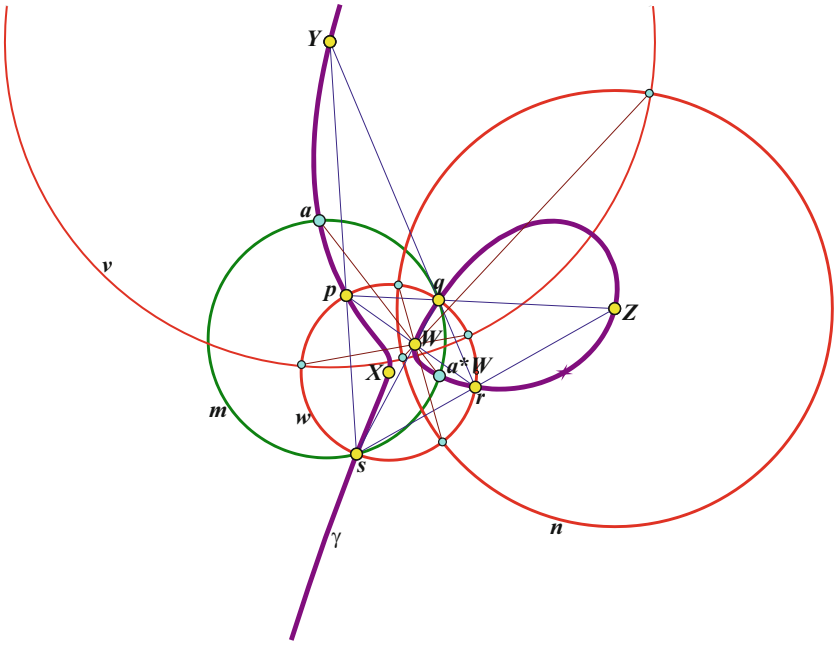


Fig. 9.11 Three mutually orthogonal circles invert self-inversive Type 2 cubic onto itself

in the previous theorem, that υ inverts γ onto itself. The points p*q, s*r, s*q, p*r must lie on the oval and in the section of the plane which lies between the tangent lines [X,q], [X,r], and consequently p*q, s*r must lie in the exterior of ω, and s*q, p*r must lie in the interior of ω. All four of these points square to ∞, but there are exactly two points, say, Z, W on the oval which lie in $\sqrt{\infty}$. If Z lies in the exterior of ω, and W in the interior of ω, then we must have p*q = s*r = Z and p*r = s*q = W. Let η be the circle with center Z which is orthogonal to ω. Again we can argue as in Lemma 12 that η inverts γ onto itself. It follows from Lemma 14 that circles η,υ are orthogonal and that W is the radical center of ω,υ,η. By Lemma 16, there exists a circle ξ with center W which antiinverts each of the mutually orthogonal circles ω,υ,η onto themselves. It remains to show that ξ antiinverts the cubic curve γ onto itself. Let a be a point on γ different from W. Consider: δ(a,q,s)=X*{(X*a)*(s*q)} = (X*∞)*{(X*a)*W} = {X*(X*a)}*(W*∞) = a*W, so the points a,s,q,W*a lie on a circle μ. Segments [s,q], [a, a*W] meet at point W in the interior of μ. Since ξ antiinverts ω onto itself, and s,q are points of ω with s,q,W collinear, the points s,q form an antiinverse pair wrt ξ. It follows that ξ antiinverts circle μ onto itself, when we can conclude that a, a*W form an antiinverse pair wrt ξ.

Now suppose that ω ∩ γ is empty as in Fig. 9.12, then the center X of ω lies on the oval and ω inverts the oval onto the bell and the bell onto the oval. Let Y be one of the points on the bell which satisfies Y*Y = ∞, and let

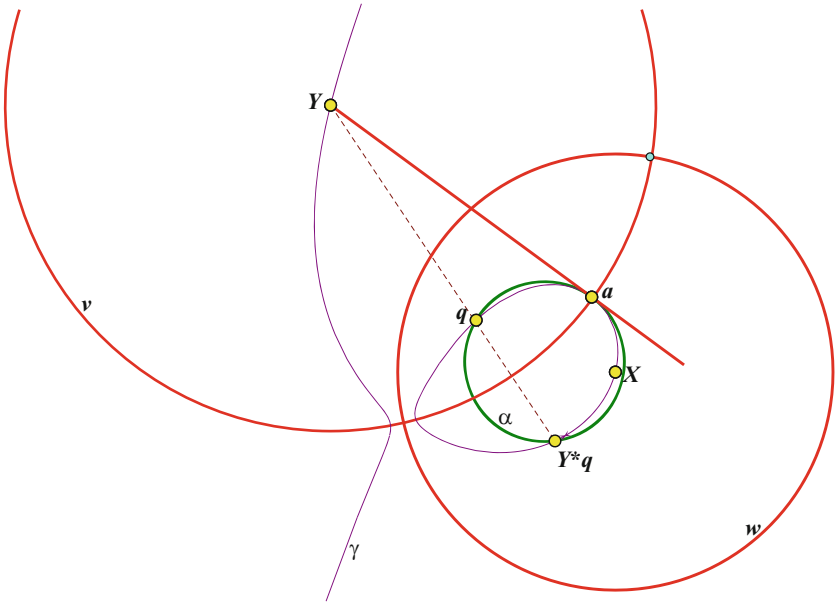


Fig. 9.12 Type 2 cubic γ which inverts onto itself via a circle w which has empty intersection with γ

a be a point on the oval which satisfies a*a = Y. If q is any point on γ, then δ(q,a,a) = X*{(X*q)*(a*a)} = X*{(X*q)*Y} = (X*∞)*{(X*q)*Y} = {X*(X*q)}*(Y*∞) = q*Y. Thus the circle α through q and tangent to γ at a also contains the point q*Y. Let υ denote the circle with center Y and point a. Since a*a = Y, the line [a,Y] is tangent to γ at a, and also tangent to α at a. As a consequence circles α,υ are orthogonal, and it must be that the points q, Y*q are inverses wrt υ. Since q was chosen arbitrarily on γ, it follows that υ inverts the entire curve γ onto itself. So there always exists a circle which inverts γ onto itself, and which has nonempty intersection with γ, and thus the proof is complete by the first paragraph. □

9.7 Automorphisms

Let γ be an irreducible nonsingular cubic curve which inverts onto itself through a circle ω with center X, and let (γ,δ) denote the ternary hypercommutative algebra defined on the points of γ. A *translate of* γ is a mapping ϕ_P: γ → γ defined by $\phi_P(x) = P*x$ where P is a fixed point of γ. For each P ∈ γ, the mapping ϕ_P is a bijection. Consider: $\phi_X(\delta(a,b,c)) = X*\{X*((X*a)*(b*c))\} = (X*a)*(b*c)$, and $\delta(\phi_X(a), \phi_X(b), \phi_X(c)) = \delta(X*a, X*b, X*c) = X*\{(X*(X*a))*((X*b)*(X*c))\} = X*\{a*((X*X)*(b*c))\} = \{X*(X*X)\}*\{a*((X*X)*(b*c))\} = (X*a)*(b*c)$. So we see

that for each point X on γ which serves as the center of a circle which inverts γ onto itself, the translate ϕ_X is an automorphism of (γ,δ). In our next result we determine all translates of γ which are automorphisms of (γ,δ).

Theorem 18. *Let γ be an irreducible nonsingular cubic curve which inverts onto itself via a circle with center X. Then the translate ϕ_P is an automorphism of (γ,δ) iff $P \in \sqrt{\sqrt{\infty * \infty}}$.*

Proof. Consider: ϕ_P is an automorphism of (γ,δ) iff $\phi_P(\delta(a,b,c)) = \delta(\phi_P(a), \phi_P(b), \phi_P(c))$ for any points a,b,c of γ. This holds iff $P * \delta(a,b,c) = \delta(P*a, P*b, P*c)$. By Lemma 12 this is equivalent to $-P-(\infty-a-b-c) = \infty-(-P-a)-(-P-b)-(-P-c)$, i.e., $4P = -\infty-\infty$. This is equivalent, in turn, to $(P*P)*(P*P) = \infty*\infty$, i.e., $P \in \sqrt{\sqrt{\infty * \infty}}$. □

If γ is Type 1, then $\sqrt{\infty * \infty} = \{\infty, X*Y\}$ where X are the centers of the pair of orthogonal circles which invert γ onto themselves. So $\sqrt{\sqrt{\infty * \infty}}$ is the union of the sets $\sqrt{\infty}$, $\sqrt{X * Y} = \{X,Y\}$, $\sqrt{X * Y}$. If $\sqrt{X} = \{p,q\}$ and $\sqrt{Y} = \{s,r\}$ as in Fig. 9.13, then $p*q = Y$ and $r*s = X$. Consider: $p*s = (q*Y)*(r*X) = (q*X)*(r*Y) = q*r$, and similarly $s*q = p*r$. Let $U = p*s = q*r$, and let $V = s*q = p*r$. Then $U*U = (p*s)*(p*s) = (p*p)*(s*s) = X*Y$, and $V*V = (p*r)*(p*r) = (p*p)*(r*r) = X*Y$. Thus $\sqrt{X * Y} = \{V,U\}$, and $\sqrt{\sqrt{\infty * \infty}} = \{X,Y,U,V\}$. These points and their relationships are illustrated in Fig. 9.13. Now suppose γ is Type 2 as in Fig. 9.14. The equation $(P*P)*(P*P) = \infty*\infty$ implies $P*P \in \{\infty, X*Y\}$, where X, Y are the centers of orthogonal circles ω, υ, respectively, which invert γ onto themselves, and which lie on the bell. Then P lies in the union of the sets $\sqrt{\infty}$, $\sqrt{X * Y} = \{X,Y,Z,W\}$, $\sqrt{X * Y}$. In Fig. 9.14 we have $\sqrt{X} = \{p,q,r,s\}$ with $p,s \in$ bell and $q,r \in$ oval and $p*q = s*r = Z$; $s*p = q*r = Y$, and $p*r = s*q = W$. Also in Fig. 9.14 we have $\sqrt{Y} = \{g,h,k,j\}$ with g,j on the bell and h,k on the oval and $h*k = g*j = X$. Consider: $g*q = (X*j)*(Y*r) = (X*r)*(Y*j) = r*j = (Z*s)*(Y*j) = (Y*s)*(Z*j) = p*k = (X*p)*(Y*k) = (X*k)*(Y*p) = h*s$, and similarly $k*s = h*p = j*q = g*r$; $k*r = h*q = g*s = j*p$, and $g*p = j*s = h*r = k*q$. If we set $R = k*p$; $S = k*s$; $T = k*r$, and $U = k*q$, then $\sqrt{X * Y} = \{R,S,T,U\}$. These points and their relationships are illustrated in Fig. 9.14.

Theorem 19. *Let γ be a nonsingular irreducible self-inversive cubic curve and let A_γ denote the group of automorphisms of (γ,δ) generated by the set of translate automorphisms. Then A_γ is isomorphic to the octic group D_4 if γ is Type 1, and A_γ is isomorphic to the group $Z_2 \times D_4$ if γ is Type 2.*

Proof. First suppose γ is Type 1. The conditions: $|a| = 4$; $|b| = 2$, and $aba = b$ completely determine the dihedral group D_4, also known as the octic group. In the notation of Theorem 18, let $a = \phi_X\phi_U$, and $b = \phi_U$. Then we have immediately: $b^2 = I$, and $aba = \phi_X\phi_U\phi_U\phi_X\phi_U = \phi_U = b$, so it remains only to show that $a^4 = I$,

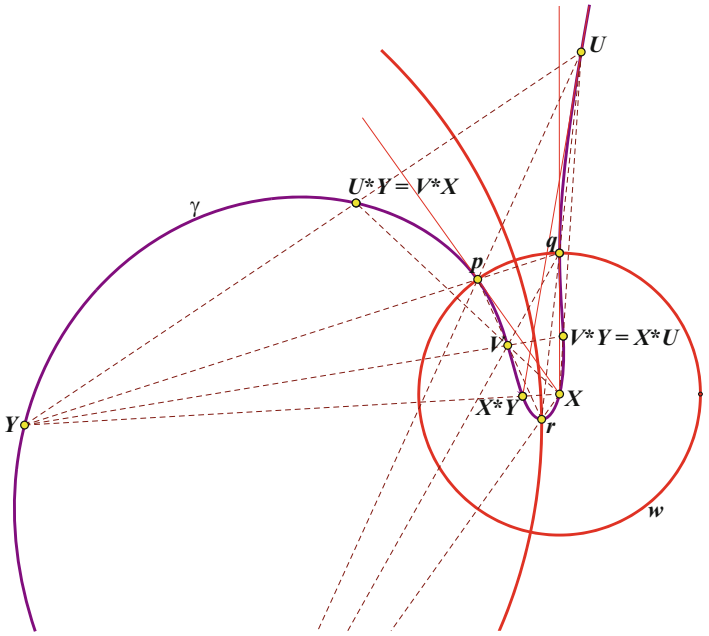


Fig. 9.13 Centers of translate automorphisms on Type 1 cubic

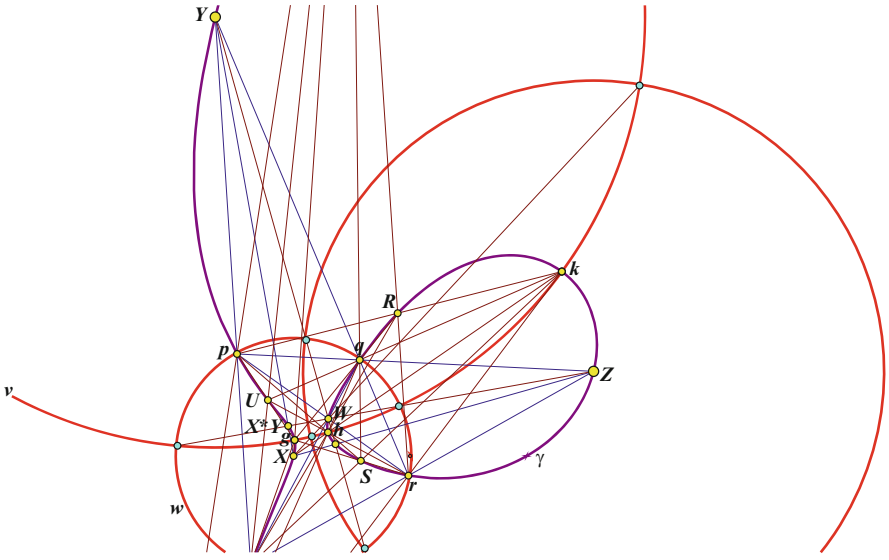


Fig. 9.14 Centers of translate automorphisms on Type 2 cubic

i.e., that $(\phi_X\phi_U)^4 = i$. Since $(\phi_X\phi_Y)^2 = i$, it suffices to show $(\phi_X\phi_U)^2 = \phi_X\phi_Y$. Consider:

$$(\phi_X\phi_U)^2 = \phi_X\phi_Y \leftrightarrow \phi_U\phi_X\phi_U = \phi_Y \leftrightarrow \phi_X\phi_U = \phi_U\phi_Y \leftrightarrow \phi_X\phi_U\,(a) = \phi_U\phi_Y\,(a)$$
$$\leftrightarrow X * (U * a) = U * (Y * a) \leftrightarrow \{s * (X * s)\} * \Big\{(p * s) * a\Big)\Big\} = (p * s)$$
$$* (Y * a) \leftrightarrow (s * a) * \{(X * s) * (p * s)\} = (s * a) * (p * Y) \leftrightarrow (X * s) * (p * s)$$
$$= p * Y \leftrightarrow (X * p) * (s * s) = p * Y.$$

Since the last equation is true, the proof is complete in case γ is Type 1. Now suppose γ is Type 2. In the notation of the previous theorem A_γ is generated by the translates $\{\phi_X, \phi_Y, \phi_Z, \phi_W, \phi_R, \phi_S, \phi_T, \phi_U\}$. By Lemmas 15 and 16, the eight elements: $\{i, \phi_X, \phi_Y, \phi_Z, \phi_W, \phi_X\phi_Y, \phi_X\phi_Z, \phi_X\phi_Z\}$ form commutative subgroup of A_γ, where $\phi_W = \phi_X\phi_Y\phi_Z$. The group $Z_2 \times D_4$ is generated by the three elements $\{(0,a), (1,a), (0,b)\}$ where $|(0,a)| = 4$; $|(0,b)| = 2$. If we let $A = (0,a)$; $B = (0,b)$, and $C = (1,a)$, then $Z_2 \times D_4$ is completely determined by the relations: (1) $|A| = 4$; (2) $|B| = 2$; (3) $ABA = B$; (4) $C^2 = A^2$; (5) $AC = CA$; and (6) $CAB = BCA$. An isomorphism $\Psi: Z_2 \times D_4 \to A_\gamma$ is obtained by setting $\Psi(A) = \phi_X\phi_R$; $\Psi(C) = \phi_Z\phi_R$, and $\Psi(B) = \phi_Z$. To establish this claim we need the following results:

(i) $\phi_R\phi_X = \phi_Y\phi_R$. Consider: $\phi_R\phi_X(a) = R*(X*a) = (p*k)*(X*a)$, and $\phi_Y\phi_R(a) = Y*(R*a) = ((Y*X)*X)*((p*k)*a) = \{(Y*X)*(p*k)\}*(X*a) = \{(X*p)*(Y*k)\}*(X*a) = (p*k)*(X*a)$.

(ii) $\phi_R\phi_Z = \phi_W\phi_R$. Consider: $\phi_W\phi_R(a) = W*(R*a) = \{(W*Z)*Z\}*\{(p*k)*a\} = \{(W*Z)*(p*k)\}*(a*Z) = \{((p*r)*(p*q))*(p*k)\}*(a*Z) = \{((p*p)*(r*q))*(p*k)\}*(a*Z) = \{((X*Y)*(p*k))\}*(a*Z) = ((X*p)*(Y*k))*(a*Z) = (p*k)*(a*Z) = R*(Z*a) = \phi_R\phi_Z(a)$.

Note that (i) and (ii) immediately imply

(iii) $\phi_X\phi_R = \phi_R\phi_Y$
(iv) $\phi_Z\phi_R = \phi_R\phi_W$

Similarly, we obtain the identities:

(v) $\phi_Z\phi_R = \phi_Y\phi_T$
(vi) $\phi_S\phi_Y = \phi_Y\phi_R$
(vii) $\phi_Y\phi_U = \phi_T\phi_Y$

The elements $\phi_X\phi_R$, $\phi_Z\phi_R$, ϕ_Z are generators for A_γ since

$$\phi_Z\,(\phi_Z\phi_R) = \phi_R, (\phi_X\phi_R)\,\phi_R = \phi_X, \phi_X\phi_Y\phi_Z = \phi_W,$$
$$(\phi_X\phi_R)\,(\phi_Z\phi_R)\,\phi_Z = (\phi_X\phi_R)\,(\phi_R\phi_W)\,\phi_Z = \phi_X\phi_W\phi_Z = \phi_X\phi_X\phi_Y\phi_Z\phi_Z$$
$$= \phi_Y\phi_Z\phi_Z = \phi_Y,$$

and the identities (v), (vi), and(vii) allow us to obtain ϕ_S, ϕ_T, ϕ_U. To complete the proof we must show that

$$|\phi_X \phi_R| = 4 \tag{9.15}$$

$$|\phi_Z| = 2 \tag{9.16}$$

$$(\phi_X \phi_R)\, \phi_Z\, (\phi_X \phi_R) = \phi_Z \tag{9.17}$$

$$(\phi_Z \phi_R)\, (\phi_Z \phi_R) = (\phi_X \phi_R)\, (\phi_X \phi_R) \tag{9.18}$$

$$(\phi_X \phi_R)\, (\phi_Z \phi_R) = (\phi_Z \phi_R)\, (\phi_X \phi_R) \tag{9.19}$$

$$(\phi_Z \phi_R)\, (\phi_X \phi_R)\, \phi_Z = \phi_Z\, (\phi_Z \phi_R)\, (\phi_X \phi_R) \tag{9.20}$$

To show (9.15), consider: $(\phi_X\phi_R)(\phi_X\phi_R) = (\phi_X\phi_R)(\phi_R\phi_Y)$ by (iii), which, in turn, equals $\phi_X\phi_Y$ which has order 2, and thus $|\phi_X\phi_R| = 4$. Consider: $(\phi_Z\phi_R)(\phi_Z\phi_R) = (\phi_Z\phi_R)(\phi_R\phi_W)$ by (iv), and $(\phi_Z\phi_R)(\phi_R\phi_W) = \phi_Z\phi_W = \phi_Z\phi_X\phi\phi_Y\phi_Z = \phi_X\phi_Y$, and so $\phi_Z\phi_R$ also has order 4. To show (9.17), consider: $(\phi_X\phi_R)\ \phi_Z(\phi_X\phi_R) = (\phi_R\phi_Y)\ \phi_Z(\phi_X\phi_R) = \phi_R\phi_W\phi_R = \phi_Z\phi_R\phi_R = \phi_Z$ where we have used (iii) and (iv) again. To show (9.18), consider: $(\phi_Z\phi_R)(\phi_Z\phi_R) = (\phi_Z\phi_R)(\phi_R\phi_W) = \phi_Z\ \phi_W = \phi_X\phi_Y = (\phi_X\phi_R)(\phi_R\phi_Y) = (\phi_X\phi_R)(\phi_X\phi_R)$, and to show (9.19), consider: $(\phi_X\phi_R)(\phi_Z\phi_R) = (\phi_X\phi_R)(\phi_R\phi_W) = \phi_X\phi_W = \phi_Y\phi_Z = \phi_Z\phi_Y = (\phi_Z\phi_R)(\phi_R\phi_Y) = (\phi_Z\phi_R)(\phi_X\phi_R)$. Finally, to show (9.20), consider: $(\phi_Z\phi_R)(\phi_X\phi_R)\ \phi_Z = \phi_Z\ (\phi_R\phi_X)\phi_R\phi_Z = \phi_Z(\phi_Y\phi_R)\phi_R\phi_Z$ by (i), which, in turn, equals $\phi_Z\phi_Y\phi_Z = \phi_Y$, and $\phi_Z(\phi_Z\phi_R)(\phi_X\phi_R) = \phi_R\phi_X\phi_R = \phi_Y\phi_R\phi_R = \phi_Y$. □

It is an open question as to whether A_γ comprises the full set of automorphisms of (γ,δ).

9.8 Subalgebras

The subalgebras of (γ,δ) and their realization as group circle systems is an extensive topic we shall only introduce here primarily as an application of the foregoing results. We begin by identifying the δ-idempotent elements and showing that they form a subalgebra. This is our first example of a *root subalgebra.* In general, a subalgebra of (γ,δ) of the form $\sqrt{P}$, $\sqrt{\sqrt{P}}$, …, $\sqrt[n]{P}$ where $P \in \gamma$, we call a *root subalgebra.*

Theorem 20. *Let γ be an irreducible nonsingular self-inversive cubic curve. The set I of δ-idempotent elements of (γ,δ) consists precisely of the set of points in* $\sqrt{\sqrt{\infty}}$.

Proof. Suppose $X \in \gamma$, $X*X = \infty$, and $\sqrt{X}$ is nonempty. If $p \in \sqrt{X}$, then $\delta(p,p,p) = X*\{(X*p)*(p*p)\} = X*\{p*(p*p)\} = X*p = p$, so $p \in I$. Conversely, if $\delta(p,p,p) = p$, then $X*\{(X*p)*(p*p)\} = p$ implies $(X*p)*(p*p) = X*p$, $p*p = (X*p)*(X*p) = (X*X)*(p*p) = \infty*(p*p)$, and thus $(p*p)*(p*p) = \infty$, i.e., $p \in \sqrt{\sqrt{\infty}}$. □

If S is a subalgebra of (γ,δ) and there exists an abelian group G; an element $g \in G$, and a bijection ϕ: $S \to G$ such that the points in S form a (G,g) circle system under the labeling determined by ϕ, then we say that S can be *realized as a (G,g) circle system*.

Theorem 21. *Let γ be an irreducible nonsingular self-inversive cubic curve. The set I of δ-idempotent elements of (γ,δ) forms a subalgebra of (γ,δ). If γ is Type 1, then I can be realized as a $(Z_2 \times Z_2, (0,0))$ circle system, and if γ is Type 2, then I can be realized as a $(Z_2 \times Z_4, (0,0))$ circle system.*

Proof. Let $p,q,r \in I$, then $\delta(p,q,r) = \delta(\delta(p,p,p),\delta(q,q,q),\delta(r,r,r)) = \delta(\delta(p,q,r),\delta(p,q,r),\delta(p,q,r))$ which implies $\delta(p,q,r) \in I$. In the last equation we have used ternary hypercommutativity. The realization of I as a $(Z_2 \times Z_4, (0,0))$ circle system on a Type 2 cubic is indicated in Fig. 9.15. □

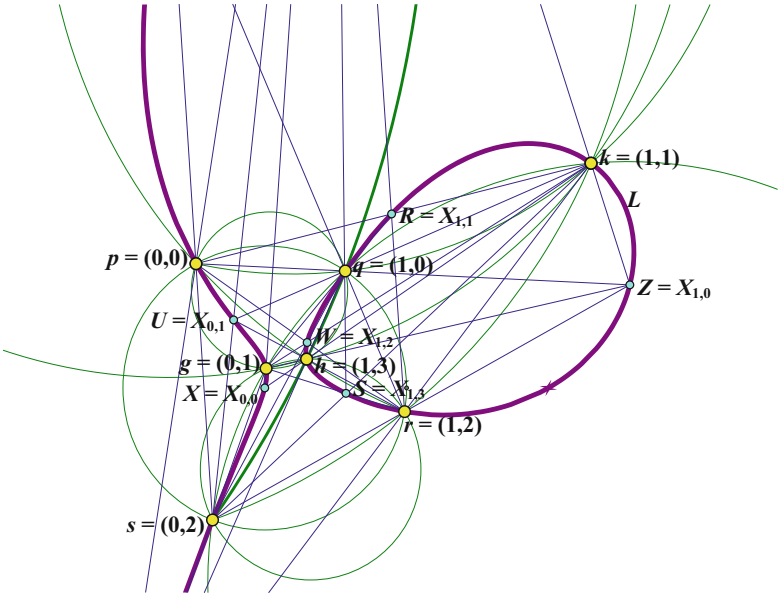


Fig. 9.15 The subalgebra I of idempotents realized as a $(Z_2 \times Z_4, (0,0))$ circle system

Let G be an abelian group and ψ: $G \to \Pi$ an injective mapping from G into the projective plane such that no four points in $\psi(G)$ are collinear. For each $g \in G$, let $W_g = \{[a,b]: a + b = g\}$. If for each g the lines in W_g are concurrent, then we call the set of points $\psi(G)$ and the associated lines, a *perfect polygon with base G,* or simply a *perfect G-gon.* The theory of perfect polygons is given in [4]. Let X_g denote the point where the lines in W_g concur, then the set of points $\{X_g: g \in G\}$ is called the set of *perspective points* of the perfect G-gon. A (G,g) circle system is a special type of perfect G-gon as we show next.

Theorem 22. *Let Ω be a (G,g) circle system with cubic envelope γ, and suppose γ is an irreducible nonsingular cubic curve which inverts onto itself via a circle with center X. Then Ω is a perfect G-gon.*

Proof. Suppose $a,b,c,d \in G$ with $a + b = c + d = q$. Then $\delta(0,a,b) = g - (a + b) = g - (c + d) = \delta(0,c,d)$. Thus, $X*\{(X*0)*(a*b)\} = X*\{(X*0)*(c*d)\}$ which implies, by cancellation, that $a*b = c*d$. If we set $a*b = c*d = X_q$, then it follows that every line in $W_q = \{[a,b]: a + b = q\}$ contains the point X_q. Consequently Ω is a perfect G-gon. □

In Fig. 9.15 we see that the points of I form a perfect $Z_2 \times Z_4$ – gon whose perspective set consists precisely of the eight points {X,Y,Z,W,R,S,T,U} which we have previously identified as the centers of the eight translation automorphisms of (γ,δ), when γ is Type 2. We note that since automorphisms preserve subalgebras, every automorphism of (γ,δ) must permute the one-element subalgebras, i.e., the δ-idempotents, and thus the subalgebra I is fixed under the action of any automorphism.

We will now extend the algebra of Fig. 9.15 to a subalgebra of (γ,δ) which is double in size. First we identify the points P on γ with the property that $\sqrt{P}$ is a subalgebra of (γ,δ).

Theorem 23. *Let γ be a nonsingular irreducible cubic which inverts onto itself* via *a circle with center X, and let P be a point on γ. Then $\sqrt{P}$ is a subalgebra of (γ,δ) iff $P \in \sqrt{\sqrt{\infty * \infty}}$.*

Proof. Let $a,b,c \in \sqrt{P}$, so $a*a = b*b = c*c = P$, or in terms of the group operation on γ, $-2a = -2b = -2c = P$. Consider: $\sqrt{P}$ is a subalgebra of $(\gamma,\delta) \leftrightarrow \delta(a,b,c) * \delta(a,b,c) = P \leftrightarrow (\infty - a - b - c) * (\infty - a - b - c) = P \leftrightarrow -\infty - \infty + 2a + 2b + 2c = P \leftrightarrow 4P = -\infty - \infty \leftrightarrow P \in \sqrt{\sqrt{\infty * \infty}}$ □

By comparing Theorems 18 and 23, we see that $\sqrt{P}$ is a subalgebra of (γ,δ) iff the translate Φ_P is an automorphism of (γ,δ).

Now let γ be a Type 2 self-inversive cubic curve. Then, in the notation of Theorem 18, $\sqrt{\sqrt{\infty * \infty}}$ consists of the points {X,Y, U,T}, and the subalgebra:

$$\sqrt{\sqrt{\sqrt{\infty * \infty}}} = \{p, q, r, s, g, h, k, j,\ a, b, c, d, e, f, m, n\}$$

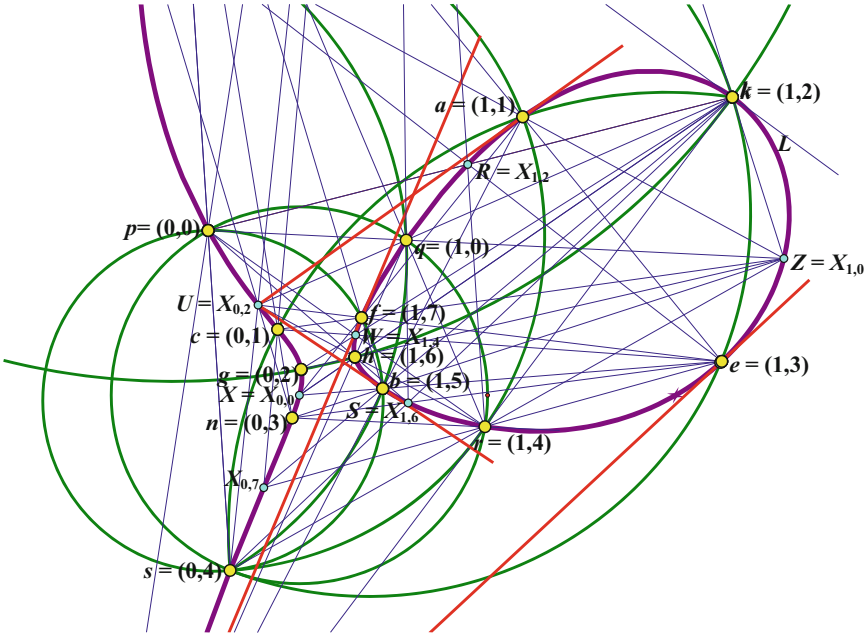


Fig. 9.16 $(Z_2 \times Z_8, (0,0))$ circle system

is illustrated in Fig. 9.16, where we indicate that it can be realized as a $(Z_2 \times Z_8, (0,0))$ circle system. Theorem 23 is generalized as follows. For ease of notation we define a mapping ζ: $\gamma \to \gamma$ by $\zeta(x) = x^*x$. This mapping, and thus ζ^n, is an endomorphism of $(\gamma,*)$ as well as $(\gamma,+)$.

Theorem 24. *Let γ be a nonsingular irreducible self-inversive cubic curve and let $P \in \gamma$. Then $\sqrt[n]{P}$ is a subalgebra of (γ,δ) iff $P \in \sqrt{\sqrt{\zeta^n(\infty)}}$.*

Proof. Let $a,b,c \in \sqrt[n]{P}$, so $\zeta^n(a) = \zeta^n(b) = \zeta^n(c) = P$. Then $\sqrt[n]{P}$ is a subalgebra of (γ,δ) iff $P = \zeta^n(\delta(a,b,c)) = \zeta^n(\infty - a - b - c) = \zeta^n(\infty) - \zeta^n(a) - \zeta^n(b) - \zeta^n(c) = \zeta^n(\infty) - 3P$ iff $4P = \zeta^n(\infty)$ iff $P \in \sqrt{\sqrt{\zeta^n(\infty)}}$. □

Now suppose $n = 2$ in Theorem 24, and γ is Type 2. The equation $(P^*P)^*(P^*P) = (\infty^*\infty)^*(\infty^*\infty)$ is satisfied by $P = \infty$ and $P = X^*Y$ where X,Y are the points on the bell which serve as centers of the two orthogonal circles which invert γ onto itself, and $P \in \text{bell} \cap \sqrt{Q}$ where Q is the element besides $\infty^*\infty$ which lies in $\text{bell} \cap \sqrt{(\infty * \infty) * (\infty * \infty)}$. If $P = \infty$, then $\sqrt{\sqrt{P}} = I$, and if $P = X^*Y$, then $\sqrt{P} \cap \text{bell} = \{U,T\}$, in terms of the notation of Theorem 18, and $\sqrt{\sqrt{P}} = \sqrt{U} \cup \sqrt{T}$ is a subalgebra of $\sqrt{\sqrt{\sqrt{\infty * \infty}}}$ which can be realized as a $(Z_2 \times Z_4, (0,2))$ circle system, as in Fig. 9.17. In Fig. 9.18 we illustrate the subalgebra $\sqrt{\sqrt{P}}$ that results if we choose $P \in \text{bell} \cap \sqrt{Q}$. (Of the two points in $\text{bell} \cap \sqrt{Q}$, only one: $P = X_{(0,0)}$ is illustrated.) It can be realized as a $(Z_2 \times Z_4, (0,1))$

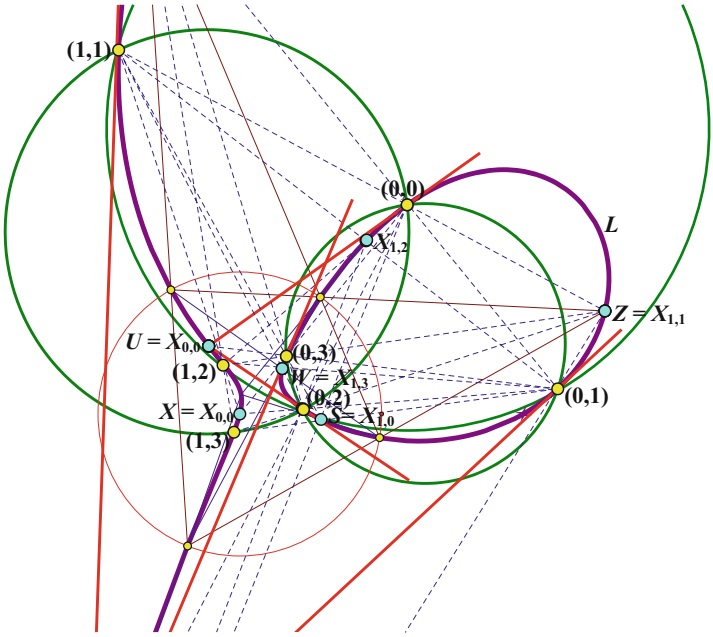


Fig. 9.17 $\sqrt{\sqrt{X * Y}}$ realized as a $(Z_2 \times Z_4, (0,2))$ circle system

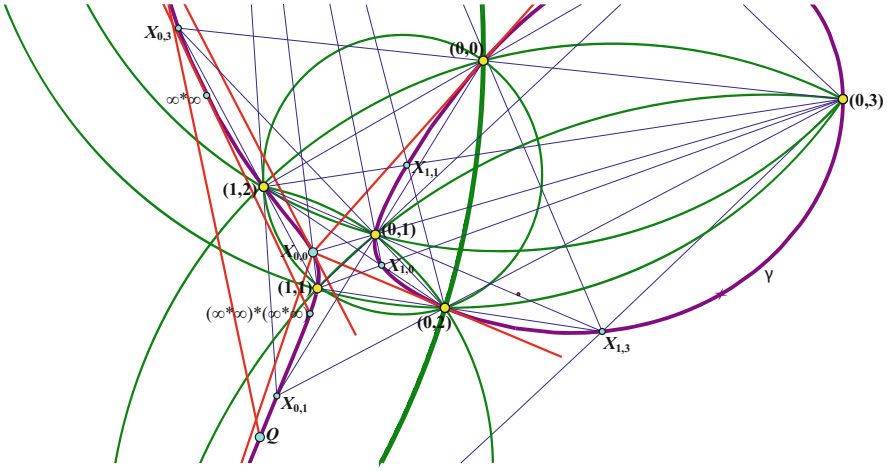


Fig. 9.18 $\sqrt{\sqrt{P}}$ realized as a $(Z_2 \times Z_4, (0,1))$ circle system

circle system. Next we develop a simple result which allows us to build larger subalgebras from a given subalgebra of (γ,δ).

Theorem 25. *Let γ be a nonsingular irreducible cubic curve which inverts onto itself via a circle with center X, and let Ω be a subalgebra of (γ,δ). If $a,b,c \in \Omega$, then $a*(b*c) \in \Omega$.*

Proof. Since Ω is a subalgebra, the element $d = \delta(a,b,c)$ lies in Ω. Then we have $d = X*\{(X*a)*(b*c)\} \leftrightarrow X*d = \{(X*a)*(b*c)\} \leftrightarrow b*c = (X*d)*(X*a)$. Thus, $a*(b*c)=a*\{(X*d)*(X*a)\}=\{a*(a*a)\}*\{(X*d)*(X*a)\}=\{a*(X*a)\}*\{(X*d)*(a*a)\}$ $= X*\{(X*d)*(a*a)\} = \delta(d,a,a) \in \Omega$. □

We call $a*(b*c)$ a *triple product,* and say Ω is *closed under taking triple products*. As an immediate consequence we obtain a result similar to the theorem of Lagrange in the theory of groups.

Theorem 26. *Let γ be a nonsingular irreducible cubic curve which inverts onto itself via a circle with center X, and let Ω be a subalgebra of (γ,δ). Let $Z = \Omega*\Omega = \{a*b: a,b \in \Omega\}$ and let Γ be a subalgebra of (Ω,δ). Then the distinct sets in the collection $\{g*\Gamma: g \in Z\}$ form a partition of Ω. If Ω is finite, then $|\Gamma|$ divides $|\Omega|$.*

Proof. Let $g,h \in Z$ and suppose $g*\Gamma \cap h*\Gamma$ is nonempty. Then $\exists$ $s,t \in \Gamma$ such that $g*s = h*t$. Suppose $p \in g*\Gamma$, then $p = g*a$ for some $a \in \Gamma$ and we obtain $p=\{s*(h*t)\}*a=\{s*(h*t)\}*\{t*(a*t)\}=\{(h*t)*t\}*\{s*(a*t)\}=h*\{s*(a*t)\} \in h*\Gamma$ since Γ is closed under triple product by Theorem 25. Thus $g*\Gamma \subseteq h*\Gamma$, and similarly $h*\Gamma \subseteq g*\Gamma$, so $g*\Gamma = h*\Gamma$. Thus the sets $\{g*\Gamma: g \in Z\}$ are either identical or disjoint. Note that $g*\Gamma \subseteq \Omega$ since $x \in g*\Gamma$ implies $x = (c*d)*s$ is a triple product with c,d,s in Ω. Let $a \in \Omega$, let $s \in \Gamma$, and let $g = a*s$. Then $g \in \Omega*\Omega$ and $a \in g*\Gamma$. The sets $g*\Gamma$ are all translates of Γ, so $|g*\Gamma| = |\Gamma|$, and if $|\Omega|$ is finite, it must follow that $|\Gamma|$ divides $|\Omega|$. □

Let a_0 be a fixed point of Ω, and suppose $a,b \in \Omega$. Then $a*b=a_0*\{a_0*(a*b)\} \in a_0*\Omega$ since Ω is closed under taking triple products. Consequently $\Omega*\Omega = a_0*\Omega$, and $|\Omega*\Omega| = |a_0*\Omega| = |\Omega|$. We call $\Omega*\Omega$ the *perspective set of* Ω. We have shown that the perspective set of Ω has the same cardinality as Ω. If g lies in the perspective set of Ω, then the collection of sets $\Pi_g = \{\{a, a*g\}: a \in \Omega\}$ form a partition of Ω. The perspective set of Ω thus resembles the perspective set of a perfect G-gon, the difference being that the elements of Ω have not been labeled with elements from an abelian group. In the above examples we see that each root subalgebra can be realized as a group circle system, and consequently as a perfect G-gon for some abelian group G. An important open question is whether every subalgebra, or at least every finite subalgebra of (γ,δ), can be realized as a group circle system. Our next result gives a method for building a larger subalgebra from a given subalgebra of (γ,δ).

Theorem 27. *Let γ be a nonsingular irreducible cubic curve which inverts onto itself via a circle with center X, and let Ω be a subalgebra of (γ,δ). Then $\Omega \cup (X*\Omega)$ is also a subalgebra of (γ,δ).*

Proof. The mapping ϕ_X, introduced in Sect. 9.7, is an automorphism of (γ,δ), and thus $X*\Omega$ is a subalgebra of (γ,δ) isomorphic to Ω. Let $a,b,c \in \Omega \cup (X*\Omega)$. If $a,b,c \in \Omega$, or $a,b,c \in X*\Omega$, then $\delta(a,b,c) \in \Omega \cap (X*\Omega)$ as desired. So suppose $a,b \in \Omega$, and $c \in X*\Omega$, then $\delta(a,b,c) = \delta(a,b, X*c_0)$ for some $c_0 \in \Omega$. Then $\delta(a,b,c) = X*\{(X*(X*c_0))*(a*b)\} = X*\{c_0*(a*b)\} \in X*\Omega$, since $c_0*(a*b)$ is a triple

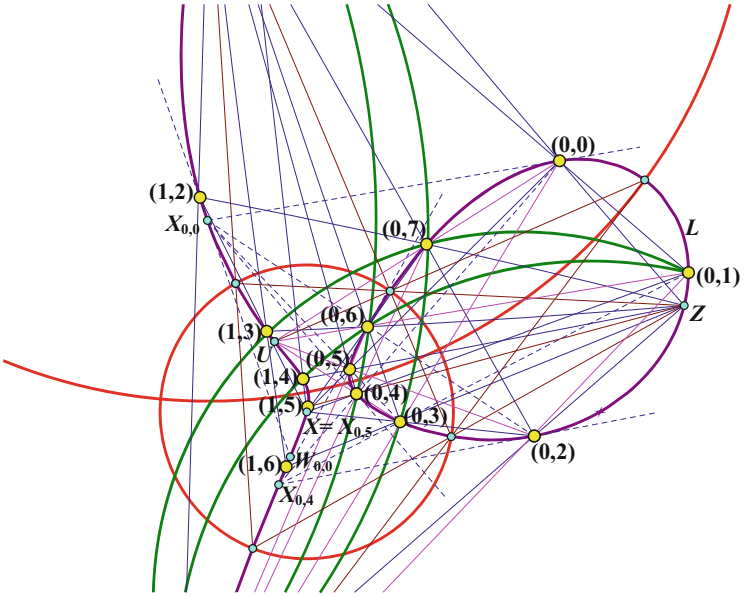


Fig. 9.19 ($Z_2 \times Z_8$, (0,2)) circle system

product of elements from Ω. The remaining possibility to consider is $a \in \Omega$ and $b,c \in X^*\Omega$. In this case $\delta(a,b,c) = \delta(a, X^*b_0, X^*c_0)$ for some $b_0, c_0 \in \Omega$. Then, $\delta(a,b,c) = X^*\{(X^*(X^*b_0))^*(a^*(X^*c_0))\} = X^*\{b_0^*(a^*(X^*c_0))\} = \{c_0^*(X^*c_0)\}^*\{b_0^*(a^*(X^*c_0))\} = a^*(c_0^*b_0)$ which is a triple product and so lies in Ω. □

Let Γ denote the ($Z_2 \times Z_4$, (0,1)) circle system constructed after Theorem 24 and illustrated in Fig. 9.18. Then by Theorem 27, $\Gamma \cup (X^*\Gamma)$ is also a subalgebra of (γ,δ). In Fig. 9.19 this subalgebra is realized as a ($Z_2 \times Z_8$, (0,2)) circle system. We will show that it can also be represented as the root subalgebra $\sqrt{\sqrt{\sqrt{(X * Y) * \infty}}}$. An important result which we will make use of is the following. The proof, for the special case $G = Z_n$, can be found in Fletcher R (Perfect Polygons, unpublished).

Theorem 28. *Let Γ be a perfect G-gon with nonsingular irreducible cubic envelope γ, and let Γ_G denote the set $\{X_g: g \in G\}$ of perspective points of Γ. Then, if no three points of Γ_G are collinear, Γ_G is also a perfect G-gon, i.e., for each $g \in G$, the set of lines $\{[X_a, X_b]: a + b = g\}$ is concurrent at a point W_g. If Γ_G contains three collinear points, say, X_a, X_b, X_c, then Γ_G is a (G, m) triple system where $m = a + b + c$.* □

Proof. Let $s,t,p,q \in G$ with $s + t = p + q$. Then $X_s^*X_t = (p^*(s - p))^*(0^*t) = (p^*0)^*((s - p)^*t) = X_p^*X_q$. Thus, if $g \in G$, all lines of the form $\{[X_s,X_t]: s + t = g\}$ meet at the same point W_g. We then obtain, with the further provision, that no three perspective points of Γ are collinear, that Γ_G is a perfect G-gon under the

identification $g \leftrightarrow X_g$. Now suppose Γ_G contains three collinear points: X_a, X_b, X_c, so $X_b * X_c = X_a$, and suppose s,t,p $\in G$ with $s + t + p = a + b + c$. We want to show $X_s * X_t = X_p$. Consider: $X_s * X_t = (a*(s-a))*(0*t) = (a*0)*((s-a)*t) = X_a * X_{s+t-a} = (X_b * X_c) * X_{b+c-p} = (X_b * X_c)*(b*(c-p)) = (X_b*b)*(X_c*(c-p)) = 0*p = X_p$. It now follows that Γ_G is a (G,m) triple system with $m = a + b + c$. □

From this theorem we might say that a geometric triple system is a *degenerate perfect polygon*. Now when the points of the $(Z_2 \times Z_4, (0,1))$ circle system Ω of Fig. 9.16 are relabeled as in Fig. 9.19, we have $(0,0)*(0,0) = (0,4)*(0,4) = X_{(0,0)}$, and $(0,2)*(0,2) = (0,6)*(0,6) = X_{(0,4)}$. In accordance with Theorem 27, we have $X_{(0,0)} * X_{(0,0)} = X_{(0,4)} * X_{(0,4)} = W_{(0,0)}$ where $W_{(0,0)} \in \sqrt{\sqrt{\zeta^2(\infty)}}$. Let $V_{(0,0)} = W_{(0,0)} * W_{(0,0)}$, then $V_{(0,0)} \in \sqrt{\zeta^2(\infty)} \cap \text{bell} = \{\infty*\infty, (X*Y)*\infty\}$, and since $V_{(0,0)} \neq \infty*\infty$, we have $V_{(0,0)} = (X*Y)*\infty$. Note that $\Omega = \sqrt{\sqrt{W_{(0,0)}}} = \{(0,0),(0,2),(0,4),(0,6),(1,0),(1,2),(1,4),(1,6)\}$. Now consider the subalgebra $\Omega \cap (X*\Omega)$ constructed after Theorem 26. We want to show that $\Omega \cup (X*\Omega) = \sqrt{\sqrt{\sqrt{V_{(0,0)}}}} = \sqrt{\sqrt{\sqrt{(X*Y)*\infty}}}$. We define the points in $X*\Omega$ by $(0,1) = X*(0,4)$; $(0,3) = X*(0,2)$; $(0,5) = X*(0,0)$; $(1,1) = X*(1,4)$; $(1,3) = X*(1,2)$; $(1,5) = X*(1,0)$; $(1,7) = X*(1,6)$. Suppose $t \in \{1,3,5,7\}$, then $(0,t)*(0,t) = (X*(0,\ 5-t))*\ (X*(0,\ 5-t)) = \infty * X_{(0,\ 2-2t)}$, and $\zeta^2((0,t)) = (\infty*\infty)*W_{(4-4t)}$. Finally, $\zeta^3((0,t)) = \zeta^2(\infty) * V_{(8-8t)} = \zeta^2(\infty) * V_{(0,0)} = (V_{(0,0)} * V_{(0,0)}) * V_{(0,0)} = V_{(0,0)}$. So $(0,t) \in \sqrt{\sqrt{\sqrt{V_{(0,0)}}}}$ and similarly $(1,t) \in \sqrt{\sqrt{\sqrt{V_{(0,0)}}}}$ for each $t \in \{1,3,5,7\}$.

In our final result we present yet another way to build subalgebras of (γ,δ).

Theorem 29. *Let γ be a nonsingular irreducible cubic curve which inverts onto itself via a circle with center X, and let Ω be a subalgebra of (γ,δ). Then $\sqrt{\Omega * \Omega}$ is also a subalgebra of (γ,δ).*

Proof. Let $a,b,c \in \sqrt{\Omega * \Omega}$, then $a*a = s_1*s_2$; $b*b = t_1*t_2$; and $c*c = q_1*q_2$, where $s_1,s_2,\ t_1,t_2,\ q_1,q_2 \in \Omega$. Then $\delta(a,b,c) * \delta(a,b,c) = (X*X)*\{\{(X*X)*(a*a)\}*\{(b*b)*(c*c)\}\}$ as in the proof of Theorem 22. Thus $\delta(a,b,c) * \delta(a,b,c) = (X*X)*\{\{(X*X)*(s_1*s_2)\}*\{(t_1*t_2)*(q_1*q_2)\}\} = \delta(s_1, t_1, t_2)\ * \delta(s_2, q_1, q_2) \in \Omega*\Omega$. □

In Fig. 9.20 we have constructed a $(Z_2 \times Z_{12}, (0,0))$ circle system on a self-inversive Type 2 cubic curve γ by extending a $(Z_6, 0)$ circle system. If Ω denotes the subalgebra of (γ,δ) determined by the $(Z_6, 0)$ circle system, then the $(Z_2 \times Z_{12}, (0,0))$ circle system represents the subalgebra $\sqrt{\Omega * \Omega}$. The vertices of Ω have been relabeled according to the formula $x \to (0,2x)$ for $x \in \{0,1,2,3,4,5\}$, then the remaining labels are assigned to realize $\sqrt{\Omega * \Omega}$ as a $(Z_2 \times Z_{12}, (0,0))$ circle system.

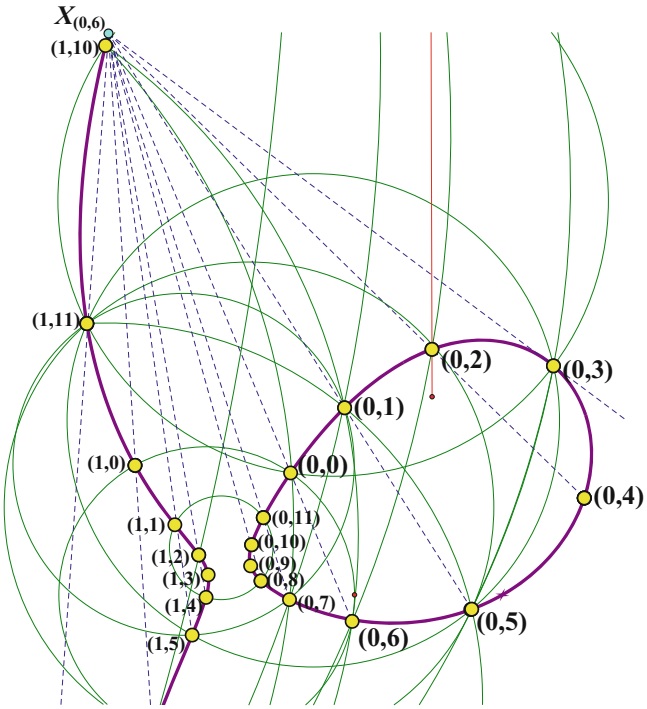


Fig. 9.20 ($Z_2 \times Z_{12}$, (0,0)) circle system

9.9 Conclusion

We have used the theory of geometric triple systems (Fletcher R. Geometric Triple Systems, unpublished) and group circle systems to determine properties of a self-inversive cubic curve γ. The key result involves the creation of an algebra (γ,δ) with a single ternary operation δ defined on the points of γ. Group circle systems reemerge in the study of finite subalgebras of (γ,δ), leading to a major open question which we state as a Conjecture:

Conjecture. *Let γ be an irreducible nonsingular self-inversive cubic curve. Then every finite subalgebra of (γ,δ) can be realized as a group circle system (G,g) for some abelian group G and some $g \in G$.*

If it is shown that the vertices of a given (G,g) circle system must lie on a self-inversive cubic curve γ, then the limitations imposed by the structure of γ limit the possibilities for G and g. It can be shown, for example, that the THAs with base group $Z_2 \times Z_8$ and ternary operations defined by $\delta(a,b,c) = (1,0) - a - b - c$ or $\delta(a,b,c) = (1,2) - a - b - c$ cannot be realized as group circle systems on a self-inversive cubic. In accordance with the opening paragraph of Sect. 9.4, an inversion can be used to convert a given group circle system into a geometric triple system.

So the determination of exactly which abelian groups can be used as a basis for a circle system is relayed to the same question for geometric triple systems. This latter question is a main topic in (Fletcher R. Geometric Triple Systems, unpublished), which is still unfinished at this writing. It appears that the groups Z_n, $Z_2 \times Z_n$, $Z_2 \times Z_2 \times Z_2$, Z^n, $Z_2 \times Z^n$ can all serve as bases for group circle systems. The THA $(Z_2 \times Z_2 \times Z_2, \delta)$ with $\delta(a,b,c) = (0,0,0) - (a + b + c)$ cannot exist on a self-inversive cubic since all eight points are δ-idempotent, but the δ-idempotent points on a Type 2 cubic form the subalgebra I which is realized as a $(Z_2 \times Z_4, (0,0))$ circle system in Theorem 21.

Addendum

Here we include an alternative proof of Theorems 9 and 10 suggested by the referee.

Theorem 30. *Let γ be a nonsingular irreducible cubic curve in the real projective plane which contains the points $I = (1,i,0)$ and $J = (1,-i,0)$ in the complex projective plane and also the point $T = (0,1,0)$ at infinity on vertical lines. Then we can define a ternary operation δ on set the points of γ by setting $\delta(a,b,c)$ equal to the unique fourth point on γ and on circle (a,b,c). In case a,b,c are collinear set $\delta(a,b,c) = T$. Then (γ,δ) is a THA.*

Proof. Using the group structure on γ let $d = T - a - b - c$ for any points a,b,c of γ, and suppose O is a flex of γ and the group operation on γ is defined by $a + b = O*(a*b)$. Since I,J lie on the vertical line $x = 1$, we have $I*J = T$, and thus $a + b + c + d + I + J = T + I + J = T + (O*(I*J)) = T + (O*T) = T + (-T) = O$. Now, in accordance with Exercise 14.12 in [2], the six points a,b,c,d,I,J lie on a unique complex curve ρ of degree 2. This curve has the form:

$$Ax^2 + Bxy + Cy^2 + Dxz + Eyz + Fz^2 = 0. \tag{9.21}$$

Since ρ contains the points I,J we must have $A + Bi - C = 0$ and $A - Bi - C = 0$, and thus $A = C$ and $B = 0$. So (9.21) reduces to

$$Ax^2 + Ay^2 + Dxz + Eyz + Fz^2 = 0. \tag{9.22}$$

Since the coefficients of x^2 and y^2 are identical, (9.22) represents the equation of a circle if the points a,b,c are noncollinear. In this case $\delta(a,b,c) = d$. If a,b,c are collinear, then (9.22) must represent the product of two lines, and we set $\delta(a,b,c) = T$. Note that if a,b,c are collinear, then $b*c = a$ and $d = T - a - b - c = (T + (-a)) + (-b + -c) = (O*(T*-a)) + (O*(-b * -c)) = (O*(T*(O*a))) + (O*((O*b)*(O*c))) = (O*O)*(T*(O*a)) + (O*((O*O)*(b*c))) = (O*T)*a + (b*c) = a*(O*T) + a = O*(a*(a*(O*T))) = T$, so in all cases $\delta(a,b,c) = T - a - b - c$.

Then $\delta(\delta(a,b,c),\delta(d,e,f),q) = T - (T - a - b - c) - (T - d - e - f) - q = -T - q + a + b + c + d + e + f$, and $\delta(\delta(a,b,d),\delta(c,e,f),q) = T - (T - a - b - d) - (T - c - e - f) - q = -T - q + a + b + c + d + e + f$.

The identity (9.12) thus holds, and it follows that (γ,δ) is a THA. □

Theorem 30 does not represent a generalization of Theorems 9 and 10 as we will now show. Let γ denote the nonsingular irreducible cubic of Theorem 30. The homogenized version of a general cubic is given by

$$ax^3 + bx^2y + cxy^2 + dy^3 + ex^2z + fxyz + gy^2z + hxz^2 + kyz^2 + jz^3 = 0 \tag{9.23}$$

Now suppose (9.23) represents the equation of γ. Since $T = (0,1,0)$ lies on γ, we must have $d = 0$, and since I,J lie on γ we must have $a + bi - c = 0$ and $a - bi - c = 0$ which yields $a = c$ and $b = 0$. So (9.23) reduces to

$$ax\left(x^2 + y^2\right) + ex^2z + fxyz + gy^2z + hxz^2 + kyz^2 + jz^3 = 0, \tag{9.24}$$

and the restriction to the Euclidean plane is

$$ax\left(x^2 + y^2\right) + ex^2 + fxy + gy^2 + hx + ky + j = 0. \tag{9.25}$$

The equation:

$$(ax + by)\left(x^2 + y^2 + z^2\right) + \left(ex^2 + fxy + gy^2\right)z = 0. \tag{9.26}$$

is the homogenized version of (9.5) which represents our general equation for an irreducible cubic curve which inverts onto itself via the unit circle. We showed in Sect. 9.2 that this curve has exactly one point at infinity. We may rotate (9.26) so that $T = (0,1,0)$ is the unique point at infinity, while remaining self-inversive through the unit circle. But T lies on the curve with Eq. (9.26) iff $b = 0$, so (9.26) reduces to

$$ax\left(x^2 + y^2 + z^2\right) + \left(ex^2 + fxy + gy^2\right)z = 0, \tag{9.27}$$

with restriction to the Euclidean plane given by

$$ax\left(x^2 + y^2 + 1\right) + \left(ex^2 + fxy + gy^2\right) = 0, \tag{9.28}$$

We then recognize that (9.25) is just a translation, or a dilation followed by a translation, of (9.28). Since neither of these transformations spoils the self-inversive property of the curve, we conclude that the curve in Theorem 30 is self-inversive and thus Theorem 29 does not represent a generalization of our Theorems 9 and 10. In Fig. 9.21 we illustrate a cubic curve γ of the type described in Theorem 30 where T denotes the point at infinity on γ. The points a,b,c are any points on γ, and it can be seen that $T - a - b - c$ is the fourth point on γ and on the circle (a,b,c). We have also indicated two orthogonal circles c_1 and c_2 which invert γ onto itself.

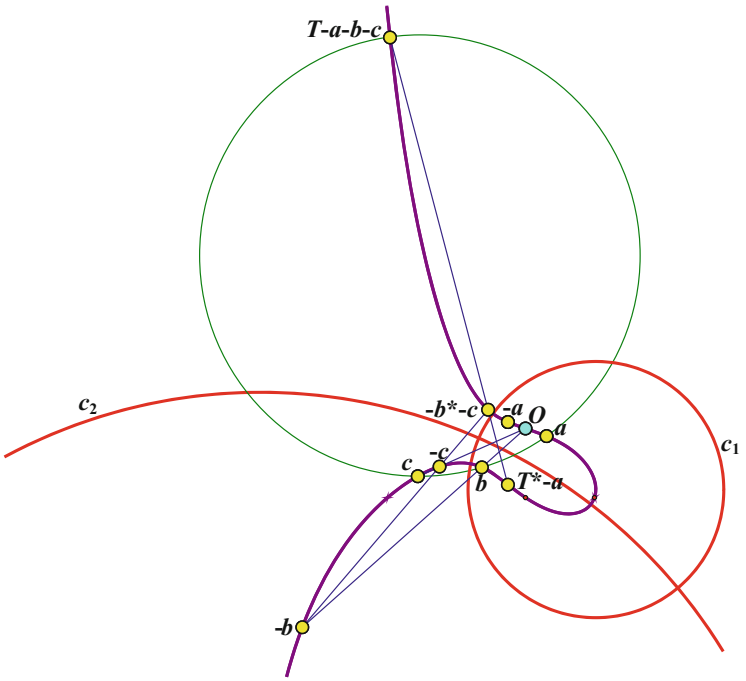


Fig. 9.21 δ(a,b,c) = T − a − b − c

References

1. Fletcher, R.: Circle chains. J. Appl. Global Res. **1**(3), Intellectbase International Consortium (2008)
2. Bix, R.: Conics and Cubics. Springer, USA (2006)
3. Fletcher, R.: Group Circle Systems on Conics, New Frontiers of Multidisciplinary Research in STEAM-H. Springer International Publishing, Switzerland (2014)
4. Robinson, D.: Self-Inversive Properties of Group Circle Systems. Masters Thesis, Virginia State University Library (2015)

Chapter 10
Elasticity Imaging

Lilí Guadarrama

Abstract This chapter is devoted to summarize different approaches for the imaging technique of elastography: Quasi-Static, Harmonic and Transient elastography, Models for viscoelasticity. This promising technique is a good example of interdisciplinary mathematical research and applications.

Keywords Tissue motion • Palpation • Elastography • Sonoelasticity • Hookian materials • Bulk waves and modulus • Shear waves and modulus • Young's modulus • Lamé parameters • Magnetic resonance imaging • Acoustic radiation force impulse • Shear wave elasticity imaging • Supersonic shear imaging • Optical coherence tomography • Crawling wave imaging • Helmholtz equation • Aixplorer • Rayleigh damping model • Voigt model • Ultrasound methods • Fibroscan • Phase-contrast method

10.1 Introduction

The aim of this chapter is to provide a survey to the different approaches to assess mechanical properties of biological tissues. The motivation for looking at these properties is because physicians can detect changes in the mechanical properties in tissues due to pathologies in a palpation examination. The elastography is a promising imaging technique that aims to take the palpation examination to the next level. In the last two decades, a lot of work has been done in theoretical research using different approaches of this technique, in optimization methods and even the development of technology that has been used in clinical trials.

The principle of elastography is the following, a mechanical excitation (quasi-static compression, harmonic or transient vibration) is generated (externally or internally) in the tissue. Then the resulting tissue displacements are measured by different imaging modalities (MRI or ultrasound) and from this information the mechanical properties are estimated.

L. Guadarrama (✉)
CONACyT/CIMAT Aguascalientes, Fray Bartolomé de la Casas 314,
Aguascalientes 20259, AGS, Mexico
e-mail: lili.guadarrama@cimat.mx

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_10

For each approach to elastography we will discuss about the way mechanical excitation is generated, the methods for measuring the displacements induced, signal processing and the different inversion schemes for estimating the elastic properties.

The chapter is organized as follows. Section 10.2 is a brief presentation of the different techniques that study the tissue motions. We present the relations that govern displacement and motion principles of tissues in Sect. 10.3. Section 10.4 is dedicated to quasi-static elastography. Section 10.5 is devoted to harmonic modalities. The transient elastography is presented in Sect. 10.6. In each of these sections we will present a review of inverse reconstruction schemes to estimate the elastic parameters and how they have been used in actual clinical situations. In the last section Sect. 10.7 some more complex models for elastography are presented as well as some direction for future work.

10.2 Overview Studies of Tissue Motion

Palpation has been used by physicians for diagnosing diseases or illness in patients. Although it is considered as an effective method for detecting tumors and other pathologies, palpation has several important limitations as qualitative information of the stiffness of tissues or the physical limitation to reach and palpate them. One alternative to address these limitations is the use of elastography, it is an imaging technique that maps the elastic parameters of soft tissue.

Elastography is a field that has evolved during the past two decades, many approaches have been developed and studied. To mention some of them we have vibration elastography imaging, compression elastography, magnetic resonance elastography (MRE), shear wave imaging, transient elastography, acoustic radiation force imaging, crawling wave imaging, spatially modulated ultrasound radiation force. Some of these techniques have been commercialized as FibroScan by EchoSens, Aixplorer by Supersonic Imagine, Acuson S2000/S3000 by Siemens, Hitachi EUB-8500 by Hitachi, Sonix Elastography by BK Ultrasound and others.

The use of ultrasound in medical imaging was first applied in the late 1940s but it was not used in the study of tissue motions until the beginning of the 1980s where Dickinson and Hill [23], Wilson and Robinson [98] studied the displacement and deformation of tissue near blood vessels due to the cardiac contractions and pulsation of blood flow to estimate elastic parameters. They used M-modes and A-scans analysis, respectively, to obtain a rough estimation of the elastic parameters. Following this technique, Tristan et al. [91] studied the correlation between normal and cancerous liver. In this decade, studies to estimate qualitatively the stiffness of fetal lungs were made by using B-scans and M-modes [1, 12]. Doppler ultrasound was also used to measure tissue elasticity [21, 36, 48].

Fatemi and Greenleaf [30] presented an ultrasound imaging technique, based on oscillatory radiation force, to map the acoustic response of tissues, that they called ultrasound-stimulated vibro-acoustic spectrography. They reported promising results in calcification within the arteries.

In the late 1980s Lerner and Parker [51] presented a method in which external mechanical excitation is generated and the resulting tissue displacements are measured by Doppler ultrasound detection, they called it *sonoelasticity*. They were able to obtain real-time images and by the beginning of 1990s they could measure the shear speed of sound in the tissue and estimate the elastic parameters [67]. A couple of years later a theoretical model of sonoelasticity was proposed by Gao et al. in [32]. Further approaches to sonoelasticity were made in [37–39, 53, 89, 100, 101].

The name of *elastography* is due to Ophir et al. [64]. They present a method to imaging strain and elastic modulus, this technique is also called strain imaging or compression imaging. In this technique the displacements are obtained by comparing the data of B-scan ultrasound before and after compression tissue. The strain is then estimated from the derivative of the displacements.

The waves in an elastic material have two components, bulk and shear waves, both of them are used in medical images; the bulk waves have been used for more than half a century, the B-mode imaging is an example of the use of this type waves. In the past two decades the use of shear waves in imaging techniques has increased. In [76] they discuss the differences between these two basic modes of waves in medical imaging. Since shear waves have a larger dynamical range than bulk waves a better characterization of tissues can be done, most of the modalities of elastography use shear wave imaging.

In the next sections will refer as quasi-static elastography the techniques that have quasi-static compression as mechanical excitation and as dynamic elastography those that have harmonic or transient vibration.

10.3 Governing Principles

The most common mathematical model for soft tissues is the model for Hookian materials, that is, it is assumed soft tissues have linear, isotropic, purely elastic mechanical behavior. However this model is not appropriate for some type of tissues, we will discuss other models in Sect. 10.7.

There are quantities used to characterize Hookian materials such as the Young 's modulus (E) which measures the stiffness of an elastic material. Young's modulus is the ratio of stress to strain. The bulk modulus (K) is the resistance to uniform compression, the shear modulus (μ) is the ratio of shear stress over the shear strain. The Young's modulus is three times the shear modulus. Another physical property is Poisson's ratio (ν) which is the negative ratio of transverse to axial strain; for soft tissues, it is in the range of 0.490–0.499, which is very close to the water [55]. All these parameters are related and any two of them can be calculated from the knowledge of the other two.

The constitutive equations of the relationship between the stress ($\sigma = \{\sigma_{ij}\}$) and strain ($\epsilon = \{\epsilon_{ij}\}$) tensor under these assumptions are

$$\sigma = E\epsilon \tag{10.1}$$

which is equivalent to

$$\sigma_{ij} = 2\mu\epsilon_{ij} + \lambda\delta_{ij}\nabla\cdot\mathbf{u} \tag{10.2}$$

where δ_{ij} is the Kronecker's delta, $\mathbf{u}$ is the displacement vector in $\mathbb{R}^n (n = 2, 3)$, and λ, μ are the Lamé parameters. Lamé's first parameter λ is related to the shear modulus and the bulk modulus by the equation $\lambda = \frac{3K-2\mu}{3}$.

In practice stress cannot be measured, but instead we measure the displacement $\mathbf{u}$. The strain tensor is then related to $\mathbf{u}$ by the equation,

$$\epsilon_{ij} = \frac{1}{2}(\partial_{x_j} u_i + \partial_{x_i} u_j)$$

where $\partial_{x_j} u_i$ denotes $\frac{\partial u_i}{\partial x_j}$. Then the equilibrium equations (10.2) are given in terms of the displacement by

$$\nabla\cdot\mu\nabla\mathbf{u} + \nabla(\lambda+\mu)\nabla\cdot\mathbf{u} = \rho\partial_t^2\mathbf{u} \tag{10.3}$$

where ρ is the density of the material, t is the time, and $\partial_t^2 = \frac{\partial^2}{\partial t\partial t}$. The inverse problem of elastography is to estimate ρ, λ, μ from the measurement of the displacement vector $\mathbf{u}$.

By considering different modalities, Eq. (10.3) take specific forms, in the case of quasi-static elastography they are reduced to

$$\nabla\cdot\mu\nabla\mathbf{u} + \nabla(\lambda+\mu)\nabla\cdot\mathbf{u} = 0 \tag{10.4}$$

while for harmonic modalities we get

$$\nabla\cdot\mu\nabla\mathbf{u} + \nabla(\lambda+\mu)\nabla\cdot\mathbf{u} = \rho\omega^2\mathbf{u} \tag{10.5}$$

In the case of harmonic motion, the Lamé parameters are complex quantities.

On the other hand, the direct problem of equations (10.3) models the wave propagation in an elastic material. As we mentioned before, there are two types of elastic body waves, the shear waves and the bulk waves and they both obey the Helmholtz equation [90]

$$\partial_t^2 v = c_i^2\Delta v \qquad i = s, p \tag{10.6}$$

where c_s, c_p are the shear and bulk wave speed, respectively, and they are given by

$$c_s = \sqrt{\frac{\mu}{\rho}} \qquad c_p = \sqrt{\frac{(\lambda+2\mu)}{\rho}}$$

Imaging technologies use these two types of waves. Historically the bulk waves were the first ones to be used for imaging biological materials, an example of this

is B-mode imaging which is the basic ultrasound method; this imaging technique is based on the differences in the acoustic impedance I, which is related with the bulk wave speed by

$$I = \rho c_p,$$

In it, ultrasound pulses are sent into the tissue and from the resulting echoes from the scattering anomalies a tomographic image is done.

One of the advantages of working with shear waves is that they have a larger dynamical range than bulk waves (for a compendium of the variation of compressional and shear wave speeds in biological tissue, see [76]). Because of this variability a better characterization of tissues can be done and the modalities of elastography that use shear wave imaging have a better quantification of the elastic parameters. The optimal would be to work with both waves, however the big difference of order of magnitude $\lambda >> \mu$ makes it very difficult to estimate both parameters λ, μ at the same time.

Usually quasi-static and harmonic elastography assume local homogeneity which considerably simplifies their numerical implementation. On the downside, we do not get accurate estimations near boundaries. However, there are techniques such as the ones that use radiation force to induce the displacement, that can estimate the elastic parameters without the knowledge of the values in the boundaries. Since biological tissues are not entirely homogeneous there have been efforts to avoid assuming local homogeneity, however the results obtained so far have a high computational cost [92].

10.4 Quasi-Static Elastography

In 1991 Ophir et al. [64] presented a method for imaging the elasticity of biological tissues that they called *elastography*. Basically the method is described as follows: A transducer is coupled to the body to acquire an echo signal for a given time period, then the transducer is pressed into the body and another echo signal is recorded. The displacements in the tissue are then estimated from these two echo signals (ultrasound B-scans), then using the relation (10.3) between strain and displacement the longitudinal strain is calculated and so a strain image is formed. This technique was conceptualized through the use of springs, the stiffest spring will compress the least.

For measuring displacements using magnetic resonance there are two methods, saturation tagging and phase-contrast. In 1995, Fowlkes et al. [31] presented an approach to MRE using the saturation tagging method to measure displacement. At the same time Plewes et al. [71] studied MRE that uses a phase-contrast imaging method. The approaches of quasi-static elastography using magnetic resonance imaging have longer acquisition times than ultrasound approaches, up to 15 times more, but the images obtained show better contrast.

In general, in quasi-static approaches by ultrasound methods, one measures the axial strain induced by an external, quasi-static source. The displacements are of order 2 % of the axial dimension, and they are measured using cross-correlation from echo signals before and after the compression [26]. To map the strain from the displacement a finite difference or a least-squares strain estimator is used [43], this gray-scale strain images are called elastograms.

To compute the elastic parameters there are two principal inversion schemes, the direct and the iterative inversion. We begin with the direct ones, in 1994 Raghavan and Yagle [73] derived a linear system rearranging the equation of the forward problem which includes the hydrostatic pressure and they solved the inverse problem using LU decomposition, which is faster than the iterative methods presented in [82]. To solve the linear system, the shear modulus and the hydrostatic pressure must be known at the boundary, but in practice there is a complication since hydrostatic pressure at the boundary cannot be measured. In order to eliminate the hydrostatic pressure of the linear system an analytic method was used by Skovoroda et al. [82], which is independent of global boundary conditions. The resulting system can be solved if the shear modulus is known at the boundary of the region of interest, it can be computed by the stress-continuity properties of soft tissues. A disadvantage of this system is that it contains high order derivatives that amplify the measurement noise. However, they designed a reconstruction procedure, which they called hybrid, that considerably reduces the artifacts in elastograms thus obtained. We will treat this artifacts further along. Studies on phantoms and an ex vivo kidney were performed using this inversion scheme by MRE [20] to show its performance.

Sumi et et al. [86] proposed a direct inversion solution, in which the system to solve has the spatial derivatives of Young's modulus as the unknowns and has the strain and their spatial derivatives as coefficients. The effectiveness of this method was verified in agar phantoms and in liver carcinoma [84].

The iterative inversion schemes propose to optimize a functional which minimizes the difference between measured displacement and the ones obtained by solving the forward problem, by iterative techniques [96]. We can classify these techniques [24] in the following, the Hessian base method which was used by Kallel and Bertrand [42], Doyley et al. [25], and Richards et al. [74], Harrigan el al. [35] each of them used different regularization methods to ameliorate the behavior of the ill-conditioned Hessian matrix. It has been reported that for clinical purposes the use of the FMINCON function, MATLAB solver for optimization problems, has been very useful [41] so the design of an iterative inversion scheme custom-made can be avoided. Another type of iterative inversion is the gradient optimization method where it is used the adjoint method to compute the gradient of the objective function, Oberai et al. [63] were the first to used this inversion. Among the noniterative inversion schemes we can mention the genetic algorithms used by Zhang et al. [103].

There are artifacts in elastograms that are well identified [83], an example is the darkening in central or brightening in boundary area of compression due to non uniform stress distribution that may compromise the diagnosis [47]. The assumption of that the internal stress distribution, σ, is constant (see Eq. (10.1)) produces another artifact called target hardening, which is the darkening of deep areas and

is dependent on the external source [72]. Quasi-static methods need knowledge of boundary conditions outside of the region under investigation, strain image may exhibit significant artifacts due to global boundary conditions [82]. In addition to the artifacts presented in elastograms, quasi-static elastography has the disadvantage that deeper organs cannot be reached with this technique and sometimes the section under investigation moves out of plane during compression.

Doyley et al. [27] studied the quality of modulus elastograms by solving directly the inverse elasticity problem and in that way avoid the mechanical artifacts caused by the assumption of stress uniformity and compared them with the ones computed on the assumption. The evaluation of the elastograms was made by the contrast-to-noise ratio and the contrast transfer efficiency performance metrics, they found that the elastograms were statistically equivalent in both evaluators but at high modulus, contrast-to-noise ratio of elastograms by solving the inverse elasticity problem was superior.

Some studies to improve the displacement estimations are done [40, 69, 105] but the computational cost is high. Improvements in signal processing have been done, examples of them are the strain filter which is a nonlinear filtering process [95], the multidimensional autocorrelation method and the multidimensional Doppler method [85], the combined autocorrelation method [79], to mention some of them. In fact the world's first commercialized equipment for elastography, Hitachi EUB-8500 is based on the last method mentioned and it was released on the market in 2003 [78].

The uniqueness of the solition of the inverse problem in quasi-static elastic modulus imaging in two dimension was estudied by Barbone and Bamber [10]. This problem consists in the following given the equilibrium strain field in an incompressible elastic material, determine the shear modulus. They showed that in order to estimate the elastic parameters either the stress distribution or the elastic stiffness must be measured along a sufficient portion of the boundary, the knowledge of the displacement field everywhere and the displacement boundary conditions are not enough to determine the shear modulus. They also found that stress boundary conditions give more quantitative reconstructions than displacement boundary conditions.

Despite the disadvantages discussed here, it has been shown that this approach of elastography gives a relative good estimation of the elastic parameters [44] and it has been used for many applications, an example of this is the application of MRE to strain measurements in the carotid arteries and aortic wall strain measurements [46, 60].

10.5 Harmonic Elastography

Study of this elastography modality began with the ultrasound methods presented by Lerner et al. [51, 52] and Yamakoshi et al. [101]. This modality consists of a low-frequency acoustic wave that is induced in tissues through a sinusoidal mechanical

source, the displacements are observed by Doppler or MR imaging. In order to estimate the elastic parameters harmonic modality uses propagating mechanical waves rather than quasi-static stress to excite the tissue so it is not necessary knowledge of the static stress distribution.

MRE is the most studied approach to harmonic elastography, although is more expensive and has an acquisition time longer than ultrasound approaches. MRE has the advantage that it measures all three spatial componets of the tissue displacements with high accuracy, precision and resolution, while ultrasound approaches only measure the axial component of the displacements in a plane, and in some cases the lateral component with very poor accuracy. The displacement patterns corresponding to harmonic shear waves have amplitudes of microns or less and it is difficult for ultrasound methods to measure these small quantities; all these advantages give MRE a promising potential for a quantitative estimation of the elastic parameters [55]. As we mentioned before in magnetic resonance approaches there are two methods to measure the induced tissue motion, saturation tagging and phase contrast method.

From the phase and amplitude of the shear waves, an estimation of the shear modulus, μ, can be computed from following the relation between shear velocity (c_s), shear modulus and density of the tissue (ρ),

$$c_s = \sqrt{\frac{\mu}{\rho}}$$

The most used model for harmonic elastography is the system (10.5), nevertheless sometimes it is considered that longitudinal wave varies slowly so $\nabla \cdot \mathbf{u} = 0$, then the model simplifies to the Helmholtz equation,

$$\mu \Delta \mathbf{u} = -\rho \omega^2 \mathbf{u} \tag{10.7}$$

In 1995, Muthupillai et al. [59] were the first to present a method of magnetic resonance elasticity where a harmonic mechanical source was coupled to the surface to induce shear waves in the tissue. They used the phase contrast imaging method to measure displacement. Estimation of shear stiffness of liver tissue using this method was compared with the estimates in the literature with good results [49]. Also in [50] it is presented a study where estimates of the shear modulus of human cerebral tissue in vivo were made.

Sinkus et al. [80] presented also an approach to MRE using the phase contrast imaging method. They used a direct inversion scheme to estimate the shear modulus, they inverse a linear system of partial differential equations with regularization techniques. In this system the spatial derivatives of the displacement appear as coefficients. This technique is very sensitive to noise. For validation of the approach, finite element simulations and phantom experiments were performed [81].

Manduca et al. [56] presented an approach to MRE also using a phase-contrast MRI technique and a direct inversion scheme that they called algebraic inversion of differential equation (AIDE), the equation could be solved separately at each pixel using only data from a local neighborhood to estimating local derivatives. Because

of the big difference in magnitude of the Lamé parameters, it was considered the Eq. (10.7). Using AIDE μ was estimated from single polarization of motion by

$$\mu = \rho\omega \frac{u_i}{\Delta u_i}$$

Studies on phantoms and animal and human tissue using this technique were presented.

Park and Maniatty [66] proposed a shear modulus reconstruction for time harmonic excitation in a nonhomogeneous medium. This finite element based method is a direct inversion. They computed both the shear modulus and the hydrostatic stress distribution. No boundary conditions are required but it is required the derivatives of the displacements. They present some numerical examples of shear modulus reconstruction from MR measured data.

It is important to notice that both the direct inversion and the AIDE require data smoothing and the calculation of second derivatives from the noisy data.

Weaver et al. [97] presented an approach to MRE using the phase contrast imaging method and an iterative inversion technique to estimating the shear modulus where no displacements differentiation are required, but this approach has a high computational cost because they solved the three-dimensional inverse problem on a highly resolved finite element mesh.

Van Houten et al. [93] showed that the displacement by an oscillation excitation cannot be accurately characterized using two-dimensional approximation, without symmetries the three-dimensional case cannot be approximated by an ultrasound method. They presented finite element based method using an iterative inversion technique. In order to solve the full 3D elasticity problem at high resolution MR data, they divided the reconstruction field of view into a series of overlapping subzones and minimized over all the subzones an optimization operator. The subzones are deployed in a random and overlapping manner. This method is relatively slow because it is needed the full three-dimensional forward solution in each iteration and requires boundary condition for the forward solution step. This technique was parallelized by Doyley et al. [26].

Doyley et al. [28] showed that subzone based reconstruction algorithm increase linearly with the increases of subzones. Study in vivo of breast tissue was performed to compare the elastic parameters obtained by this technique and those reported in literature [94], with satisfactory results.

Doyley et al. [29] presented a study where it is evaluated the clinical efficacy of breast MRE under the assumption that soft tissues exhibit a Hookian material behavior. They used linear elastic MR reconstruction methods applied to phantoms that were fabricated using viscoelastic materials. They showed that although small viscoelastic inclusion can be detected, artifacts that degrade spatial and contrast resolution will be incurred when viscoelastic materials such as breast tissues are reconstructed using linear elastic reconstruction methods. They suggested that improvements in both the accuracy and quality of MR elastograms of the breast will be performed if other constitutive laws are considered.

Barbone and Gokhale [11] studied the uniqueness in two dimensions of the linear elastic form of the inverse problem for dynamic sinusoidal excitation cases, that is given determine the shear modulus given the displacement field in an incompressible linear elastic solid throughout a region. They presented the condition for uniqueness and give several examples of nonuniqueness. They found that with two displacement fields given, the shear modulus distribution is determined uniquely with four or fewer a priori known values of the shear modulus. With four different displacement fields, the shear modulus is determined uniquely up to a multiplicative constant.

Ammari et al. [8] presented stability results for the mathematical model for MRE, they considered the Stokes system for this analysis as a simplification of the time-harmonic system.

10.6 Transient Elastography

In this section we review the approaches to transient elastography, in this modality the mechanical excitation is produced by an impulsive or short time lasting burst. An interesting approach to assessing elasticity is to use the acoustic radiation force of an ultrasonic focused beam to remotely generate mechanical vibrations in organs [34]. The acoustic force is generated by the momentum transfer from the acoustic wave to the medium. The radiation force essentially acts as a dipolar source. See [68], Fig. 12 for a graphic representation of radiation force excitation.

Nightingale et al. [61] were the first using radiation force in medical images. They presented experimental result in breast tissue phantoms of the technique called remote palpation. A single transducer was used both to generate radiation force and to record the resulting tissue displacements. They found that the displacement images they obtained had higher contrast than the corresponding B-mode images. They presented the remote palpation method as similar to elastography (quasi-static elastography) but emphasized the advantage of the internal localized application of the excitation versus the global external compression and the real-time implementation.

Nightingale et al. [62] studied the liver fibrosis with a method also based on acoustic radiation, acoustic radiation force impulse (ARFI), that consists in focusing an ultrasonic beam deep in tissues for short durations, the resulting displacements at focus are measured by ultrasonic correlation-based techniques.

Sarvazyan et al. [75] proposed a method based on acoustic radiation force to generate shear waves, the shear wave elasticity imaging (SWEI). This technique consists in focusing an ultrasonic beam deep in tissues, the high attenuation of shear waves induces mechanical oscillations within a very limited area of tissue. Nightingale et al. [61] studied this technique experimentally. Application on liver [65], prostate [102], cardiac tissue [15] was performed.

Catheline et al. [18] devised a low frequency (50 Hz) external mechanical vibrator integrated with an ultrasound M-mode system. The Fibroscan manufactured by Echosense is based on this technique. The procedure with the Fibroscan takes around 3 min and has been very useful in studying liver diseases.

One of the disadvantages of the radiation force techniques is the weak effect in the induced displacements, these are of order of dozen of microns ($\sim$10 μm). To overcome this difficulty, Bercoff et al. [13] present an ultrasound-based technique that they called supersonic shear imaging (SSI), this technique generates two plane shear waves by the multiple ARFIs moving at a supersonic speed, the displacements are recorded by an ultra-fast scanner, 5000 frames/s. They reported with this technique mechanical displacements of order 100 μm in phantoms and 40 μm in vivo. Inversion algorithms are used to estimate the elastic parameter. A spatio-temporal sequence of the propagation of the induced transient wave can be acquired, leading to a quantitative estimation of the viscoelastic parameters of the studied medium in a source-free region [13, 14]. One of the strengths of SSI is the acquisition time, a mapping of the tissue elasticity is done in less than 30 ms. This technique has been used in experimental studies in breast [9, 88], muscle [22], liver [58]. This technique was implemented in the Aixplorer by Supersonic imagine.

Ammari et al. [2, 3, 33] presented a series of theoretical and numerical studies for the reconstruction of small anomalies. They designed an asymptotic imaging method leading to a quantitative estimation of physical and geometrical parameters of the anomaly. They used Helmholtz equation and quasi-incompressible elasticity model. Different algorithms were proposed to locate the anomaly such as Back-propagation, Kirchhoff, time-reversal, MUSIC. These algorithms are simple and do not require significant computational resources. In [4] it is presented topological derivative detection algorithm which is more robust with respect to the noise than the algorithms mentioned before, numerical examples are presented. The results were very promising.

In the recent book by Ammari et al. [6], they present a deep analysis exclusively of elastic and viscoelastic equation. They study the elasticity imaging of cracks and inclusions (small and extended anomalies) with boundary or internal data. They discuss the location algorithms mentioned before and also present a mathematical analysis of topological derivative based imaging and propose modifications of this algorithm in order to ameliorate its performance, they present numerical examples.

McLaughlin and Yoon [57] investigated the role of the boundary conditions in determining the uniqueness of the compressible form of the elastography inverse problem for transient approaches. They found that there exist at most one pair (ρ,μ) if μ is given on the boundary. This result is obtained with a single interior time dependent displacement measurement.

10.7 Models for Viscoelasticity, Related Techniques, and Future Direction

Elastography is an imaging technique that is continuously evolving. In the last sections we cover the elastography where the model for Hookian materials is used and the displacement measurements are made by ultrasound or magnetic resonance imaging. In this section we present a brief review of the different models, techniques to measure the displacement and other types of tissue excitation. We mention also some related techniques to elastography.

Petrov et al. [70] studied elastographic brain imaging in vivo, they proposed a parametric identification approach to estimate the elastic parameters. They used Rayleigh damping (RD) model, which is an alternative model for soft tissue attenuation for elastic materials. The results indicated limited applicability of the parametric RD model to accurately characterize viscoelastic properties of the brain tissue, nevertheless the values obtained are in agreement with literature.

Another model that has been considered to describe the viscoelastic properties of tissues is the Voigt model. Chateline et al. [19] have shown that this model is well adapted to describe the viscoelastic response of tissue to low frequency excitation. Ammari et al. [5] presented a mathematical analysis of time reversal algorithm for viscoelastic model and presented numerical results.

Bretin et al. [16] presented an explicit expression for the Green function in a viscoelastic medium. To model the viscoelastic properties they used the power law model described by Szabo and Wu [87], which is a generalization of the Voigt model. They presented numerical reconstruction of the Green function. Using this function, localization and estimation of the elastic parameters of a small anomaly in a visco-elastic medium can be performed following the work presented by Ammari et al. [2, 3, 33]

Optical coherence tomography (OCT) is an imaging technique that uses light to capture micrometer resolution. It uses a relatively long wavelength light that allows penetration into the tissue. Penetration in tissue of the optical wavelengths is limited but the resolution achieved is very high (sub-micrometer). Schmitt [77] presented a quasi-static elastography by OCT. Studies of OCT in the characterization of tissue have been done [45, 54]. Ammari et al. [7] presented a mathematical analysis of OCT in elastography using as model the Stoke system in heterogeneous medium. They presented a reconstruction algorithm and numerical experiments.

Wu et al. [99] presented a method called crawling wave imaging, where sonoelastography is used to image slowly moving interference patterns produced by two opposing shear vibration sources with a slight difference in frequency or phase. This technique has been used in vivo and ex vivo experiments of hepatic lesions and prostate [17, 104].

References

1. Adler, R.S., Barbosa, D.C., Cosgrove, D.O., Nassiri, D.K., Bamber, J.C., Hill, C.R.: Quantitative tissue motion analysis of digitized M-mode images: gestational differences of fetal lung. Ultrasound Med. Biol. **16**, 561–569 (1990)
2. Ammari, H., Garapon, P., Guadarrama Bustos, L., Kang, H.: Transient anomaly imaging by the acoustic radiation force. J. Differ. Equ. **249**, 1579–1595 (2010)
3. Ammari, H., Guadarrama Bustos, L., Kang, H., Lee, H.: Transient elasticity imaging and time reversal. Proc. R. Soc. Edinb. **141A**, 1–20 (2011)
4. Ammari, H., Bretin, E., Garnier, J., Jing, W., Kang, H., Wahab, A.: Localization, stability, and resolution of topological derivative based imaging functionals in elasticity. SIAM J. Imag. Sci. **6**(4), 2174–2212 (2013)
5. Ammari, H., Bretin, E., Garnier, J., Wahab, A.: Time-reversal algorithms in viscoelastic media. Eur. J. Appl. Math. **24**, 565–600 (2013)
6. Ammari, H., Bretin, E., Garnier, J., Kang, H., Lee, H., Wahab, H.: Mathematical Methods in Elasticity Imaging. Princeton University Press, Princeton (2015)
7. Ammari, H., Bretin, E., Millien, P., Seppecher, L., Seo, J.-K.: Mathematical modeling in full-field optical coherence elastography. SIAM J. Appl. Math. **75**(3), 1015–1030 (2015) http://epubs.siam.org/doi/abs/10.1137/140970409
8. Ammari, H., Waters, A., Zhang, H.: Stability analysis for magnetic resonance elastography. J. Math. Anal. Appl. **430**, 919–931 (2015)
9. Athanasiou, A., Tardivon, A., Tanter, M.I., SigalZafrani, B., Bercoff, J., Deffieux, T., Gennisson, J.L., Fink, M., Neuenschwander, S.: Breast lesions: quantitative elastography with supersonic shear imaging: preliminary results. Radiology **256**, 297–303 (2010)
10. Barbone, P.E., Bamber, J.C.: Quantitative elasticity imaging: what can and cannot be inferred from strain images. Phys. Med. Biol. **47**, 2147–2164 (2002)
11. Barbone, P.E., Gokhale, N.H.: Elastic modulus imaging: on the uniqueness and nonuniqueness of the elastography inverse problem in two dimensions. Inverse Probl. **20**, 283–296 (2004)
12. Birnholz, J.C., Farrell, E.E.: Fetal lung development: compressibility as a measure of maturity. Radiology **157**, 495–498 (1985)
13. Bercoff, J., Tanter, M., Fink, M.: Supersonic shear imaging: a new technique for soft tissue elasticity mapping. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 396–409 (2004)
14. Bercoff, J., Tanter, M., Muller, M., Fink, M.: The role of viscosity in the impulse diffraction field of elastic waves induced by the acoustic radiation force. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **51**, 1523–1536 (2004)
15. Bouchard, R.R., Hsu, S.J., Wolf, P.D., Trahey, G.E.: In vivo cardiac, acoustic-radiation-force-driven, shear wave velocimetry. Ultrason. Imaging **31**, 201–213 (2009)
16. Bretin, E., Guadarrama Bustos, L., Wahab, A.: On the Green function in visco-elastic media obeying a frequency power-law. Math. Methods Appl. Sci. **34**, 819–838 (2011)
17. Castaneda, B., Hoyt, K., Westesson, K., An, L., Baxter, L., Joseph, J., Strang, J., Rubens, D., Parker, K.: Performance of three-dimensional sonoelastography in prostate cancer detection: a comparison between ex vivo and in vivo experiments. In: IEEE Ultrasonics Symposium, 20–23 September 2009, Rome, pp. 519–522
18. Catheline, S., Wu, F., Fink, M.: A solution to diffraction biases in sonoelasticity: the acoustic impulse technique. J. Acoust. Soc. Am. **105**, 2941–2950 (1999)
19. Catheline, S., Gennisson, J.-L., Delon, G., Sinkus, R., Fink, M., Abouelkaram, S., Culioli, J.: Measurement of viscoelastic properties of solid using transient elastography: an inverse problem approach. J. Acoust. Soc. Am. **116**, 3734–3741 (2004)

20. Chenevert, T.L., Skovoroda, A.R., O'Donnell, M., Emelianov, S.Y.: Elasticity reconstructive imaging by means of stimulated echo MRI. Magn. Reson. Med. **39**, 482–490 (1998)
21. Cox, M., Rogers, P.H.: Automated noninvasive motion measurement of auditory organs in fish using ultrasound. J. Vib. Acoust. Stress Reliab. Des. **109**, 55–59 (1987)
22. Deffieux, T., Montaldo, G., Tanter, M., Fink, M.: Shear wave spectroscopy for in vivo quantification of human soft tissues visco-elasticity. IEEE Trans. Med. Imaging **28**, 313–322 (2009)
23. Dickinson, R.J., Hill, C.R.: Measurement of soft tissue motion using correlation betweenA-scans. Ultrasound Med. Biol. **8**, 263–271 (1982)
24. Doyley, M.M.: Model-based elastography: survey of approaches to the inverse elasticity problem. Phys. Med. Biol. **57**, R35–R37 (2012)
25. Doyley, M.M., Meaney, P.M., Bamber, J.C.: Evaluation of an iterative reconstruction method for quantitative elastography. Phys. Med. Biol. **45**, 1521–1540 (2000)
26. Doyley, M.M., Van Houten, E.E., Weaver, J.B., Poplack, S., Duncan, L., Kennedy, F., Paulsen, K.D.: Shear modulus estimation using parallelized partial volumetric reconstruction. IEEE Trans. Med. Imaging **23**, 1404–1416 (2004)
27. Doyley, M.M., Srinivasan, S., Pendergrass, S.A., Wu, Z., Ophir, J.: Comparative evaluation of strain-based and model-based modulus elastography. Ultrasound Med. Biol. **31**, 787–802 (2005)
28. Doyley, M.M., Feng, Q., Weaver, J.B., Paulsen, K.D.: Performance analysis of steady-state harmonic elastography. Phys. Med. Biol. **52**, 2657–2674 (2007)
29. Doyley, M.M., Perreard, I., Patterson, A.J., Weaver, J.B., Paulsen, K.M.: The performance of steady-state harmonic magnetic resonance elastography when applied to viscoelastic materials. Med. Phys. **37**(8), 3970–3979 (2010)
30. Fatemi, M., Greenleaf, J.F.: Ultrasound-stimulated vibro-acoustic spectrography. Science **280**, 82–85 (1998)
31. Fowlkes, J.B., Emelianov, S.Y., Pipe, J.G., Skovoroda, A.R., Carson, P.L., Adler, R.S., Sarvazyan, A.P.: Magnetic resonance imaging techniques for detection of elasticity variation. Med. Phys. **22**, 1771–1778 (1995)
32. Gao, L., Alam, S.K., Lerner, R.M., Parker, K.J.: Sonoelasticity imaging: theory and experimental verification. J. Acoust. Soc. Am. **97**, 3875–3886 (1995)
33. Gdoura, S., Guadarrama Bustos, L.: Transient wave imaging of anomalies: a numerical study. Contemp. Math. **548**, 31–44 (2011)
34. Greenleaf, J.F., Fatemi, M., Insana, M.: Selected methods for imaging elastic properties of biological tissues. Annu. Rev. Biomed. Eng. **5**, 57–78 (2003)
35. Harrigan, T.P., Konofagou, E.E.: Estimation of material elastic moduli in elastography: a local method, and an investigation of Poisson's ratio sensitivity J. Biomech. **37**(8), 1215–1221 (2004)
36. Holen, J., Waag, R., Gramiak, R.: Representation of rapidly oscillating structures on Doppler display. Ultrasound Med. Biol. **11**, 267–272 (1985)
37. Hoyt, K., Parker, K.J., Rubens, D.J.: Real-time shear velocity imaging using sonoelastographic techniques Ultrasound Med. Biol. **33**, 1086–1097 (2007)
38. Hoyt, K., Castaneda, B., Parker, K.J.: Two-dimensional sonoelastographic shear velocity imaging. Ultrasound Med. Biol. **34**, 276–288 (2008)
39. Huang, S.R., Lerner, R.M., Parker, K.J.: Time domain Doppler estimators of the amplitude of vibrating targets. J. Acoust. Soc. Am. **91**, 965–974 (1992)
40. Insana, M.F., Cook, L.T., Bigen, M., Chaturvede, P., Zhu, Y.: Maximum-likelihood approach to strain imaging using ultrasound. J. Acoust. Soc. Am. **107**, 1421–1434 (2000)

41. Jiang, J., Varghese, T., Brace, C.L., Madsen, E.L., Hall, T.J., Bharat, S., Hobson, M.A., Zagzebski, J. A., Lee Jr., F.T., Young's modulus reconstruction for radio-frequency ablation electrode-induced displacement fields: a feasibility study. IEEE Trans. Med. Imaging **28**, 1325–1334 (2009)

42. Kallel, F., Bertrand, M.: Tissue elasticity reconstruction using linear perturbation method. IEEE Trans. Med. Imaging **15**(3), 299–313 (1996)

43. Kallel, F., Ophir, J.: A least-squares strain estimator for elastography. Ultrason. Imaging **19**, 195–208 (1997)

44. Kallel, F., Ophir, J., Magee, K., Krouskop, T.: Elastographic imaging of low-contrast elastic modulus distributions in tissue. Ultrasound Med. Biol. **24**, 409–425 (1998)

45. Kennedy, B.F., Hillman, T.R., McLaughlin, R.A., Quirk, B.C., Sampson, D.D.: In vivo dynamic optical coherence elastography using a ring actuator. Opt. Express **17**, 21762–21772 (2009)

46. Kolipaka, A., Woodrum, D., Araoz, P.A., Ehman, R.L.: MR elastography of the in vivo abdominal aorta: a feasibility study for comparing aortic stiffness between hypertensives and normotensives. J. Magn. Reson. Imaging **35**, 582–586 (2012)

47. Konofagou, E., Dutta, P., Ophir, J., Cespedes, I.: Reduction of stress nonuniformities by apodization of compressor displacement in elastography. Ultrasound Med. Biol. **22**, 1229–1236 (1996)

48. Krouskop, T.A., Dougherty, D.R., Levinson, S.F.: A pulsed Doppler ultrasonic system for making noninvasive measurements of the mechanical properties of soft tissue. J. Rehabil. Res. Biol. **24**, 1–8 (1987)

49. Kruse, S.A., Smith, J.A., Lawrence, A.J., Dresner, M.A., Manduca, A., Greenleaf, J.F., Ehman, R.L.: Tissue characterization using magnetic resonance elastography: preliminary results. Phys. Med. Biol. **45**, 1579–1590 (2000)

50. Kruse, S.A., Rose, G.H., Glaser, K.J., Manduca, A., Felmlee, J.P., Jack, C.R., Ehman, R.L.: Magnetic resonance elastography of the brain. Neuroimage **39**, 231–237 (2008)

51. Lerner, R.M., Parker, K.J.: Sonoelasticity in ultrasonic tissue characterization and echographic imaging. In: Proceedings of the European Communities Workshop, Nijmegen, 7th October 1987

52. Lerner, R.M., Parker, K.J., Holen, J., Gramiak, R., Waag, R.C.: Sono-elasticity: medical elasticity images derived from ultrasound signals in mechanically vibrated targets. Acoust. Imaging **16**, 317–327 (1988)

53. Levinson, S.F., Shinagawa, M., Sato, T.: Sonoelastic determination of human skeletal muscle elasticity. J. Biomech. **28**, 11445–11454 (1995)

54. Liang, X., Adie, S.G., Renu, J.R., Boppart, S.A.: Dynamic spectral-domain optical coherence elastography for tissue characterization. Opt. Express **18**, 14183–14190 (2010)

55. Manduca, A., Dutt, V., Borup, D.T., Muthupillai, R., Greenleaf, J.F., Ehman, R.L.: An inverse approach to the calculation of elasticity maps for magnetic resonance elastography. Proc. SPIE **3338**, 426–436 (1998)

56. Manduca, A., Oliphant, T.E., Dresner, M.A., Mahowald, J.L., Kruse, S.A., Amromin, E., Felmlee, J.P., Greenleaf, J.F., Ehman, R.L.: Magnetic resonance elastography: non-invasive mapping of tissue elasticity. Med. Image. Anal. **5**, 237–254 (2001)

57. McLaughlin, J.R., Yoon, J.R.: Unique identifiability of elastic parameters from time-dependent interior displacement measurement. Inverse Probl. **20**, 25–45 (2004)

58. Muller, M., Gennisson, J.L., Deffieux, T., Tanter, M., Fink, M.: Quantitative viscoelasticity mapping of human liver using supersonic shear imaging: preliminary in vivo feasability study. Ultrasound Med. Biol. **35**, 219–229 (2009)

59. Muthupillai, R., Lomas, D.J., Rossman, P.J., Greenleaf, J.F., Manduca, A., Ehman, R.L.: Magnetic resonance elastography by direct visualization of propagating acoustic strain waves. Science **269**, 1854–1857 (1995)
60. Nederveen, A.J., Avril, S., Speelman, L.: MRI strain imaging of the carotid artery: present limitations and future challenges. J. Biomech. **47**, 824–833 (2014)
61. Nightingale, K.R., Soo, M.S., Nightingale, R., Trahey, G.: Acoustic radiation force impulse imaging: in vivo demonstration of clinical feasibility. Ultrasound Med. Biol. **28**, 227–235 (2002)
62. Nightingale, K.R., Zhai, L., Dahl, J.J., Frinkley, K.D., Palmeri, M.L.: Shear wave velocity estimation using acoustic radiation force impulsive excitation in liver in vivo. In: Proceedings of IEEE Ultrasonic Symposium, Vancouver, pp. 1156–1160 (2006)
63. Oberai, A.A., Gokhale, N.H., Feijoo, G.R.: Solution of inverse problems in elasticity imaging using the adjoint method. Inverse Probl. **19**, 297–313 (2003)
64. Ophir, J., Cespedes, I., Ponnekanti, H., Yazdi, Y., Li, X.: Elastography: a quantitative method for imaging the elasticity of biological tissues. Ultrason. Imaging **13**, 111–134 (1991)
65. Palmeri, P.L., Wang, M.H., Dahl, J.J., Frinkley, K.D., Nightingale, K.R.: Quantifying hepatic shear modulus in vivo using acoustic radiation force. Ultrasound Med. Biol. **34**, 546–558 (2008)
66. Park, E., Maniatty, A.M.: Shear modulus reconstruction in dynamic elastography: time harmonic case. Phys. Med. Biol. **51**, 3697–3721 (2006)
67. Parker, K.J., Lerner, R.M.: Sonoelasticity of organs: shear waves ring a bell. J. Ultrasound Med. **11**, 387–392 (1992)
68. Parker, K.J., Doyley, M.M., Rubens, D.J.: Imaging the elastic properties of tissue: the 20 year perspective. Phys. Med. Biol. **56**, R1–R29 (2011)
69. Pellot-Barakat, C., Mai, J.J., Kargel, C., Hermen, A., Trummer, B., Insana, M.F.: Accelerating ultrasonic strain reconstructions by introducing mechanical constraints. Proc. SPIE **4684**, 323–333 (2002)
70. Petrov, A.Y., Sellier, M., Docherty, P.D., Chase, J.G.: Parametric-based brain Magnetic Resonance Elastography using a Rayleigh damping material model. Comput. Methods Programs Biomed. **116**(3), 328–339 (2014)
71. Plewes, D.B., Betty, I., Urchuk, S.U., Soutar, I.: Visualizing tissue compliance with MR imaging. J. Magn. Reson. Imaging **5**, 733–738 (1995)
72. Ponnekanti, H., Ophir, J., Cespedes, I.: Ultrasonic-imaging of the stress-distribution in elastic media due to an external compressor. Ultrasound Med. Biol. **20**, 27–33 (1994)
73. Raghavan, K.R., Yagle, A.E.: Forward and inverse problems in elasticity imaging of soft-tissues. IEEE Trans. Nucl. Sci. **41**, 1639–1648 (1994)
74. Richards, M.S., Barbone, P.E., Oberai, A.A.: Quantitative three-dimensional elasticity imaging from quasi-static deformation: a phantom study. Phys. Med. Biol. **54**, 757–779 (2009)
75. Sarvazyan, A.P., Rudenko, O.V., Swanson, S.D., Fowlkes, J.B., Emelianov, S.Y.: Shear wave elasticity imaging—a new ultrasonic technology of medical diagnostic. Ultrasound Med. Biol. **20**, 1419–1436 (1998)
76. Sarvazyan, A.P., Urban, M.W., Greenleaf, J.F.: Acoustic waves in medical imaging and diagnostics. Ultrasound Med. Biol. **39**(7), 1133–1146 (2013)
77. Schmitt, J.M.: OCT elastography: imaging microscopic deformation and strain of tissue. Opt. Express **3**, 199–211 (1998)
78. Shiina, T.: Ultrasound elastography: development of novel technologies and standardization. Jpn. J. Appl. Phys. **53**, 07KA02 (2014)
79. Shiina, T., Nitta, N., Ueno, E., Bamber, J.C.: Real time elasticity imaging using the combined autocorrelation method. J. Med. Ultrason. **29**, 119–128 (2002)

80. Sinkus, R., Lorenzen, J., Schrader, D., Lorenzen, M., Dargatz, M., Holz, D.: High-resolution tensor MR elastography for breast tumor detection. Phys. Med. Biol. **45**, 1649–1664 (2000)

81. Sinkus, R., Tanter, M., Xydeas, T., Catheline, S., Bercoff, J., Fink, M.: Viscoelastic shear properties of in vivo breast lesions measured by MR elastography. Magn. Reson. Imaging **23**, 159–165 (2005)

82. Skovoroda, A.R., Emelianov, S.Y., Lubinski, M.A., Sarvazyan, A.P., O'Donnell, M.: Theoretical analysis and verification of ultrasound displacement and strain imaging. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **41**, 302–313 (1994)

83. Skovoroda, A.R., Emelianov, S.Y., O'Donnell, M.: Tissue elasticity reconstruction based on ultrasonic displacement and strain images. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **42**, 747–765 (1995)

84. Sumi, C.: Spatially variant regularization for tissue strain measurement and shear modulus reconstruction. J. Med. Ultrason. **34**, 125–131 (2007)

85. Sumi, C.: Displacement vector measurement using instantaneous ultrasound signal phase-multidimensional autocorrelation and Doppler methods. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **55**, 24–43 (2008)

86. Sumi, C., Suzuki, A., Nakayama, K.: Phantom experiment on estimation of shear modulus distribution in softtissue from ultrasonic measurement of displacement vector field. IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **78**, 1655–1664 (1995)

87. Szabo, T.L., Wu, J.: A model for longitudinal and shear wave propagation in viscoelastic media. J. Acous. Soc. Am. **107**, 2437–2446 (2000)

88. Tanter, M., Bercoff, J., Athanasiou, A., Deffieux, T., Gennisson, J.L., Montaldo, G., Muller, M., Tardivon, A., Fink, M.: Quantitative assessment of breast lesion viscoelasticity: initial clinical results using supersonic shear imaging. Ultrasound Med. Biol. **34**, 1373–1386 (2008)

89. Taylor, L.S., Porter, B.C., Rubens, D.J., Parker, K.J.: Three-dimensional sonoelastography: principles and practices. Phys. Med. Biol. **45**, 1477–1494 (2000)

90. Timoshenko, S.P., Goodier, J.N.: Theory of Elasticity. McGraw-Hill, New York (1970)

91. Tristam, M., Barbosa, D.C., Crosgrove, D.O., Nassire, D.K., Bamber, J.C., Hill, C.R.: Ultrasonic study of in vivo kinetic characteristics of human tissue. Ultrasound Med. Biol. **12**, 927–937 (1986)

92. Van Houten, E.E.W., Paulsen, K.D., Miga, M.I., Kennedy, F.E., Weaver, J.B.: An overlapping subzone technique for MR-based elastic property reconstruction. Mag. Reson. Med. **42**, 779–786 (1999)

93. Van Houten, E.E.W., Miga, M.I., Weaver, J.B., Kennedy, F.E., Paulsen, K.D.: Three-dimensional subzone-based reconstruction algorithm for MR elastography. Magn. Reson. Med. **45**, 827–837 (2001)

94. Van Houten, E.E.W., Doyley, M.M., Kennedy, F.E., Weaver, J.B., Paulsen, K.D.: Initial in vivo experience with steady-state subzone-based MR elastography of the human breast. J. Magn. Reson. Imaging **17**, 72–85 (2003)

95. Varghese, T., Ophir, J.: A theoretical framework for performance characterization of elastography: the strain filter. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **44**, 164–172 (1997)

96. Vogel, C.: Computational Methods for Inverse Problems. SIAM, Philadelphia (2002)

97. Weaver, J.B., Van Houten, E.E.W., Miga, M.I., Kennedy, F.E., Paulsen, K.D., Magnetic resonance elastography using 3D gradient echo measurements of steady-state motion. Med. Phys. **28**, 1620–1628 (2001)

98. Wilson, L.S., Robinson, D.E.: Ultrasonic measurement of small displacements and deformation tissue. Ultrason. Imaging **4**, 71–82 (1982)

99. Wu, Z., Taylor, L.S., Rubens, D.J., Parker, K.J.: Sonoelastographic imaging of interference patterns for estimation of the shear velocity of homogenous materials. Phys. Med. Biol. **49**, 911–922 (2004)

100. Wu, Z., Hoyt, K., Rubens, D.J., Parker, K.J.: Sonoelastographic imaging of interference patterns for estimation of shear velocity distribution in biomaterials. J. Acoust. Soc. Am. **120**, 535–545 (2006)

101. Yamakoshi, Y., Sato, J., Sato, T.: Ultrasonic-imaging of internal vibration of soft-tissue under force vibration. IEEE Trans. Ultrason. Ferroelectr. Freq. Control **37**, 45–57 (1990)

102. Zhai, L., Madden, J., Mouraviev, V., Polascik, T., Nightingale, K.: Correlation between SWEI and ARFI image findings in ex vivo human prostates. In: IEEE Ultrasonics Symposium, Rome, pp. 523–526 (2009)

103. Zhang, Y., Hall, L.O., Goldgof, D.B., Sarkar, S.: A constrained genetic approach for computing material property of elastic objects. IEEE Trans. Evol. Comput. **10**, 341–357 (2006)

104. Zhang, M., Castaneda, B., Wu, Z., Nigwekar, P., Joseph, J.V., Rubens, D.J., Parker, K.J.: Congruence of imaging estimators and mechanical measurements of viscoelastic properties of soft tissues. Ultrasound Med. Biol. **33**, 1617–1631 (2007)

105. Zhu, Y., Chaturvedi, P., Insana, M.F.: Strain imaging with a deformable mesh. Ultrason. Imaging **20**, 127–146 (1999)

Chapter 11
Affine Complete Algebras

Gérard Kientega

Abstract We study affine completeness of algebras. In the first part of the work, we use a generalized metric to prove an extension theorem. This extension theorem plays a key role in proving new results. In the second part, we show that in some situations, we can skip the extension theorem. This idea allows us to answer a question of Karrli and Pixley.

Keywords Algebra • Compatible operation • Affine completeness • Generalized metric • Congruence lattice • Torsion free module

11.1 Introduction

Jawhari, Misane, and Pouzet [1] introduced the notion of a generalized metric and found an extension theorem applicable in languages, ordered sets, and graph theory. However, it needs hyperconvexity, a very restrictive hypothesis and in many cases infinite family of balls are used. Although that generalized metric spaces are not essentially of topological nature certain similarities with the usual metric spaces are useful. Kaarli [2] gave an extension theorem based on the arithmetical properties of the lattice of the equivalence relations and a new affine completeness result. Following his ideas we define a generalized metric applicable in universal algebra. This leads to new proofs of some affine completeness results. In the last part of this work, we give an answer to a problem raised by Kaarli and Pixley [3]. For this purpose we skip using the extension theorem that was useful in other contexts.

Definition 1. Let $V = \langle \mathrm{V}; +, 0, ^{-}, \leq \rangle$ where

(i) $\langle \mathrm{V}; +, 0 \rangle$ is a monoïd (not necessarily commutative), with the neutral element 0,

G. Kientega (✉)
UFR des Sciences exactes et appliquées, Université de Ouagadougou, 03 B.P. 7021, Ouagadougou 03, Burkina Faso
e-mail: gkienteg@ictp.it; kientdia@yahoo.fr

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_11

(ii) $\leq$ is a (partial) order on V compatible with $+$ (i.e., $p + r \leq q + r$ and $r + p \leq r + q$ whenever $p \leq q$) such that 0 is the least element of V (i.e., $0 \leq p$ for all $p \in V$), and
(iii) $p \to \overline{p}$ is an involutive automorphism of the order $\leq$ (i.e., $\overline{\overline{p}} = p$ and $\overline{p} \leq \overline{q}$ whenever $p \leq q$).

A *V-metric space* or a *V-metric* is a pair (E,d) where E is a set and $d: E^2 \to V$ is a map such that for all $x, y, z \in E$:

$$(d1)\ d(x, y) = 0 \iff x = y,$$

$$(d2)\ \overline{d(x, y)} = d(y, x),$$

$$(d3)\ d(x, y) \leq d(x, z) + d(z, y)\ (\textit{the triangle inequality}).$$

We call *symmetric* a V-metric provided $d(x,y) = d(y,x)$ for all $x,y \in E$. Notice that the usual metric space is a symmetric V-metric space for $V = \langle R_+; +, 0, -, \leq \rangle$ where R_+ is the set of nonnegative reals with the usual sum and order and $\overline{r} = r$ for all $r \in R_+$. For another natural example consider $V = \langle V, o, \Delta, {}^{-1}, \subseteq \rangle$ where V is a set of binary reflexive relations on a set E such that:

(1) V is closed under arbitrary intersections, contains $\Delta = \{(x,x) : x \in E\}$ and whose union is E^2,
(2) ros is the relational product $\{(x,y) \in V : (x,z) \in r, (z,y) \in s \text{ for some } z \in V\}$, r^{-1} is the converse $\{(y,x) : (x,y) \in r\}$, and $\subseteq$ is the set theoretical intersection, and
(3) V is closed under o and $^{-1}$.

Then define $d: E^2 \to V$ by setting $d(x,y) = \cap \{r \in V : (x,y) \in r\}$.

In the sequel V will be fixed. Let (E,d) be a V-metric space. By a *V-ball* with *radius* r and *center* x we mean the set $B(x,r) = \{y \in E : d(x,y) \leq r\}$. Let (E',d') be also a V-metric space. A function $f: E' \to E$ is *nonexpansive* if $d(f(x), f(y)) \leq d'(x,y)$ for each pair (x,y) in E^2. The V-metric space (E,d) is *convex* if for any x, y in E and r, s in V, the relation $d(x,y) \leq r + s$ implies the existence of z in E such that $d(x,z) \leq r$ and $d(z,y) \leq s$. The *2-Helly property* (respectively, the *2-Helly property for finite families of balls*) means that each family (respectively, finite family) of V-balls has a nonempty intersection as soon as any pair of them has a nonempty intersection [1] §II.2. Let F be a subset of E' and $f: F \to E$ be nonexpansive. The function f is said to have the *one-point extension property* if for each $x \in E' \backslash F$ there is a nonexpansive extension of f on $F \cup \{x\}$. When the last property is satisfied for all subsets F of E', all V-metric spaces (E',d'), and all nonexpansive $f: F \to E$, the space (E,d) is *called a hyperconvex space*. The V-metric space (E,d) has the *finite extension property* if for every V-metric space (E',d') and each finite subset F of E', every nonexpansive function from (F,d') to (E,d) has the one-point extension property. According to [1] § II.2 hyperconvexity is equivalent to convexity and the 2-Helly property.

11.2 The Finite Extension Theorem

In this section the order of V is a join semilattice on V with a least element 0 and whose join operation is denoted $\vee$. Now, let (E,d) be a V-metric space. The join operation is not necessarily the addition $+$ of V but we always have $x \vee y \leq x + y$ for all $x, y \in E$ since $x + y \geq x + 0 = x$ and similarly $x + y \geq y$. For each natural number $n \geq 1$, define a V-metric d_n on E^n by the formula

$$d_n((x_1, \ldots, x_n), (y_1, \ldots, y_n)) = d(x_1, y_1) \vee \cdots \vee d(x_n, y_n).$$

Definition 2. Let E be a set and X a subset of E. An *a-function* on X is a ternary function $f_X: X^3 \to E$ such that for all x, y in X, $f_X(x,y,y) = f_X(x,y,x) = f_X(y,y,x) = x$. An *m-function* on X is a ternary function $m_X: X^3 \to E$ such that for all x,y in X $m_X(x,x,y) = m_X(x,y,x) = m_X(y,x,x) = x$. The V-metric space (E,d) has the *a-property* (respectively, the *m-property*) if for each finite subset X of E there exists a nonexpansive a-function (respectively, a nonexpansive m-function) f: $(X^3, d_3) \to (E,d)$ on X.

A symmetric V-metric space (E,d) is *ultrametric* if for all x,y, and z in E we have

$$(*) \quad d(x,y) \leq d(x,z) \vee d(z,y).$$

In this case (E,d) is also a V'-metric space where $V' = \langle V; \vee, 0, id_V, \leq \rangle$ and the triangle inequality is the above inequality.

Remark 1. Let E be a set and let Eqv(E) denote the set of equivalence relations on E. Let V be a subset of Eqv(E) such that

(i) $\Delta = \{(x,x) : x \in E\} \in V$,
(ii) V is closed under arbitrary intersections and the join $\vee$ of equivalence relations, and
(iii) $E^2 = \cup V$.

Now let $V = \langle V; \vee, \Delta, id_V, \subseteq \rangle$ and define $d: E^2 \to V$ by setting $d(x,y) = \cap \{r \in V : (x,y) \in r\}$ for all $x,y \in E$. Clearly the V-metric space $V_{eq} = (E,d)$, which is implicit in Kaarli [3], is ultrametric.

Lemma 2.1. *Let* (E,d) *be a V-metric space. If for each subset* X of E *such* that $|X| = \min(3,|E|)$ *there exists a nonexpansive* a-*function on* X, *then* (E,d) *is an ultrametric. In particular,* (E,d) *is an ultrametric if it has the a-property.*

Proof. Since the case $|E| = 1$ is trivial, we may suppose that $|E| \geq 2$. Let x, y, z be elements of E, X consists of x,y,z, and let f_X be a nonexpansive a-function on X. Then

$$d(x,y) = d(f_X(y,y,x), f_X(y,x,x)) \leq d_3\Big((y,y,x), \Big(y,x,x\Big)\Big). =$$
$$d(y,y) \vee d(y,x) \vee d(x,x) = d(y,x)$$

showing that $d(x,y) \leq d(y,x)$. By symmetry $d(y,x) \leq d(x,y)$ and thus, $d(x,y) = d(y,x)$ proving the symmetry. For the inequality (*)

$$d(x,y) = d(f_X(z,z,x), f_X(y,x,x)) \leq d_3\left((z,z,x), \left(y,x,x\right)\right) = d(z,y) \vee d(z,x) \vee d(x,x) = d(x,z) \vee d(z,y). \quad \square$$

Let E^n denote the n^{th} power of E and set

$$D_n = \{(x_1, \ldots, x_n) \in E^n : x_1 = \cdots = x_{i-1} = x_{i+1} = x_n \quad \text{for some } 1 \leq i \leq n-1\}$$

Lemma 2.2. *Let* (E, d) *be a V-metric. If* (E,d) is *ultrametric, then the ternary function* $f: D_3 \to E$ *defined by setting* $f(x,x,y) = f(y,x,y) = f(y,x,x) = y$ *for all* x,y *in* E *is a nonexpansive map from* (D_3,d_3) *into* (E,d).

Proof. Let $\overline{x} = (x_1, x_2, x_3)$ and $\overline{y} = (y_1,y_2,y_3)$ be two elements of D_3. We have two cases.

(1) Suppose $f(\overline{x}) = x_i$ and $f(\overline{y}) = y_i$ for some $i, 1 \leq i \leq 3$.
Then $d(f(\overline{x}), f(\overline{y})) = d(x_i,y_i) \leq d_3(\overline{x}, \overline{y}) = d(x_1,y_1) \vee d(x_2,y_2) \vee d(x_3,y_3)$.
(2) Suppose $x_i = x_j$ and $y_j = y_k$ where $\{i,j,k\} = \{1,2,3\}$.
Then $f(\overline{x}) = x_k$ and $f(\overline{y}) = y_i$ and $d(f(\overline{x}),f(\overline{y})) = d(x_k,y_i) \leq d(x_k,y_k) \vee d(y_k,y_i) = d(x_k,y_k) \vee d(y_j,y_i) \leq d(x_k,y_k) \vee d(y_j,x_j) \vee d(x_j,y_i) = d(x_k,y_k) \vee d(x_j,y_j) \vee d(x_i,y_i) = d_3(\overline{x}, \overline{y})$ using the symmetry and the fact that $x_i = x_j$ and $y_j = y_k$. This shows that f is nonexpansive. $\square$

Theorem 2.3. (The Finite Extension Theorem): *Let the order* $\leq$ of V *be a join semilattice such that 0 is the least element of* $\leq$. *The following conditions* (i)–(iii) *are equivalent for a V-metric space* (E,d):

(i) *The space* (E,d) *has the finite extension property,*
(ii) *The space* (E,d) *is convex and has the m-property*,
(iii) *The space* (E,d) *is convex and possesses the* 2*-Helly property for finite families of V-balls. Moreover, if* (E,d) *is an ultrametric space, then the conditions* (i)–(iii) *are equivalent to*
(iv) *The space* (E,d) *is convex and has the* a*-property*.

Proof. (i) $\Rightarrow$ (ii) Suppose that (E,d) has the finite extension property. The convexity follows easily by the same argument as in [1] Thm II.2.1. Now, let X be a finite subset of E. On $Y = D_3 \cap X^3$ define the function f by setting $f(x,x,y) = f(x,y,x) = f(y,x,x) = x$ for all x, y in X. We show that f is a nonexpansive map from (Y,d_3) to (E,d). Let $\overline{x} = (x_1, x_2, x_3)$ and $\overline{y} = (y_1,y_2,y_3)$ be arbitrary elements of Y. Then $f(\overline{x}) = x_i$ and $f(\overline{y}) = y_i$ for some $1 \leq i \leq 3$. Furthermore $d(f(\overline{x}),f(\overline{y})) = d(x_i, y_i) \leq d(x_1,y_1) \vee d(x_2,y_2) \vee d(x_3,y_3) = d_3(\overline{x}, \overline{y})$.

Then, by the finite extension property there exists a nonexpansive extension m of f on X^3. Clearly the function m is an m-function on X.

(ii) $\Rightarrow$ (iii) Suppose that (E,d) is convex and has the m-property. By induction on $n = 2,3,\ldots$ we prove the following statement S_n: Let $\{B(x_i,r_i)\colon 1 \le i \le n\}$ be a set of V-balls satisfying the condition: if $B(x_i,r_i) \cap B(x_j,r_j) \neq \emptyset$ for all $1 \le i < j \le n$, then $B(x_1,r_1) \cap \ldots \cap B(x_n,r_n) \neq \emptyset$. Since S_2 is true, we suppose that S_n holds for some $n \ge 2$. Let us prove the result for $n+1$. Let $\{B(x_1,r_1),\ldots,B(x_{n+1},r_{n+1})\}$ be a set of pairwise intersecting V-balls on E. For $j = 1,2,3$ there exists $y_j \in B(x_1,r_1) \cap \ldots \cap B(x_{j-1},r_{j-1}) \cap B(x_{j+1},r_{j+1}) \cap \ldots \cap B(x_{n+1},r_{n+1})$.

Set $X = \{x_1,\ldots,x_{n+1},y_1,y_2,y_3\}$. By the m-property there exists a nonexpansive m-function m from (X^3,d_3) into (E,d). We show that $u = m(y_1,y_2,y_3)$ is an element of the intersection of all the $n+1$ balls. We have $d(x_1,u) = d(m(y_1,x_1,x_1)),m(y_1,y_2,y_3)) \le d_3((y_1,x_1,x_1),(y_1,y_2,y_3)) = d(y_1,y_1) \vee d(x_1,y_2) \vee d(x_1,y_3) \le r_1 \vee r_1 = r_1$ and by symmetry $d(x_2,u) \le r_2$ and $d(x_3,u) \le r_3$. Now for $j > 3$, $d(x_j,u) = d(m(x_j,x_j,x_j), m(y_1,y_2,y_3)) \le d_3((x_j,x_j,x_j),(y_1,y_2,y_3)) = d(x_j,y_1) \vee d(x_j,y_2) \vee d(x_j,y_3) \le r_j$ since each y_k for $k = 1,2,3$ belongs to the ball of center x_j and radius r_j. This proves that u is in the intersection of the $n+1$ balls and (iii).

(iii) $\Rightarrow$ (i) Suppose that (iii) holds for a V-metric space (E,d) and let (E',d') be a V-metric space. We prove (i) by induction on the size n of a finite subset X of E'.

1. Let $n = 1$; i.e., $X = \{x\}$. Let $f\colon\{x\} \to E$. Clearly the constant map $f_1\colon E' \to E$ with value f(x) is nonexpansive and so the statement holds for $n = 1$.
2. Suppose that (i) holds for some $n \ge 1$ such that $n+1 < |E'|$. Let $\{x_1,\ldots,x_{n+2}\} \subseteq E'$ and let g be a nonexpansive map from $(\{x_1,\ldots,x_{n+1}\}, d')$ to (E,d). For $i = 1,\ldots,n+1$ set $x_i' = g(x_i)$, $r_i = d'(x_i, x_{n+2})$ and denote the balls $B(x_i',r_i)$ (of (E,d)) by B_i. For $1 \le i < j \le n+1$ the balls B_i and B_j intersect. Indeed $d(x_i',x_j') \le d'(x_i,x_j) \le d'(x_i,x_{n+2}) + d'(x_{n+2},x_j) = r_i + \overline{r_j}$ and so by convexity there exists $u \in E$ with $d(x_i',u) \le r_i$, $d(u, x_j') \le \overline{r_j}$ proving $u \in B_i \cap B_j$. By the 2-Helly property there exists $v \in B_1 \cap \ldots B_{n+1}$. Extend g to $g_1\colon \{x_1,\ldots,x_{n+2}\} \to E$ by setting $g_1(x_{n+2}) = v$. For $i = 1,\ldots,n+1$ from $v \in B_1 \cap \ldots B_{n+1}$ and the definition of r_i clearly $d(g_1(x_i), g_1(x_{n+2})) = d(x_i',v) \le r_i = d'(x_i, x_{n+2})$ proving that g_1 is indeed nonexpansive. This concludes the induction and hence (i) holds.

Let the space (E,d) be ultrametric. By Lemma 2.2 clearly (i) $\Rightarrow$ (iv). We prove that (iv) $\Rightarrow$ (ii). Let X be a finite subset of E and let f_X be an a-function on X. We define $Y = \text{im}\,(f_X)$. Then $X \subseteq Y$ and Y is finite. The function m defined on Y^3 by $m(x,y,z) = f_X(x,f_X,(x,y,z),z)$ is an m-function on Y and hence also on X.

The last theorem is a generalization of a result of Kaarli [2] Theorem 3. In fact, Kaarli's proof can be used when the metric (E,d) is symmetric and the semilattice $\langle V;\le\rangle$ has a special structure as is shown in the following proposition. Call an element r of V *idempotent* if $r + r = r$.

Proposition 2.4. *Let the order of V be a meet semilattice whose meet is denoted* $\wedge$. *Suppose that for all* that u, v, *and* w *in* V *we have* $u \wedge (v + w) = (u \wedge v) + (u \wedge w)$ (i.e., $\wedge$ *distributes over* $+$). *If* (E,d) *is a symmetric and convex V-*metric *space, then it*

possesses the 2*-Helly property for finite families of balls whose radii are idempotent. In particular, if all elements of* V *are idempotent, then* (E,d) *has the finite extension property*.

Proof. A direct adaptation of the proof of the Theorem 3 of Kaarli [2].

Definition 3. A *V*-metric space (E,d) is *quasi-compact* if every family of V-balls with empty intersection contains a finite subfamily with empty intersection.

We are now able to characterize hyperconvex spaces.

Theorem 2.5. *A V-metric space is hyperconvex if and only if it is quasi-compact and has the finite extension property.*

Proof. ($\Rightarrow$) Clear.

($\Leftarrow$) Let (E,d) be a quasi-compact *V*-metric space with the finite extension property. To show that it is hyperconvex, let (E',d') be a *V*-metric and let f: (F,d') $\rightarrow$ (E,d) a nonexpansive map where F is a subset of E'. Consider the set $\Im$ of pairs (X,g) where each X is a subset of E' containing F and g is a nonexpansive map (X,d') $\rightarrow$ (E,d) which is an extension of f. Clearly $\Im$ is nonempty. The set $\Im$ is (partially) ordered by setting (X,g) $\leq$ (X',g') if X $\subseteq$ X' and g' is an extension of g to X'. We show that the order $\leq$ is inductive. Let T = $\{(X_i,g_i); i \in I\}$ be a totally ordered subset of $\Im$. Set X = $\cup$ $\{X_i : i \in I\}$ and define g from (X,d') to (E,d) as the map whose restriction to each X_i coincides with g_i. Clearly (X,g) is the least upper bound of T in $\Im$. By Zorn's lemma there exists a maximal element (Y,h) of $\Im$. We show that Y = E'. If not, choose x in E'\Y and consider the set $\{B(h(y),r_y): y \in Y\}$ where $r_y = d'(y,x)$. Let $Y_k = \{y_1, \ldots, y_k\}$ be a finite subfamily of Y and let h_k be the restriction of h to Y_k. By the finite extension property there exists a nonexpansive extension $\overline{h}_k$ of h to $Y_k \cup \{x\}$. Clearly for j = 1, ...,k, d(f(y_j), $\overline{h}_k$ (x)) = d($\overline{h}_k$ (y_j), $\overline{h}_k$ (x)) $\leq r_j$. This shows that $\overline{h}_k$ (x) is in the intersection of the balls $\{B(h(y_j), r_{y_j}); j = 1, \ldots, k\}$. Thus any finite subset of the *V*-balls $\{B(h(y),r_y); y \in Y\}$ has a nonempty intersection. By the quasi-compactness of the space (E,d) the V-balls $\{B(h(y),r_y); y \in Y\}$ have a nonempty intersection. Let a be an element of this intersection. Clearly we may extend the map h to a nonexpansive map h': (Y $\cup\{x\}$,d') $\rightarrow$ (E,d) by setting h'(x) = a and h'(y) = h(y) for each y in Y. The pair (Y $\cup\{x\}$, h') is a strict majorant of (Y,d) in $\Im$ which is a contradiction. Hence Y = E' and by [1] Theorem II.2.1 clearly (E,d) is an hyperconvex space.

11.3 Generalized Metrics and Equivalence Relations

Let d be a *V*-metric on a set E. For each v $\in$ V we define the binary relation $(d)_v = \{(x,y) \in E^2 : d(x,y) \leq v\}$. We recall that for n $\geq$ 1 an n-*ary operation* on E is a function f: $E^n \rightarrow$ E. For n = 0 a 0-ary operation on E is a constant of E.

Definition 4. A (*nonindexed universal*) *algebra* on E is a pair $A = \langle E,C\rangle$ where C is a nonempty set of operations on E.

The elements of C are the *fundamental operations* of the algebra A. A subset F of E is *closed* with respect to the n-ary operation p of A if for all $x_1, \ldots, x_n \in F$ we have that $p(x_1, \ldots, x_n) \in F$. In the event that p is a constant, this means that F is closed with respect to p if and only if $p \in F$

Definition 5. A *subuniverse* of an algebra A on E is a nonempty subset S of E closed with respect to each fundamental operation of the algebra A. A *subalgebra* of the algebra $A = \langle E,C\rangle$ is an algebra $B = \langle S,C'\rangle$ where S is a subuniverse of A and C' consists of the restrictions of the functions of C to S.

Let $m \geq 1$ and let p be an n-ary fundamental operation of $A = \langle E,C\rangle$. We define an n-ary operation p_m on E^m by setting for all $a_1 = (a_{11}, \ldots, a_{1m}), \ldots, a_n = (a_{n1}, \ldots, a_{nm}) \in E^m$, $p_n(a_1, \ldots, a_n) = (p(a_{11}, \ldots, a_{n1}), \ldots, (a_{1m,\ldots,} a_{mn}))$.

Let $C_m = \{p_m;\ p \in C\}$. By the m-th power A^m of A we mean the algebra $\langle E^m, C_m\rangle$.

Definition 6. For $n \geq 1$ an n-ary *relation* on E is a nonempty subset of E^n. The n-ary relation r is *compatible* with the operation f if r is a subuniverse of $\langle E^n,\{f\}\rangle$.

We need the following useful result.

Theorem 3.1. (Pouzet–Rosenberg [6], Lemma I-3.) *Let C be a nonempty set of operations on E. For each V-metric space (E,d) the following two propositions are equivalent:*

1. *For every* $v \in V$ *the binary relation* $(d)_v$ is *a subuniverse of* A^2;
2. *For all* $n \geq 1$ *each n-ary operation from* C *is a nonexpansive map from* (E,d_n) *into* (E,d).

Of particular interest is the next corollary.

Corollary 2. Let *the order of* V *be a join semilattice, let* (E,d) *be a V-metric space, and let* $X \subseteq E^n$ *where* $n \geq 1$. Suppose *that the metric of* E^n *is* d_n. *A map f from* (X,d_n) *into* (E,d) is *nonexpansive* if *and only if for each* $v \in V$ *the function* f *is compatible with the binary relation* $(d)_v$.

Proof. ($\Leftarrow$) Suppose that f is compatible with the binary relation $(d)_v$ for each v in V. Let $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$ be elements of X. Set $v = d_n(x,y) = d(x_1,y_1) \vee \ldots \vee d(x_n,y_n)$. Then $(x_i,y_i) \in (d)_v$ for $i = 1, \ldots, n$. Since f is compatible with $(d)_v$, clearly $(f(x), f(y)) \in (d)_v$, that is, $d(f(x),f(y)) \leq v = d_n(x,y)$ proving that f is nonexpansive.

($\Rightarrow$) Suppose that f is nonexpansive and let $v \in V$. Let $(x_i,y_i) \in (d)_v$ and for $i = 1, \ldots, n$, set $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots y_n)$. Since f is nonexpansive, we have

$$d(f(x), f(y)) \leq d_n(x, y) = d(x_1, y_1) \vee \cdots \vee d(x_n, y_n) \leq v.$$

Thus $(f(x),f(y)) \in (d)_v$. □

Lemma 3.3. *Let* (E,d) *be a V-metric space and* (u,v) in V^2. *If the metric* d *is convex, we have* $(d)_u o (d)_v = (d)_{u+v}$. *If it is also symmetric, then* $(d)_v o (d)_u = (d)_u o (d)_v$.

Proof. We prove that $(d)_u o(d)_v = (d)_{u+v}$ when d is convex. Suppose $(x,y) \in (d)_{u+v}$, that is, $d(x,y) \leq u + v$. Then by convexity there exists z in E such that $d(x,z) \leq u$ and $d(z,y) \leq v$. We thus have $(x,z) \in (d)_u$ and $(z,y) \in (d)_v$ proving that $(x,y) \in (d)_u o(d)_v$. The reverse inclusion is immediate by the triangle inequality. For the equality $(d)_v o(d)_u = (d)_u o(d)_v$ it is sufficient to prove one inclusion. We prove the inclusion $(d)_v o(d)_u \subseteq (d)_u o(d)_v$. Let $(x,y) \in (d)_v o(d)_u$. Then there exists z in E such that $(x,z) \in (d)_v$ and $(z,y) \in (d)_u$. This means $d(x,z) \leq v$ and $d(z,y) \leq u$ which implies $d(x,y) \leq v + u$ by the triangle inequality. By the symmetry of d we also have $d(y,x) \leq v + u$. Now by the convexity of d, there exists $t \in E$ such that $d(y,t) \leq v$ and $d(t,x) \leq u$. By symmetry we obtain $d(x,t) \leq u$ and $d(t,y) \leq v$. Hence $(x,t) \in (d)_u$ and $(t,y) \in (d)_v$ proving that $(x,y) \in (d)_u o(d)_v$. □

For some spaces, we have a partial converse to Lemma 3.3.

Lemma 3.4. *The space* $V_{Eq} = (E,d)$ *from the remark in* §2 *is convex if and only if the elements of* V *are pairwise permuting equivalence relations on* E.

Proof. (⇒) Let (E,d) be convex. Observe that $(d)_v = v$ for each v in V. Since by the remark in §2 an ultrametric space is symmetric, Lemma 3.3 implies that the elements of V are pairwise permuting equivalence relations on E.

(⇐) Suppose that the elements of V are pairwise permuting equivalence relations. Let $x, y \in E$ and $r,s \in V$ be such that $d(x,y) \leq r \vee s$. Since, as it is well known, $r \vee s = ros$, there exists z in E such that $(x,z) \in r$ and $(z,y) \in s$. This shows that $d(x,z) \leq r$ and $d(z,y) \leq s$. □

Definition 7. We say that a *V*-metric space (E,d) has the *3-set extension property* if for every *V*-metric space (F,d'), every subset X of F with $|X| \leq \min(3, |E|)$, and every $x \in F \backslash X$ any nonexpansive function $f: (X,d') \to (E,d)$ has a nonexpansive extension to $X \cup \{x\}$.

Theorem 3.5. *Let the order of V be a join semilattice. If* (E,d) *is a symmetric V-metric space* with 3-*set extension property, then for all* $u,v,w \in V$ *we have*

(i) $(d)_v o(d)_u = (d)_u o(d)_v$.
(ii) $(d)_u o ((d)_v \cap (d)_w) = ((d)_u o(d)_v) \cap ((d)_u o(d)_w)$.

Proof. Since the 3-set extension property implies convexity [1] Thm II.2.1, clearly (i) follows by Lemma 3.4. For (ii), it is clear that for every $(u,v,w) \in V^3$,

$$(d)_u o ((d)_v \cap (d)_w) \subseteq ((d)_u o(d)_v) \cap ((d)_u o(d)_w)$$

In order to prove the reverse inclusion, let $(x,y) \in ((d)_u o(d)_v) \cap ((d)_u o(d)_w)$. There exist $p,q \in E$ such that $(x,p) \in (d)_u$, $(p,y) \in (d)_v$ and $(x,q) \in (d)_u$, $(q,y) \in (d)_w$. Set $X = \{(x,x), (y,q), (p,y)\} \subseteq E^2$. Define $f: X \to E$ by setting $f((x,x)) = x$ and $f((y,q)) = f((p,y)) = y$. It is immediate that f is a nonexpansive map from (X,d_2) into (E,d). By the 3-set extension property, there exists a nonexpansive extension $g: X \cup \{(p,q)\} \to E$ of f. Set $h = g((p,q))$. A simple verification shows that $(x,h) \in (d)_u$ and $(h,y) \in (d)_v \cap (d)_w$ and thus $(x,y) \in (d)_u o((d)_v \cap (d)_w)$. □

This result uses essentially the same method of proof as Kaarli [2] with the transitivity replaced by the triangle inequality.

Let $L = \langle L; \subseteq \rangle$ be a sublattice of $\langle Eqv(E); \subseteq \rangle$, the lattice of all equivalence relations on E. The set L is *inf-complete* if the intersection of every nonvoid family of elements of L belongs to L. Recall that L is *arithmetical* if its elements pairwise permute with respect to the product of relations and if for all θ, $\Phi, \Psi \in L$ the distributivity law $\theta \cap (\Phi o \Psi) = (\theta \cap \Phi) o (\theta \cap \psi)$ holds.

The next theorem gives a link between extension theorem and arithmetical algebras.

Theorem 3.6. *Let* E *be a set and* $L = \langle L; \subseteq \rangle$ *be an inf-complete sublattice of the lattice* $\langle Eqv(E); \subseteq \rangle$ *of equivalence relations on E such that* $\cup L = E^2$. *Then* L *is arithmetical if and only if for each finite subset* X *of* E, *there exists an a-function on* X, *compatible with* L.

Proof. ($\Rightarrow$) Let *L* be arithmetical and $V = \langle L; o, \Delta, id_L, \subseteq \rangle$. By the remark in §2 the function $d: E^2 \to L$ such that $d(x,y) = \cap\{\theta \mid (x,y) \in \theta\}$ is a *V*-ultrametric on E. As it was mentioned in the proof of Lemma 3.4, we have $(d)_v = v$ for each $v \in L$. Since the elements of L permute, (E,d) is convex by Lemma 3.4. Also (E,d) has the 2-Helly property for finite families of balls by Proposition 2.4 since the elements of L are idempotent. Since it is an ultrametric space by Theorem 2.3 (ii) $\Rightarrow$ (iv), it has the a-property; i.e., for each finite subset X of E, there exists a nonexpansive a-function on X compatible with L.

($\Leftarrow$) It follows directly from Proposition 2.4 since for each $v \in L$ we have that $(d)_v = v$ as already observed in the proof of Lemma 3.4. □

11.4 Strictly Locally Affine Complete Algebras

Let $A = \langle E\ C \rangle$ be an algebra. Denote by *Con(A)* the set of the congruences of the algebra *A*. We recall that a congruence of A is an element of Eqv(E) compatible with every fundamental operation of *A*. This implies that each congruence of *A* is a subalgebra of A^2 [3]. The set Con(*A*), ordered by inclusion of relations, is a complete sublattice of Eqv(E) containing Δ and E^2 and closed with respect to the join operation $\vee$ [4]. Set $V = \langle Con(A); \vee, \Delta, id, \subseteq \rangle$ where id is the identity function in Con(A). Then by the remark in §2 we obtain a *V*-ultrametric on E by setting for every pair (x,y) of E

$$d_A(x,y) = \cap \left\{ \theta \in Con(A) \;\middle|\; (x,y) \in \theta \right\}.$$

Definition 8. An algebra $A = \langle E; C \rangle$ is called *affine complete* if every operation on E compatible with Con(*A*) is a polynomial of *A*. The algebra *A* is called strictly *locally affine complete* if for each positive integer n and each finite subset X of E, every function $f: X^n \to E$ compatible with Con(*A*) is a restriction of a polynomial of *A*.

Theorem 4.1. *Let $A = \langle E; C\rangle$ be a strictly locally affine complete algebra and* X *a finite subset of* E. *Then there exists a polynomial* p_X *of A whose restriction to* X *is an a-function.*

Proof. Consider the ultrametric d_A on E. Denote by $d^{'}$ the ultrametric induced by V in E^3. Let X be a finite subset of E. By Lemma 2.2 there exists a nonexpansive map f from $(D_3 \cap X^3, d^{'})$ into (E,d_A) by setting $f(x,x,y) = f(y,x,y) = f(y,y,x) = y$ for all $x, y \in X$. By Theorem 3.1 the map f is compatible with Con(A). Since the algebra *A* is locally affine complete, there exists a polynomial p_X of A whose restriction to $D_3 \cap X^3$ is equal to f. The restriction of p_X to X is clearly an a-function on X. □

Corollary 4.2. (Hageman & Hermann [5]) *A strictly locally affine complete algebra is arithmetical.*

Proof. It follows easily from the above theorem and Theorem 3.6. □

Corollary 4.3. *If A is a strictly locally affine complete algebra, then any reflexive subuniverse of A^2 is a congruence of A.*

Proof. By Theorem 4.1 we know that for each finite subset X of E there is an a-function on X. Let θ be a reflexive subuniverse of A^2. We first prove that θ is a symmetric relation. Let (x,y) be an element of θ and set $X = \{x,y\}$. Then there exists a polynomial p_X of A whose restriction to X is an a-function. Since p_X is a polynomial, it preserves θ. Clearly (x,x), (x,y), and (y,y) are elements of θ and hence $(p_X(x,x,y), p_X(x,y,y)) = (y,x) \in \theta$. For the transitivity of θ let (x,y) and (y,z) be elements of θ. As (x,y), (y,y), and (y,z) are in θ, clearly $(p_X(x,y,y), p_X(y,y,z)) = (x,z)$ is in θ. This completes the proof. □

To prove the main result in this section, we need the following lemma that generalizes a result of Baker–Pixley [5].

Definition 9. Let C be a clone on E and k a positive integer. An n-ary operation f on E is *k-interpolable* by C if for any *k*-element subset U of E^n there exists $g \in C$ agreeing with f on U.

Lemma 4.4. *Let* C *be a clone on* E *such that for each finite subset X of* E *the clone* C *contains a ternary function m_X whose restriction to* X *is an* m*-function on* X. *If an* n*-ary operation* f *on* E *is* 2*-interpolable by* C, *then* f *is* k*-interpolable by* C *for any* $k > 2$.

Proof. We proceed by induction on $k > 1$. For $k = 2$ this is the hypothesis. Suppose that f is k-interpolable by C for some $2 \leq k < |E|$. Let $X = \{a_1, \ldots, a_{k+1}\}$ be a subset of E^n. For $i = 1,2,3$ let f_i be the interpolation of f on $X\backslash\{a_i\}$. Define $g: E^n \to E$ by setting

$$g(x_1, \ldots, x_n) = m_X(f_1(x_1, \ldots, x_n), f_2(x_1, \ldots, x_n), f_3(x_1, \ldots, x_n)).$$

Since m_X, f_1, f_2, f_3, are all elements of C, it follows that g is in C. Now we prove that g agrees with f on X. Let $a_m \in X$ where $1 \leq m \leq k+1$. Since X has at least three elements, there exists $1 \leq i < j \leq k+1$ such that $a_i \neq a_m \neq a_j$. This means that a_m belongs to the intersection of $X\backslash\{a_i\}$ and $X\backslash\{a_j\}$. Hence $f_i(a_m) = f_j(a_m) = f(a_m)$ and $g(a_m) = m_X(f_1(a_m), f_2(a_m), f_3(a_m)) = f(a_m)$. □

The following lemma is well known and easy to prove.

Lemma 4.5. *Let* $A = \langle E, F\rangle$ be *an algebra and* $X = \{(a_1,b_1), \ldots (a_n, b_n)\}$ *be a subset of* E^2. *Denote by* $R(X)$ *the reflexive subuniverse of* A^2 *generated by* X. *Then the pair* $(a,b) \in R(X)$ *if and only if there exists an* n*-ary polynomial f of* A *such that* $f(a_1, \ldots a_n) = a$ *and* $f(b_1, \ldots b_n) = b$.

Definition 10. Let $A = \langle E, F\rangle$ be an algebra and h an integer with $h > 2$. Let ρ be an h-ary relation on A and $1 \leq i < j \leq h$. We call $pr_{ij}\rho = \{(x_i,x_j) : (x_1, \ldots, x_h) \in \rho\}$ a *binary projection* of ρ. We say that *A* has the *property* B_2 if the subuniverses ρ and σ of A^h are equal whenever $pr_{ij}\rho = pr_{ij}\sigma$ for all $1 \leq i < j \leq h$.

We are now ready to prove the main result in this section.

Theorem 4.6. *For any algebra A on a base set* E *the following propositions are equivalent*:

(i) *A is strictly locally affine complete.*
(ii) *For each finite subset X of* E *there exists a ternary polynomial* p_X *of A whose restriction to* X *is an* a*-function on* X.
(iii) *Each reflexive subuniverse of* A^2 *is a congruence of the algebra A and for each finite subset* X *of* A *there exists a ternary polynomial* m_X *of A whose restriction to* X *is an* m*-function on* X.
(iv) Each *reflexive subuniverse of* A^2 *is a congruence of A and the algebra* $A^+ = \langle E, Pol(A)\rangle$ *has the property* B_2.

Proof. (i) ⇒ (ii). Theorem 4.1

(ii) ⇒ (iii). The fact that any reflexive subuniverse of A^2 is a congruence of *A* is already proved in Corollary 4.3. Now let X be any finite subset of E and p_X: $E^3 \to E$ a ternary polynomial of A whose restriction to X is an a-function on X. Define m: $E^3 \to E$ by setting $m(x,y,z) = p_X(x,p_X(x,y,z),z)$ for all x,y,z in E. Then m is a polynomial of *A* whose restriction to X is an m-function on X.

(iii) ⇒ (iv). Theorem of Rosenberg and Schweigert [7] Theorem 2.20.

(iv) ⇒ (i). By [7] Theorem 2 for every finite subset X of E there exists a ternary term function m_X whose restriction to X is an m-function on X. To prove that *A* is strictly locally affine complete let *f* be an n-ary operation on E compatible with *Con*(*A*). We show that f is 2-interpolable by Pol *A*. Let $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$ be arbitrary elements of E^n and let θ denote the reflexive subuniverse of *A* generated by $\{(a_1,b_1), \ldots, (a_n,b_n)\}$. By (iv) clearly θ is a congruence of *A* and so (f(a),f(b)) belongs to θ. Moreover, by Lemma 4.5 there exists an n-ary polynomial g of *A* such that $g(a) = f(a)$ and $g(b) = f(b)$. This proves that f is 2-interpolable by Pol *A*. From Lemma 4.4 we obtain that *f* is *k*-interpolable by Pol *A* for every

k > 2. Finally, let X be an arbitrary finite subset of E of cardinality *m.* As f is m^n-interpolable by Pol *A,* there exists an *n*-ary polynomial of *A* agreeing with f on X. This proves that *A* is strictly locally affine complete. □

We immediately obtain a fundamental result of Kaarli [2] Corollary 4.8.

Corollary 4.7. (Kaarli [2]) *Any arithmetical affine complete algebra that is denumerable or whose congruence lattice is finite is strictly locally affine complete.*

The following result yields examples of algebras that are not strictly locally affine complete.

Corollary 4.8. *Any algebra A whose square contains a reflexive subuniverse that is not a congruence of* A *is not strictly locally affine complete. In particular, any algebra compatible with a nontrivial order is not strictly locally affine complete.*

Proof. This follows easily by (iii). □

11.5 Affine Completeness of Modules

The question of affine completeness of modules over commutative rings has been studied by many authors. The case of abelian groups is well known thanks to W. Nöbauer, K. Kaarli, and A. Saks (see [8]). As pointed out in [4], most of the results on the affine completeness of abelian groups can be generalized to modules over commutative principal ideal domains because the abelian groups and these modules are similar due to the fact that the underlying ring structure of the ring of integers is that of a commutative principal ideal domain. According to K. Kaarli and A. Pixley, there is only one exception. Indeed, when proving that an abelian group of rank one with bounded torsion part is not affine complete, one relies on the countability of the ring of integers [4, Theorem 5.2.22]. This argument does not hold if it has to do with a ring which is uncountable. This leads to the following problem raised in [4, Problem 5.2.29].

Problem. Does there exist an affine complete torsion free module of rank 1 over a commutative principal ideal domain?

In this part we will answer this question and moreover, we will generalize some other theorems from abelian group's affine completeness theory to modules over a commutative domain. Throughout this part R will designate a commutative ring with 1, and A a left module over R. For any a in A, the annihilator of a is designated by $\{r \in R : ra = 0\}$and denoted by Ann(a). It is clear that Ann(a) is an ideal of R as well as $\mathrm{Ann}(A) = \cap a \in A\ \mathrm{Ann}(a)$. An element a of A for which Ann(a) is nontrivial will be said to be a torsion element of A. Clearly the set T of torsion elements of A is a submodule of A. The module A is bounded if Ann(A) is nontrivial and in this case any nonzero element of Ann(A) is called an exponent of A. Otherwise we say that it is unbounded.

For a subset X of the R-module M we denote by <X> the submodule of M generated by X. As noted in [5], the compatibility criterion takes the following form in the case of modules. An n-ary operation f on the R-module M is compatible with Con(M) if and only if $f(a) - f(b) \in < a_1 - b_1, \ldots, a_n - b_n >$ for all $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n) \in A^n$. Clearly this means that for $a = (a_1, \ldots, a_n)$, $b = (b_1, \ldots, b_n) \in A^n$ there exist $s_1, \ldots, s_n \in R$ depending on a and b, such that $f(a) - f(b) = s_1(a_1 - b_1) + \cdots + s_n(a_n - b_n)$. An n-ary function $f: M^n \to M$ is zero preserving if $f(0, \ldots, 0) = 0$. Obviously every Con(M)-compatible operation of the R-module M can be represented as a sum of a zero preserving function and a constant function. Also, if f is an n-ary zero preserving Con(M) compatible function on an R-module M and $(a_1, \ldots, a_n) \in M^n$, then there exist $r_1, \ldots, r_n$ in R such that $f(a_1, \ldots, a_n) = r_1 a_1 + \cdots + r_n a_n$ where the coefficients r_i depend on the n-tuple $(a_1, \ldots, a_n)$.

Besides the notion of a strictly locally affine complete algebra is the notion of a locally affine complete algebra. For every $n \in N$, an n-ary function f on an algebra A is said to be a local polynomial of A if it can be interpolated by a polynomial operation on every finite subset X of A^n, that is, there exists a polynomial operation $g: A^n \to A$ whose restriction to X is equal to f. An algebra A is said to be locally affine complete if every Con(A)-compatible function on A is a local polynomial of A.

Remark 3. It is clear from the above definition that an affine complete algebra is locally affine complete. It is easy to see, as noted in [1], that for an R-module A an n-ary polynomial operation on A is an operation $f: A^n \to A$ satisfying the following property: there exists $r_0 \in A$, $r_1, \ldots, r_n \in R$ such that the following equality holds: $f(x_1, \ldots, x_n) = r_1 x_1 + \cdots + r_n x_n + r0$ for all $(x_1, \ldots, x_n) \in A^n$.

Remark 4. It is important to notice that $r_0, \ldots, r_n$ do not depend on the n-tuple $(x_1, \ldots, x_n)$.

Since constant operations are polynomials, it follows that when studying affine completeness of modules, we may restrict to the case of zero preserving functions.

Remark 5. Polynomial operations on modules are just the linear functions. We recall here the well known notion of semisimple modules and a result on affine complete modules that will be used below. An R-module A is called semisimple if it satisfies the following equivalent conditions:

1. A is a direct sum of simple R-modules;
2. Every submodule of A is a direct summand.

The following theorem will be used. We skip its proof.

Theorem 3.9. [1] *A semisimple R-module A is locally affine complete if and only if it has no simple homogeneous component.*

11.5.1 Affine Completeness of Free Modules of Rank 1

We will now give a proof of [4, Theorem 5.2.22] which avoids the use of the compatible function extension property. We first recall that if A is a bounded abelian group, then Ann(A) is generated by a unique positive element e called the exponent of A.

Theorem 5.1. *An abelian group of rank 1 with a bounded torsion part is not locally affine complete.*

Proof. Let A be an abelian group of rank 1. We suppose that it is locally affine complete. We will construct a unary function on A, compatible with Con(A), and which cannot be interpolated by a polynomial operation in a specified finite subset of A. We will hence obtain a contradiction. Let T be the torsion part of A so that A/T is a torsion free group of rank 1; it then has a basis over the ring of integers $\mathfrak{Z}$ with one element. Let $\{a+T\}$ be a basis of A/T over $\mathfrak{Z}$. Define the function g: $A/T \to A/T$ by setting $g(k(a+T)) = k^2a + T$, for each $k \in \mathfrak{Z}$ It is easy to see that g is a zero preserving function on A/T which is Con(A/T)-compatible. Now, since T is bounded, let $\exp(T) = e$ be its exponent. It follows that $e(x+t) = ex$ for all $(x,t) \in A \times T$. Hence $e(x+T)$ is the set $\{ex\}$ for every $x \in A$. This fact allows us to define a function f: $A \to A$ by sending x to the unique element of $e(g(x+T))$. For simplicity, we choose $f(x) = e(g(x+T))$. This function is a zero preserving function and it induces a function eg on the quotient group A/T. If $ka+T$ is an element of A/T, then $eg(ka+T) = ek^2a + T$. So if f were a local polynomial on A, then eg would be a local polynomial on A/T. Let $X = \{x_1 + T, \ldots, x_n + T\}$ be a finite subset of A/T with $n > 0$. Since f is a local polynomial then there exists an integer s in R such that the restriction of the polynomial operation $y \to sy$ to $\tilde{X} = \{x_1, \ldots, x_n\}$is equal to f on $\tilde{X}$. For $i = 1, \ldots, n$, we therefore have $eg(x_i + T) = ek_i^2(a+T) = sk_i^2(a+T) = sk_ia + T$, where $x_i + T = k_ia + T$. Taking $\tilde{X} = \{ka | k = 0,1,2\}$ we obtain $sk(a+T) = ek^2(a+T)$, $k \in \{0,1,2\}$. Since $a+T$ is an element of infinite order, the latter equality would imply that $s = e$ and $2s = 4e$, which is clearly absurd since e is a nonzero integer. It now remains to prove that f is a Con(A)-compatible function on A. Let b, $c \in A$ and $g(b+T) = b_1 + T$, $g(c+T) = c_1 + T$ with $b_1, c_1 \in A$. Then $f(b) = eb_1$ and $f(c) = ec_1$. Since g is Con(A)-compatible and zero preserving, there exists an integer r such that $g(b+T) - g(c+T) = r(b-c+T)$. Consequently there also exists $t \in T$ such that $b_1 - c_1 = r(b-c) + t$. Then $f(b) - f(c) = e(r(b-c)+t) = er(b-c) \in \langle b-c \rangle$ proving that f is Con(A)-compatible. □

We will use the idea of the above proof in other situations below.

Theorem 5.2. *A torsion free module of rank 1 over a commutative domain is not locally affine complete.*

Proof. Again we proceed by contradiction. Let A be a torsion free module of rank 1 over a commutative domain R. Since by [4, Theorem 5.2.9] any 1-dimensional vector space is not affine complete, we can suppose that R is not the field with two

elements. Let (x) be a basis of A. Then each element $a \in A$ has the form $a = rx$ for some $r \in R$. Let us define the unary function f: $A \to A$, $rx \to r^2x$. Then f is Con(A)-compatible on A. Indeed if $a_1 = r_1x$ and $a_2 = r_2x$ are elements of A, then

$$f(a_1) - f(a_2) = {r_1}^2x - {r_2}^2x = (r_1 + r_2)(a_1 - a_2) \in < a_1 - a_2 > .$$

Let us assume that A is locally affine complete, which implies that f is a local polynomial on A. Since R is not the field with two elements we choose $\alpha \in R\backslash\{0,1\}$ and consider the finite set $X = \{0, x, \alpha x\}$. Since f is interpolated by a polynomial operation on X, we can find $s, t \in R$ such that $f(a) = sa + t$ for all $a \in X$. Therefore

$$\begin{cases} f(0) = t = 0 \\ f(x) = sx + t = x \\ f(\alpha x) = s\alpha^2 x + t = \alpha^2 x \end{cases}$$

Putting $t = 0$ and using the fact that (x) is a basis, we have $s = 1$ and therefore $\alpha = \alpha^2$. Now since R is a domain it follows that $\alpha = 0$ or $\alpha = 1$, which yields a contradiction. □

Corollary 5.3. *A torsion free module of rank 1 over a commutative domain is not affine complete.*

Proof. A direct application of Theorem 5.2. □

Remark 6. Actually, we do not need the compatible function extension property to prove that torsion free abelian groups of rank 1 are not locally affine complete. This is one of the main results we obtain in this work. This property is relevant because the compatible function extension property was originally introduced in order to prove that a free abelian group of rank 1 is not affine complete. Now [4, Problem 5.2.29] is solved by the above theorem. We will now prove the following theorem that gives a more general result about local affine completeness of modules with one generator.

Theorem 5.4. *A nontrivial cyclic module over a commutative principal ideal domain R is not locally affine complete.*

Proof. Let A be a nontrivial cyclic module over a commutative principal ideal domain R. Then the R-module A is isomorphic to R/(r) for some $r \in R$, and is generated by $x = 1 + (r)$. Let us define f: $A \to A$ by setting $f(a) = s^2x$ where $a = s + (r) = s(1 + (r)) = sx$ and $s \in R$. We will show by contradiction that f is Con(A)-compatible but is not a local polynomial of A. The compatibility of f with Con(A) is straightforward. To show that f is not a local polynomial, let us first assume that there exists an element $\alpha \in R$ such that $\alpha(\alpha - 1) \notin R$. Set $X = \{0, x, \alpha x\}$ and suppose that f is a local polynomial on A. Then there exist $t_1, t_2 \in R$ such that $f(a) = t_1a + t_2$ for all $a \in X$. It is clear that we must have $t_2 = 0$, so that $f(x) = t_1x = x$ and $f(\alpha x) = \alpha^2 x = t_1\alpha x$. Hence $\alpha x = \alpha^2 x$ and $\alpha - \alpha^2 \in (r)$ which is a contradiction. To complete this proof, we now suppose that for every $\alpha \in R$ we have $\alpha(\alpha - 1) \in (r)$.

Let $r = p_1^{\beta_1} \dots p_n^{\beta_n}$ be the factorization of r into irreducible elements. Suppose that $\beta_i > 1$ for some $i \in \{1, \dots, n\}$. Then, since $p_j(p_j - 1) \in (r)$ for all $j \in \{1, \dots, n\}$, there exists $t_i \in R$ such that $p_i(p_i - 1) = rt_i$. But $r = p_i u$ where $u = p_1^{\beta_1} \dots p_i^{\beta_i - 1} \dots p_n^{\beta_n}$. This yields $p_i(p_i - 1) = p_i u t_i$. Consequently $p_i - 1 = ut_i$ and thus p_i is not a factor of u, which contradicts the assumption that $\beta_i > 1$. Therefore, the factorisation of r is of the form $r = p_1 \cdots p_n$. This factorisation implies that $R/(r) = R/(p_1) \oplus \cdots \oplus R/(p_n)$ which proves that A is a semisimple module with simple homogeneous components. Hence A cannot be locally affine complete [4, Theorem 5.2.13]). □

The next question is what concerns modules of rank 1 whose torsion part is bounded. The answer is given by the following theorem which generalizes [4, Theorem 5.2.22] to modules over a commutative domain. We will use the same idea as in the proof of Theorem 5.1.

Theorem 5.5. *Let A be a module of rank 1 over a commutative domain with a nonzero torsion part T such that Ann(T) is nontrivial. Then A is not locally affine complete.*

Proof. We will construct a Con(A)-compatible unary function on A which is not a local polynomial. First, R is not a field with two elements since a vector space has no nontrivial torsion part. Denoting by T the torsion part of the R-module A, then A/T is a torsion free R-module of rank 1. Let $(a + T)$ be a basis of A/T over R, and define f: $A/T \to A/T$ by setting $f(ka + T) = k^2a + T$, $k \in R$. Then f is clearly Con(A/T)-compatible. Let $r \in Ann(T)$ be nonzero. Then for all x in A and t in T, we have $r(x + t) = rt$, so that $r(x + T)$ is a well defined element of A. We may hence define a function g: $A \to A$ by the formula $g(x) = r(f(x + T))$. This function induces a function rf on the quotient A/T. Indeed if g were a local polynomial on A, then rf would be a local polynomial on A/T. Since f is Con(A/T)-compatible rf is Con(A/T)-compatible and g is Con(A)-compatible. Suppose that A is locally affine complete. Then choose $\alpha \in R$ such that $\alpha \notin \{0,1\}$ and set $X = \{T, a + T, \alpha a + T\}$. By the local affine completeness of A, rf can be interpolated by a polynomial operation on X, so there exists $s \in R$ such that $rk^2(a + T) = sk(a + T)$, for $k = 0, 1, \alpha$. Since $a + T$ is not a torsion element, this implies that $r = s$ and $r\alpha^2 = s\alpha$, thus $r = 0$. This is impossible since we have assumed that $r \neq 0$. □

11.5.2 Affine Completeness of Modules of Rank Greater than One

In this section we generalize some affine completeness results for modules of rank 1 over a commutative domain to modules over a commutative domain. We first give preliminary lemmas that we will need for our main result given by Theorem 5.9.

Lemma 5.6. *Let A be a module over a ring R. Assume that for any d in A the annihilator of the quotient A/Rd is the same as the annihilator of A. Then A is affine complete if and only if any unary Con(A)-compatible function on A is a polynomial.*

Proof. We only have to prove that binary Con(A)-compatible operations on A are polynomials [4, Theorem 5.2.3]. Let f be a binary Con(A)-compatible operation. Without loss of generality, we may assume that $f(0,0) = 0$. Since unary Con(A)-compatible operations are polynomials, for each $x, y \in A$ there exist some elements of R, k_y, l_x, b_y, c_x such that

$$\{f(x,y) = k_y x + b_y \tag{11.1}$$

$$\{f(x,y) = l_x y + c_x \tag{11.2}$$

Since $f(0,0) = 0$, we must have $b_0 = c_0 = 0$. This shows that $f(x,0) = c_x = k_0 x$ and $f(0,y) = b_y = l_0 y$ for all x, y in A. From (11.1) and (11.2) we obtain the equality:

$$(k_y - k_0)\, x = (l_x - l_0)\, y. \tag{11.3}$$

We want to prove that the both sides of Eq. (11.3) are 0. Suppose that this condition is not satisfied, then, for example, the left side is not identically zero. Therefore there exists d in A such that the left side of (11.3) is not 0. Thus $k_d x \neq k_0 x$ for some x. In A/Rd the right side of (11.3) vanishes and the left side must also vanish. This shows that $k_d - k_0$ is in the annihilator of A/Rd. By hypothesis this annihilator is the same as the annihilator of A which contradicts the fact that the left side of (11.3) is nontrivial for $x \neq d$. □

Lemma 5.7. *Let A be a module over a commutative domain R. If A contains a free submodule of rank ≥ 2, then A is affine complete if and only if each unary Con(A)-compatible function on A is a polynomial.*

Proof. We only need to prove that for each d in A the annihilator of A is the same as the annihilator of A/Rd, that is, they are all trivial. Let a be an element of the annihilator of A/Rd. Choose a free pair {x,y} in A. We have $ax = td$ and $ay = sd$ for some t and s in R. Clearly $sax - tay = 0$. Hence, due to the freeness of {x,y} we have $ta = sa = 0$. Observe that, since R is a domain, then $a \neq 0$ implies that $t = s = 0$. We conclude that $ax = 0$ which is absurd. Hence $a = 0$ and this leads to the result that the annihilator of A/Rd is trivial. □

The next corollary is direct consequence of Lemma 5.7.

Corollary 5.8. *Let A be a module over a commutative domain R. If A contains a free submodule of rank ≥2, then every quotient of A by a cyclic submodule is unbounded.*

We are now able to prove a general result about modules containing submodules of rank greater than 2. It shows also that modules of rank greater than 2 are affine complete.

Theorem 5.9. *Let A be a module over a commutative domain. If A contains a free direct summand of rank ≥2, then A is affine complete.*

Proof. Suppose that A has a free direct summand of rank at least 2. Then A contains a submodule of rank at least 2. Hence A satisfies the hypothesis of Lemma 5.7; so it is sufficient to prove that every unary function on A, compatible with Con(A), is a polynomial operation. From the hypothesis we know that $A = A_1 \oplus F$ where F is a free module with rank ≥ 2. By the compatibility of f with Con(A), there are functions $g: A_1 \to A_1$ and $h: F \to F$ compatible with $Con(A_1)$ in A_1 and with Con(F) in F, respectively, such that

$$f(x + y) = g(x) + h(y) \tag{11.4}$$

for every $x \in A_1$ and $y \in F$. We may also suppose that f is zero preserving, which is the case for g and h. But F is affine complete by [5, Theorem 5.2.8], hence there exists r in R such that $h(y) = ry$ for every y in F. Moreover, the compatibility of f and g, respectively, with Con(A) and $Con(A_1)$ implies that, for every $x \in A_1$ and $y \in F$, there exists s_{x+y}, $t_x \in R$ such that $f(x + y) = s_{x+y}(x + y)$, $g(x) = t_x x$. Equation (11.4) thus implies that $s_{x+y}x = t_x x$, $s_{x+y}y = ry$ for all $x \in A_1$ and $y \in F$. Taking y as a nonzero element we see that $sx + y = r$ for all $x \in A_1$ and $y \in F$. We hence get that $f(x + y) = r(x + y)$ for all $x \in A_1$ and $y \in F$.

References

1. Jawhari, E., Misane, D., Pouzet, M.: Graphs and ordered sets from the metric point of view. Contemp. Math. **57**, 175–223 (1986)
2. Kaarli, K.: Compatible function extension property. Algebra Univers. **17**, 200–207 (1983)
3. Kaarli, K., Pixley, A.F.: Polynomial Completeness in Algebraic Systems. Chapman & Hall, London/New York/Washington (2000)
4. Kientega G.: Métriques généralisées et algèbres affinement completes. Ph D thesis, Université de Montréal (1992)
5. Baker, K.A., Pixley, A.F.: Polynomial interpolation and the Chinese remainder theorem for algebraic systems. Math. Z. **143**, 165–174 (1975)
6. Pouzet, M., Rosenberg, I.G.: General metrics and contracting operations. Discrete Mathematics **130**, 103–169 (1994)
7. Rosenberg I.G., Schweigert D.: Almost unanimity operations, contributions to general algebra 5. In: Proceedings of the Salzburg Conference, May 29–June 1, 1996. Verlag Hôlder-Pichler, Tempsky, Wien (1987)
8. Kaarli, K.: Affine complete Abelian groups. Math. Nachr. **107**, 235–239 (1982)
9. Kientega, G., Nonkane, I.: Affine completeness of some modules. Afr. Diaspora J. Math. **14**(1), 73–82 (2012)
10. Kientega, G., Rosenberg, I.: Extension of partial operation and relations. Math. Sci. Res. J. **8**(12), 362–372 (2004)

Chapter 12
A Codimension-Four Singularity with Potential for Action

Bernd Krauskopf and Hinke M. Osinga

Abstract We review how a conjectural codimension-four unfolding of the full family of cubic Liénard equations helped to identify the central singularity as an excellent candidate for the organizing center that unifies different types of spiking action potentials of excitable cells. This point of view and the subsequent numerical investigation of the respective bifurcation diagrams led, in turn, to new insight on how this codimension-four unfolding manifests itself as a sequence of bifurcation diagrams on the surface of a sphere.

Keywords Cubic Lienard equations • Spiking action potentials • Unfolding • Codimension-four singularity • Pseudo-plateau bursting • Bogdanov–Takens bifurcation • Homoclinic bursting

In 1952, Hodgkin and Huxley [9] formulated the first realistic mathematical model describing the flow of electric current through the surface membrane of a squid giant axon. Their system produces a sequence of single action potentials, which are equivalent to the relaxation oscillations generated by a simple RCL-circuit (involving a resistor, capacitor, and inductor) such as the Van der Pol oscillator [19, 20]. Electrically excitable cells can exhibit many other bursting patterns, which can loosely be interpreted as a series of spikes (action potentials) modulated by a slower relaxation oscillation. The bursting is related to and controlled by ionic currents through channels in the cell wall, which evolve on much slower time scales. Rinzel [15, 16] was the first to explain such bursting patterns mathematically in terms of an underlying bifurcation diagram with a hysteresis loop, which is traversed by one or more slowly varying parameters; see also [10].

The bursting pattern one finds depends on the codimension-one bifurcations that are encountered, that is, on the relative positions of saddle-node bifurcations, Hopf bifurcations, and homoclinic bifurcations that are crossed by the slowly varying parameter. These occur naturally near codimension-two Bogdanov–Takens bifurcations in two-parameter bifurcation diagrams of planar systems which,

B. Krauskopf (✉) • H.M. Osinga
Department of Mathematics, The University of Auckland, Auckland, New Zealand
e-mail: b.krauskopf@auckland.ac.nz

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_12

therefore, arise as "minimal models" of bursting patterns of action potentials. The classification of bursting patterns was formalized further by studying the transitions between them via parameter dependence of the underlying bifurcation diagram. In particular, the organization of the two-parameter bifurcation diagram under consideration changes when the Bogdanov–Takens bifurcation itself undergoes a bifurcation, which is an event of codimension-three where a higher-order normal-form term vanishes. This realization is behind the work of Bertram et al. [2], who presented many known bursting patterns as generated by horizontal parameter paths through a two-parameter bifurcation diagram of the Chay–Cook model, which is a paradigm model that retains many physiological features and is representative for a large class of realistic models of neuronal spiking. They realized that this bifurcation diagram of the Chay–Cook model can be found as a slice in the three-parameter unfolding of the degenerate Bogdanov–Takens singularity of focus type (or nilpotent cusp of order three)—one of the classic codimension-three bifurcations, with a two-dimensional center manifold, whose unfolding in planar vector fields was presented in [6]; see already case (M) of Fig. 12.1. This point of view was made explicit in the paper by Golubitsky et al. [7], who proposed a classification of bursting patterns in terms of the smallest codimension of a singularity in whose unfolding it can be generated (via a path of one or more slow parameters). In particular, they showed that the so-called fold/homoclinic or square-wave bursting, which involves a hysteresis loop generated by a saddle-node and homoclinic bifurcation, requires an underlying codimension-three singularity, such as the degenerate Bogdanov–Takens singularity of focus type considered in [2].

It emerged that one type of bursting, called pseudo-plateau bursting—first analyzed in [17] and also known as fold/subHopf bursting—could not be found in the unfolding of this codimension-three singularity. This was puzzling because, for biological reasons, it was considered to be related to fold/homoclinic bursting, which is part of the patterns found in [2]. Recent work by Osinga et al. [14] showed that all the relevant types of bursting, including fold/subHopf and fold/homoclinic bursting, can be found near a doubly degenerate Bogdanov–Takens singularity, whose conjectural unfolding was presented in 1998 by Khibnik et al. [11]. As a result, this codimension-four singularity and its unfolding has enjoyed particular interest from mathematical biologists. Quite amazingly, it emerged as a natural organizing center that unifies an entire class of different bursting patterns of electrically excitable cells.

We now proceed in Sect. 12.1 by recalling the candidate unfolding of the doubly degenerate Bogdanov–Takens bifurcation from [11] and review in Sect. 12.2 the results from [14] on the identification of fold/subHopf bursting near this singularity. Section 12.3 then presents numerical results on the nature of the codimension-four unfolding in terms of bifurcation diagrams on spheres. In particular, we show that all topologically different bifurcation diagrams can be found readily on spheres of appropriate radii; this point of view is particularly helpful for identifying two-parameter sections that feature certain bursting patterns of interest. We summarize and draw some conclusions in Sect. 12.4.

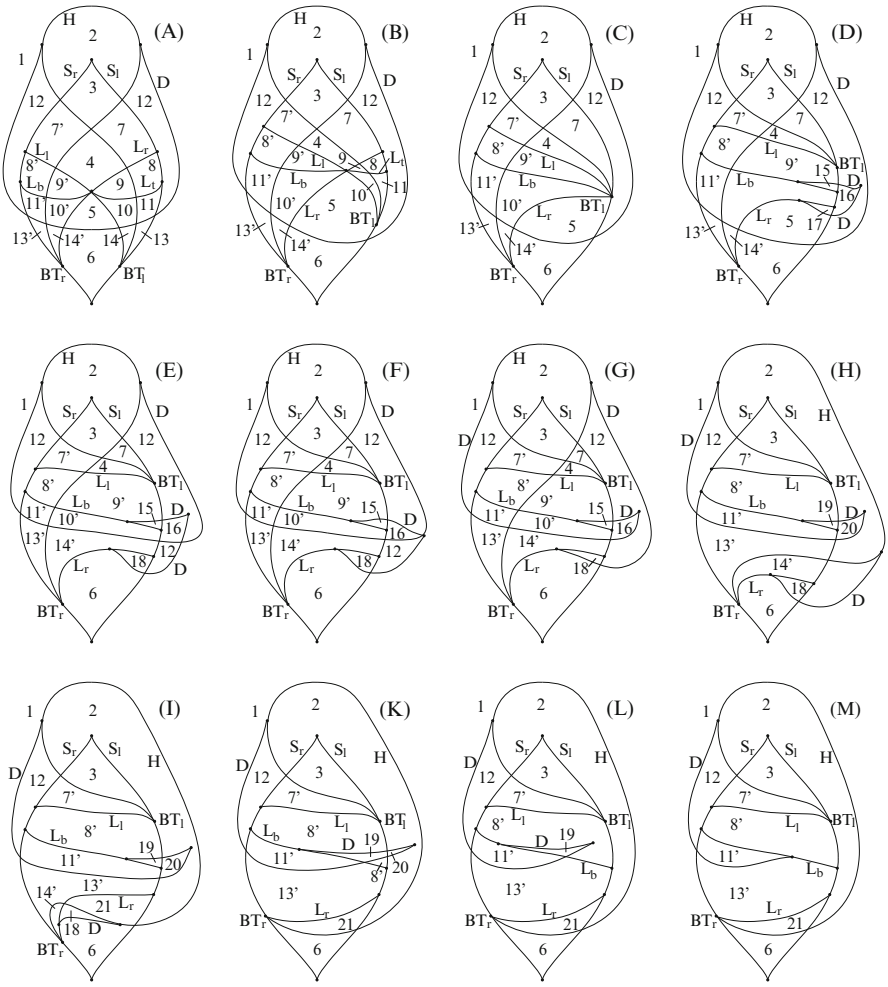


Fig. 12.1 Sketch of the suggested transition with increasing μ_4 between codimension-three unfoldings on a sphere in (μ_1, μ_2, μ_3)-space of (12.1); the associated phase portraits can be found in Fig. 12.2. Reproduced with permission from [11]. ©1998 IOP Publishing & London Mathematical Society. All rights reserved

12.1 Candidate Four-Parameter Unfolding

In the final section of the paper [11] the four-parameter planar vector field

$$\begin{cases} \dot{x} = y, \\ \dot{y} = \mu_1 + \mu_2 x + \mu_3 y + \mu_4 xy - x^3 - x^2 y, \end{cases} \tag{12.1}$$

was considered. It represents a candidate unfolding that provides a connection between two codimension-three bifurcations: the case $\mu_4 = 0$, which was the main subject of study in [11], and the case of sufficiently large μ_4, when (12.1) represents a nilpotent focus of codimension-three as studied in [6]. In fact, when the four parameters μ_i are allowed to vary over the reals, (12.1) represents the full family of cubic Liénard equations.

The point of view taken in [11] was to consider the transition of the three-parameter bifurcation diagram of (12.1) in (μ_1, μ_2, μ_3)-space as the parameter μ_4 is varied between these two known cases of $\mu_4 = 0$ and μ_4 sufficiently large. The respective three-parameter bifurcation diagram for a given value of μ_4 can be represented conveniently on the surface of a sphere in (μ_1, μ_2, μ_3)-space (due to cone structure of the unfolding); it changes qualitatively on the sphere at non-generic values of μ_4, which include different types of codimension-three singularities. Importantly, there are also quite a number of events of codimension "one-plus-two," where a bifurcation curve moves over a codimension-two bifurcation point on the sphere.

Figure 12.1 reproduces from [11] the respective series of sketched bifurcation diagrams (A) to (M) on the sphere (represented in stereographic projection), and Fig. 12.2 reproduces the associated phase portraits. The starting point is the reflectionally symmetric bifurcation diagram (A) for $\mu_4 = 0$; details and the proof of correctness can be found in [11]. There is then a first event of codimension "one-plus-two," when the curve D of double (or saddle-node) limit cycles crosses over the Bogdanov–Takens bifurcation point BT_l, yielding bifurcation diagram (B). At (C) there is a cuspidal loop formed by the separatrices of a Bogdanov–Takens point, which then gives bifurcation diagram (D). The curve D then moves up and at (F) there is a limit cycle of multiplicity four; it is unfolded by a swallow tail yielding (G). In a sequence of events of codimension "one-plus-two" the curve H of Hopf bifurcation then moves past the Bogdanov–Takens bifurcation BT_l and beyond to give bifurcation diagram (H), and then the degenerate Hopf point on H moves across the saddle-node bifurcation curve S_l to result in (I). Then there is a cusp of order three, yielding (K), after which the cusp point on D moves over S_l to yield bifurcation diagram (L). Finally, there is a homoclinic loop of order three and the final result is bifurcation diagram (M), which is that of the nilpotent focus; compare with [1, 6].

This sequence of unfoldings (A)–(M) in Fig. 12.1 takes into account the information available at the time, especially that on different codimension-three bifurcations. The existence of the cuspidal loop had been studied in [22] and, except for the limit cycle of multiplicity four, the stated codimension-three bifurcations had been noted explicitly in [3]; moreover, rigorous numerics in [8, 13] showed the existence of a small region with four limit cycles. The overall unfolding of (12.1) in Fig. 12.1 was constructed abstractly in [11] in the spirit of a "minimal model" and it is, hence, conjectural, specifically in terms of the exact sequence of codimension-three and codimension-one-plus-two bifurcations.

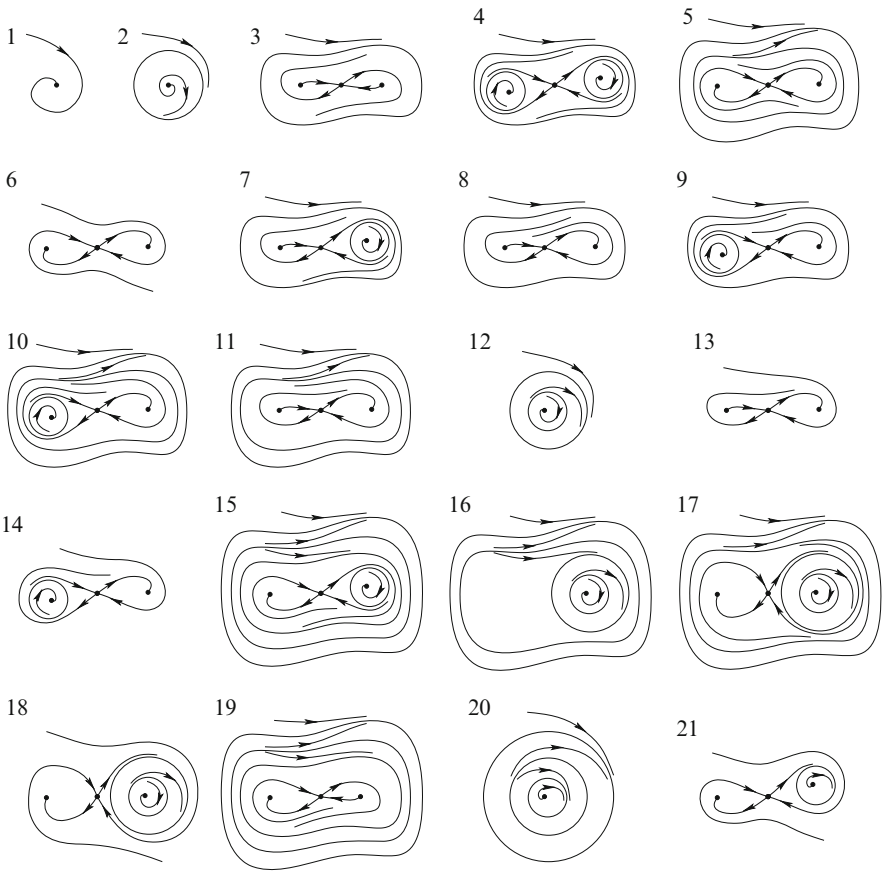


Fig. 12.2 Phase portraits of (12.1) in the open regions of the bifurcation diagrams in Fig. 12.1; adding a $'$ to the number corresponds to a rotation of the phase portrait by π. Reproduced with permission from [11]. ©1998 IOP Publishing & London Mathematical Society. All rights reserved

12.2 Identification of Fold/Sub-Hopf Bursting

Bertram et al. [2] considered a two-parameter slice near the degenerate Bogdanov–Takens singularity of focus type, where the two saddle-node curves are parallel vertical lines. This corresponds to the (μ_1, μ_3)-plane with $\mu_2 = const < 0$ and μ_4 sufficiently large in (12.1); see case (M) in Fig. 12.1. The different bursters were identified as different horizontal parameter paths in this parameter plane, along which μ_1 changes back and forth slowly.

In a similar spirit, Osinga et al. [14] were guided by the bifurcation diagrams in Fig. 12.1 and presented the fold/subHopf or pseudo-plateau burster by a suitable horizontal path on the relevant bifurcation diagram on the unit sphere in (μ_1, μ_2, μ_3)-space for $\mu_4 = 0.75$. Figure 12.3 reproduces from [14] the bifurcation

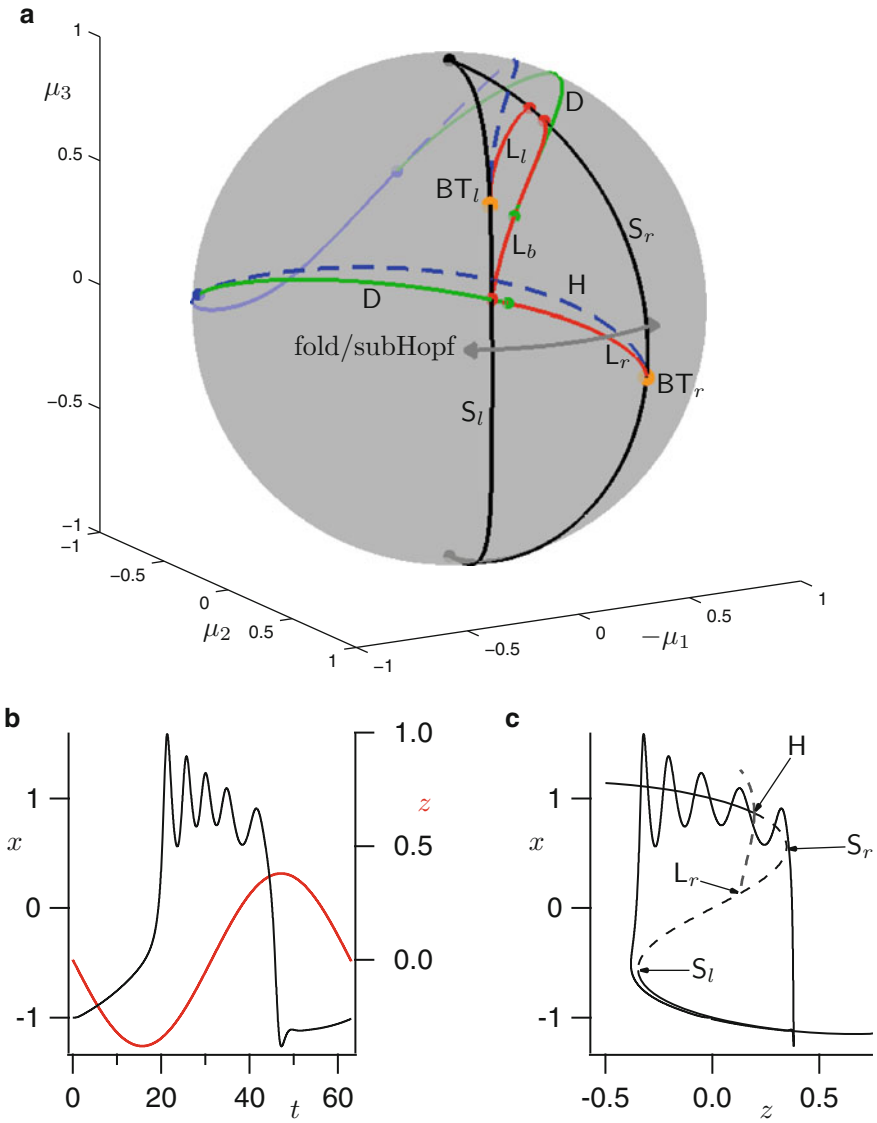


Fig. 12.3 Fold/subHopf bursting for system (12.1) as generated by a parameter path on the unit sphere in $(-\mu_1, \mu_2, \mu_3)$-space with $\mu_4 = 0.75$. Panel (**a**) shows the bifurcation diagram and the path on the unit sphere. Panels (**b**) and (**c**) show the time series and the underlying bifurcation diagram of the corresponding fold/subHopf bursting pattern. Reproduced with permission from [14].

diagram and the path on the unit sphere, as well as the time series and phase-space representation of the ensuing fold/subHopf bursting. More specifically, the path is

parameterized by $\mu_1 \in [-0.38, 0.38]$, with $\mu_2 = \sqrt{1-\mu_1^2-\mu_3^2}$, $\mu_3 = 0.1$ and $\mu_4 = 0.75$. System (12.1) exhibits along this path the saddle-node bifurcation of equilibria S_l, the homoclinic bifurcation L_r, the subcritical Hopf bifurcation H, and the other saddle-node bifurcation of equilibria S_r. For consistency of presentation, images from [14] are reproduced here with parameters and notation as used in [11]. In fact, in [14] $\mu_3 = \nu$, $\mu_4 = b$, and μ_1 has the opposite sign; moreover, the curves S_l, S_r, H, D, L_l, L_b, and L_r here are referred to in [14] as SN_l, SN_r, H_l or H_r, SNP, HC_l, HC_c, and HC_r, respectively. The relevant features of the bifurcation diagram on the sphere in Fig. 12.3a correspond qualitatively to a situation in between cases (G) and (H) in Fig. 12.1; a difference is that (G) and (H) feature a cusp bifurcation point on the curve D of double limit cycles in Fig. 12.3.

The bursting pattern is generated by introducing a slow variable defined by

$$z(t) = -\mu_1(t) := -0.38 \sin(\varepsilon t),$$

where the time-scale separation parameter $\varepsilon = 0.1 > 0$ is small (but not so small that delayed bifurcation phenomena are encountered). The x-coordinate of system (12.1) represents the membrane potential, and it exhibits the particular bursting pattern known as fold/subHopf or pseudo-plateau bursting [17]; its time series is shown in Fig. 12.3b together with the time series of the slow variable $z(t) = -\mu_1(t)$.

The biologically distinguishing aspects of fold/subHopf bursting are its relatively short period and the small amplitudes of the spikes on the plateau [17]; see also [12, 18, 21]. In contrast to fold/homoclinic or square-wave bursting, the spikes are not stable oscillations but rather transient oscillations that damp down to an upper steady state. Hence, if the time-scale separation parameter is too small, the time series will consist of relaxation oscillations instead. Fold/subHopf bursting only arises if the contraction to the upper steady states is weak relative to the speed of the slow variable.

Figure 12.3c shows the underlying periodic oscillation overlayed onto the bifurcation diagram in the (z, x)-plane. As can be checked, fold/subHopf bursting cannot be generated by any path on the two-parameter bifurcation diagram in [2].

Indeed, it has been argued in [14] that fold/subHopf or pseudo-plateau bursting can only be generated in the vicinity of a codimension-four singularity, such as that in system (12.1). However, the bursting patterns of fold/subHopf and fold/homoclinic bursting are considered very similar and it is often hard to distinguish the two types in experiments. Indeed fold/homoclinic or square-wave bursting was found in [2] near the degenerate Bogdanov–Takens singularity of focus type, that is, in system (12.1) for sufficiently large μ_4. Hence, it seems natural to expect the existence of a parameter path in the full four-dimensional parameter space of system (12.1) that generates fold/homoclinic bursting. Furthermore, it should be possible to deform and/or move this path such that the type of bursting changes to fold/subHopf bursting. In order to find such a transition, the four-dimensional $(\mu_1, \mu_2, \mu_3, \mu_4)$-space of system (12.1) was explored in [14] by setting $\mu_4 = 0.75$

and considering horizontal or vertical sections chosen appropriately relative to the bifurcation diagram on the sphere. The section for $\mu_2 = 0.0675$ (not shown; see [14]) gives an associated bifurcation diagram in the (μ_1, μ_3)-plane that is exactly the one near the degenerate Bogdanov–Takens singularity of focus type presented in [2].

Furthermore, the choice $\mu_3 = -0.09$ gives a bifurcation diagram in the $(-\mu_1, \mu_2)$-plane that features paths for both fold/subHopf and fold/homoclinic bursting, thus, providing the sought connection between the two. This is illustrated in Fig. 12.4 reproduced from [14] (with $-\mu_1$ along the horizontal axis, owing to the mentioned sign change). Panel (a) shows the section for $\mu_3 = -0.09$ relative to the unit sphere for $\mu_4 = 0.75$; panel (b) shows the corresponding bifurcation diagram in the $(-\mu_1, \mu_2)$-plane together with the paths for fold/subHopf and fold/homoclinic bursting; and panel (c) is an enlargement to highlight the transition to fold/homoclinic bursting. An important observation in Fig. 12.4b is the presence of two codimension-two Bogdanov–Takens points, denoted BT_r and $\mathsf{BT}_r^{\mathrm{far}}$, on the saddle-node bifurcation curve SN_r. The point $\mathsf{BT}_r^{\mathrm{far}}$ has the same local unfolding as BT_r in Fig. 12.3, but the Hopf bifurcation in the local unfolding of BT_r in Fig. 12.4b is supercritical. This implies that the bifurcation diagram on a sphere of sufficiently small radius $R \ll 1$ in Fig. 12.4a is, in fact, topologically that near the degenerate Bogdanov–Takens singularity of focus type, that is, case (M) of Fig. 12.1.

12.3 Transitions of Bifurcation Diagram on a Sphere

The analysis in [14] started with the hypothesis that there exists a bifurcation diagram on the unit sphere for a suitable choice of μ_4 in system (12.1) such that both fold/subHopf and fold/homoclinic bursting could be generated by paths on this sphere. As we argued above, this is not actually the case. Moreover, these initial investigations indicated that the transition from case (A) to case (M) does exist, but that the sequence of codimension-three bifurcations on a sphere in (μ_1, μ_2, μ_3)-space is not exactly as proposed in [11] and shown in Fig. 12.1. In particular, it seems that there is no cusp point on the curve D of double limit cycles that disappears in a codimension-three singularity on L_b in between case (L) and case (M) in Fig. 12.1.

As was mentioned at the end of Sect. 12.2, the bifurcation diagram on the sphere changes topologically when its radius is decreased. We now consider this aspect of the codimension-four unfolding in more detail. As was already known from [11], for sufficiently large μ_4 the bifurcation diagram on a sphere with a fixed radius is that of the nilpotent focus of codimension-three as presented in [6]. Here sufficiently large μ_4 means sufficiently large *relative to* μ_1, μ_2, and μ_3. Hence, for any given value of $\mu_4 > 0$ this is satisfied on any sphere with sufficiently small radius $R = \sqrt{\mu_1^2 + \mu_2^2 + \mu_3^2}$, which has the following interesting consequence. Suppose one considers a sphere of a given fixed radius, say, with $R = 1$, with the bifurcation diagram of case (A) in Fig. 12.1 on it. As soon as $\mu_4 > 0$, then

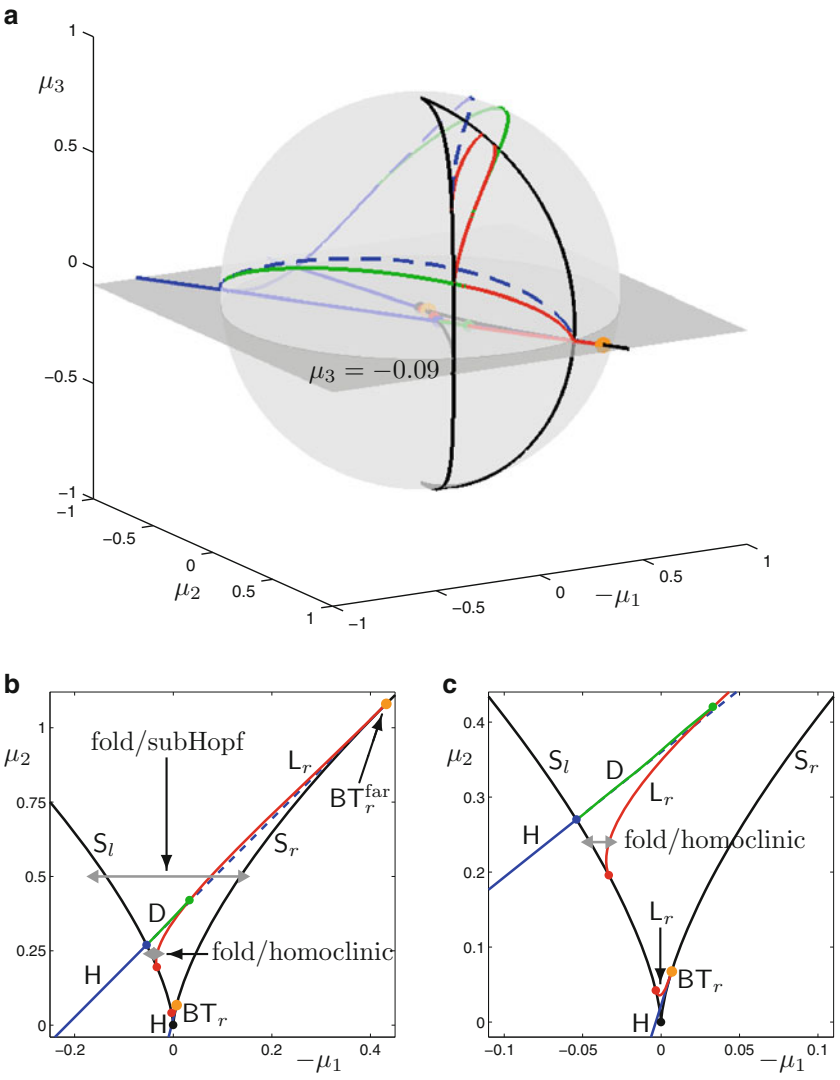


Fig. 12.4 A transition from fold/subHopf to fold/homoclinic bursting for system (12.1) can be obtained by considering the section $\mu_3 = -0.09$ for $\mu_4 = 0.75$. Panel (**a**) shows this section relative to the unit sphere, and panel (**b**) and the enlargement (**c**) show the bifurcation diagram on this section, together with parameter paths giving rise to fold/subHopf and fold/homoclinic bursting. Reproduced with permission from [14]. ©2012 American Institute of Mathematical Sciences. All rights reserved

bifurcation diagram (M) of the nilpotent focus of codimension three can already be found inside this given sphere on a sufficiently small sphere close to the central singularity! This observation means, in particular, that one finds the entire transition

of bifurcation diagrams from case (A) to case (M) on nested spheres when one reduces the radius R down to zero.

Of course, it is also natural to keep the radius of the chosen sphere of interest constant, say, again at $R = 1$. As μ_4 is increased from 0, case (M) can be found on larger and larger spheres until it can be found on the chosen sphere. Hence, the entire transition is "pushed through" the chosen sphere. In other words, increasing μ_4 while considering a sphere of a given radius is equivalent in this sense with decreasing the radius of the sphere considered while keeping $\mu_4 > 0$ constant.

Another consequence of this observation is the following. For $\mu_4 = 0$ the bifurcation diagram in (μ_1, μ_2, μ_3)-space has cone structure, so is topologically the same on any sphere. For $\mu_4 > 0$ it also has cone structure, but only in a small neighborhood of the origin, meaning that one finds case (M) of Fig. 12.1, the unfolding of the nilpotent focus of codimension three, on any sufficiently small sphere. Any of the other bifurcation diagrams (B) to (L) in Fig. 12.1, on the other hand, do not correspond to bifurcation diagrams in (μ_1, μ_2, μ_3)-space that have cone structure. In particular, this means that the exact sequence of transitions one finds from case (A) to case (M) depends on the properties of the family of closed convex surfaces around the origin (such as spheres, ellipses, or parallelepipeds).

Since it is arguably the most natural choice, we consider in what follows the bifurcation diagram on a sphere in (μ_1, μ_2, μ_3)-space, where we concentrate on the transition from about case (G) to case (M) in system (12.1); this corresponds to the transition from the sphere in Fig. 12.3, where fold/subHopf bursting was found, to the limiting case of the degenerate Bogdanov–Takens singularity of focus type.

We first present in Fig. 12.5 topological sketches of this transition, as observed numerically via the computation of bifurcation diagrams on spheres that will be presented next. In the topological sketches in Fig. 12.5 the projections are reflected with respect to the vertical axis when compared with Fig. 12.1; in other words, the view is from outside the sphere, so that the projections better resemble the bifurcation diagrams on the sphere shown in Figs. 12.3 and 12.4, and in similar figures below. The starting point in Fig. 12.5 is case (G'), which is as the bifurcation diagram in Fig. 12.3a. Case (G') lies "in between" cases (G) and (H) in Fig. 12.1 as far as the position of the Hopf curve H is concerned, but notice the absence of a cusp point on curve D. The curve H then crosses the end points of the curves L_b and L_r on S_l, yielding cases (H') and (I') of Fig. 12.5, respectively. Subsequently, there is a sign-change in the higher-order terms of the Bogdanov–Takens bifurcation BT_r to give case (K'), where the relative position of the curves H and L_r changes locally near BT_r. An important aspect is that there are now three degenerate Hopf bifurcation points on the curve H. The one inside the area bounded by S_l, and S_r then moves through S_l to give case (K"). The associated curve D of double periodic orbits then disappears when the respective two degenerate Hopf points that bound it come together and disappear; this codimension-three doubly degenerate Hopf point does not seem to involve additional bifurcations, but its further analysis is beyond the scope of this contribution. The final result is case (M), the bifurcation diagram of the degenerate Bogdanov–Takens singularity of focus type.

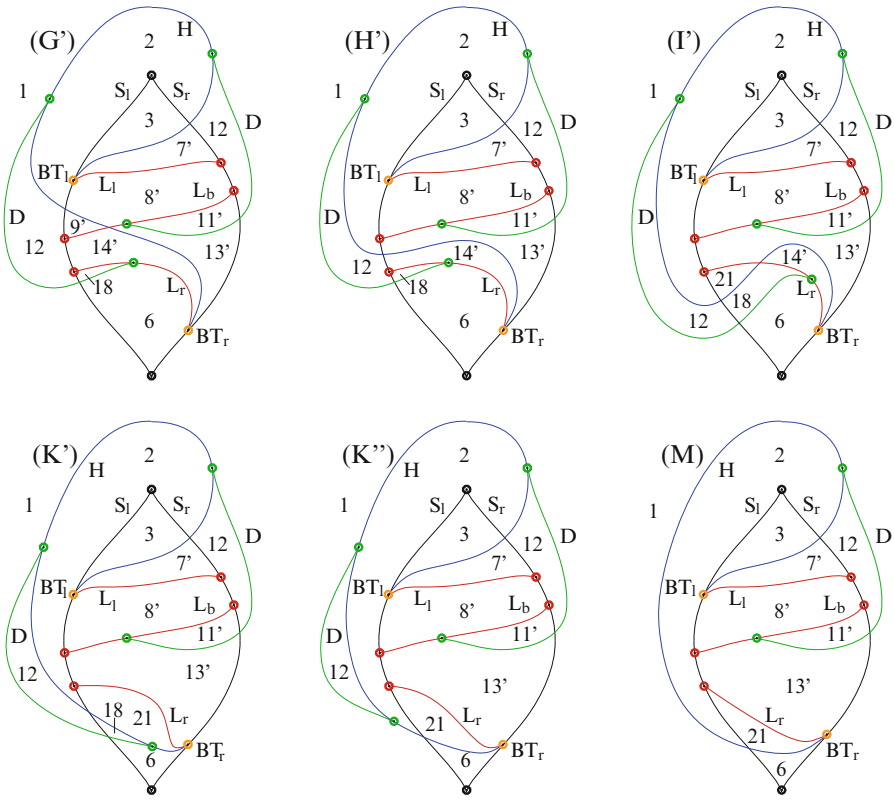


Fig. 12.5 Transition for increasing μ_4 as found numerically for system (12.1); shown are projections of unfoldings on a sphere in (μ_1, μ_2, μ_3)-space for fixed μ_4

Figure 12.6 presents numerical evidence of the transition in the form of images of computed bifurcation diagrams of system (12.1) for $\mu_4 = 1$ on spheres of radius $R = 1$, $R = 0.7$, $R = 0.5$, and $R = 0.2$; these computations were performed with the packages MATCONT [4] and AUTO [5]. The bifurcation diagram in Fig. 12.6a for $R = 1$ is as case (H') in Fig. 12.5. Figure 12.6b shows the bifurcation diagram on the sphere of radius $R = 0.7$, where the Hopf curve H has dipped below the end point of L_r on S_l, as is sketched in case (I') of Fig. 12.5. Figure 12.6c for $R = 0.5$ is past the type change of the Bogdanov–Takens point BT_r; moreover, the associated curve D is already quite short and lies entirely outside the region bounded by S_l, and S_r, as in case (K") of Fig. 12.5. Finally, for $R = 0.2$, as shown in Fig. 12.6d, we find case (M).

For illustration purposes, each sphere in Fig. 12.6 was rendered at the same size, irrespective of its actual radius. Figure 12.7, on the other hand, shows how the respective bifurcation diagrams are nested by rendering all four computed spheres in (μ_1, μ_2, μ_3)-space in one and the same image. Also shown is the vertical line of

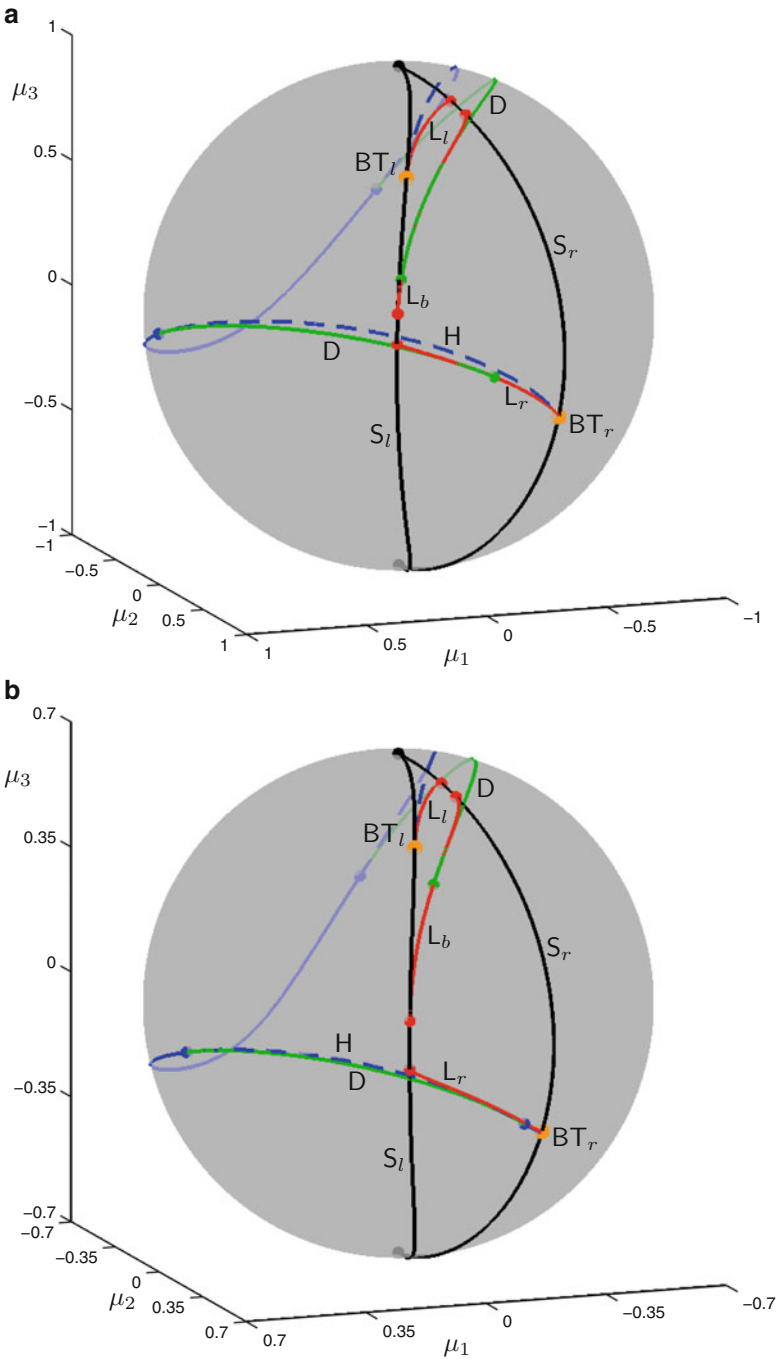


Fig. 12.6 Bifurcation diagrams of system (12.1) for $\mu_4 = 1$ on a sphere of radius R in (μ_1, μ_2, μ_3)-space; from (**a**) to (**d**), $R = 1$, $R = 0.7$, $R = 0.5$, and $R = 0.2$

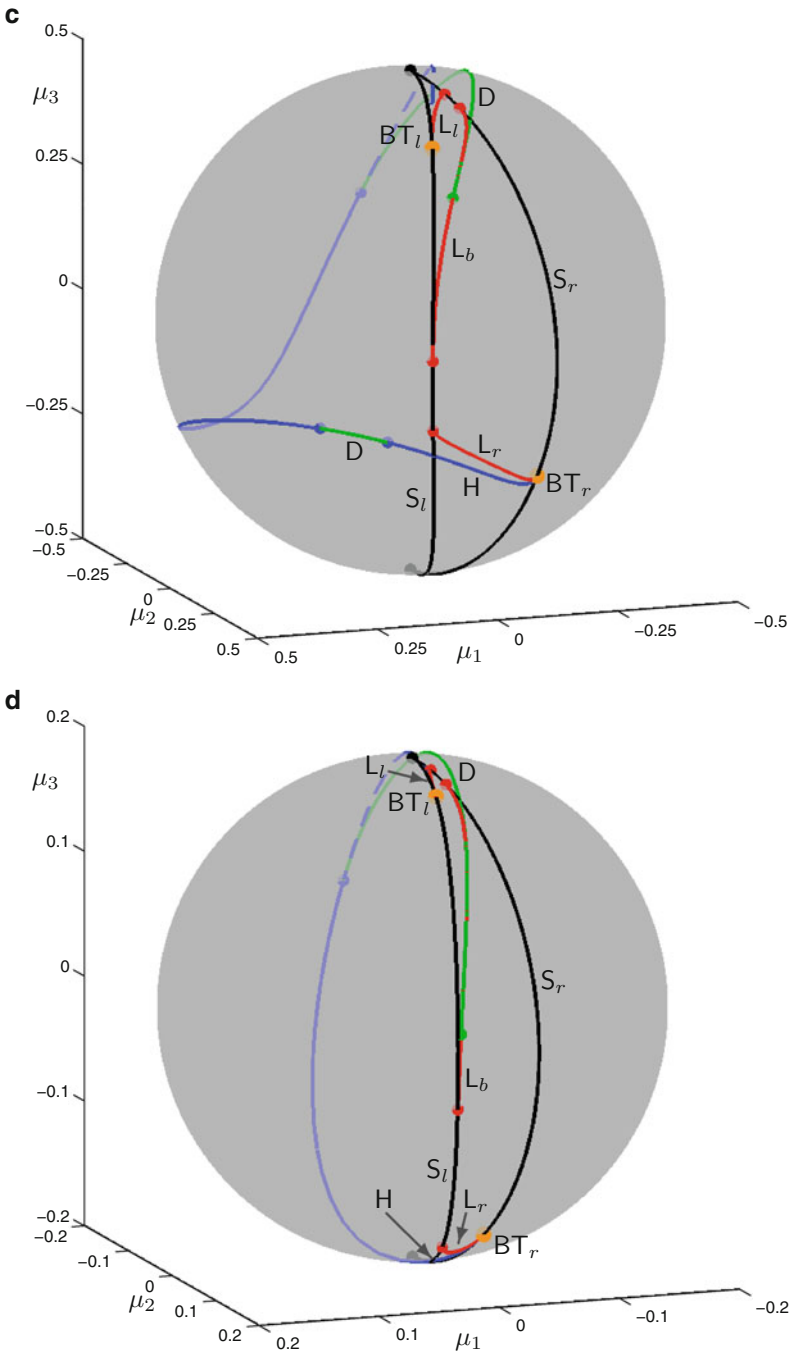


Fig. 12.6 continued

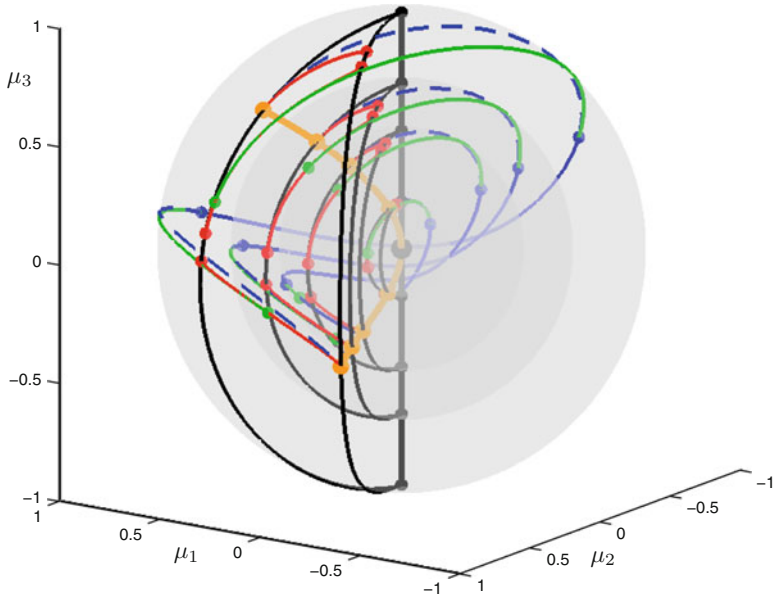


Fig. 12.7 The bifurcation diagrams of system (12.1) for $\mu_4 = 1$ on the nested spheres in (μ_1, μ_2, μ_3)-space of radius $R = 1$, $R = 0.7$, $R = 0.5$, and $R = 0.2$

cusp bifurcations and the curve of Bogdanov–Takens bifurcations, which meet in a tangency at the origin, that is, at the nilpotent focus of codimension three (since $\mu_4 = 1 > 0$). Taken together, Figs. 12.6 and 12.7 constitute numerical evidence in support of the revised transition presented in Fig. 12.5.

12.4 Conclusions

Unfoldings of codimension-four singularities of vector fields are sometimes seen as quite esoteric. The conjectural unfolding of codimension-four that was originally presented in 1998 was almost a bit of an afterthought in the paper [11], which deals with a codimension-three singularity that gives rise to symmetric bifurcation diagrams in planar sections nearby that had been found in numerous applications. Quite a number of years later, in 2012, it provided the solution presented in [14] to the question of where pseudo-plateau or fold/subHopf bursting can be found and whether and how it is connected to fold/homoclinic bursting.

The important aspect here is that the conjectural unfolding was presented in [11] as a sequence of bifurcation diagrams on spheres that constitutes the transition from the codimension-three unfolding considered in [11] to the well-known degenerate Bogdanov–Takens bifurcation of focus type that was known from [6]. As a result of the renewed interest in this transition we realized that the transition is, in some

sense, not so well defined. More specifically, the bifurcation diagrams found in the transition on convex surfaces (such as spheres or ellipses) are not uniquely defined due to the lack of cone structure. On the other hand, it is quite natural to consider spheres in parameter space, in which case an amended sequence of transitions can be determined with the help of numerical continuation tools. The associated bifurcation diagrams are encountered on nested spheres as soon as $\mu_4 > 0$ in (12.1), rather like Russian dolls. As μ_4 is increased they emerge one-by-one on a chosen fixed sphere, such as the unit sphere in (μ_1, μ_2, μ_3)-space.

We presented here only the part of the codimension-four unfolding that is relevant for generating the different types of bursting action potentials considered in [14]. Indeed, the complete transition between the codimension-three singularity for $\mu_4 = 0$ and the degenerate Bogdanov–Takens bifurcation of focus type can be represented in the same spirit in terms of bifurcation diagrams on nested spheres for $\mu_4 = 1$. The overall sequence of bifurcation diagrams, to be presented elsewhere, will shed light on the manifestation of the relevant bifurcations known from [11] and the study [3] of an alternative parameterization.

Acknowledgements The work presented here is quite directly related to work of and with Christiane Rousseau, and it is a pleasure to have this opportunity to thank her for explicit and implicit support and encouragement during many years. She was instrumental in getting us into unfoldings on spheres and compactifications of phase spaces, techniques that we keep using throughout our work. We have been enjoying meeting Christiane in many different places, including regularly during our visits to Montréal of course. We also thank our co-authors Alexander Khibnik, Arthur Sherman, and Krasimira Tsaneva-Atanasova, who have been great companions in this unfolding adventure.

References

1. Bazykin, A.D., Kuznetsov, Y.A., Khibnik, A.I.: Bifurcation Diagrams of Planar Dynamical Systems. Research Computing Center, Pushchino (in Russian) (1985, preprint)
2. Bertram, R., Butte, M.J., Kiemel, T., Sherman, A.: Topological and phenomenological classification of bursting oscillations. Bull. Math. Biol. **57**(3), 413–439 (1995)
3. Dangelmayr, G., Guckenheimer, J.: On a four parameter family of planar vector fields. Arch. Ration. Mech. **97**, 321–352 (1987)
4. Dhooge, A., Govaerts, W., Kuznetsov, Y.A.: MATCONT: a MATLAB package for numerical bifurcation analysis of ODEs. ACM Trans. Math. Softw. **29**(2), 141–164 (2003). Available via http://www.matcont.ugent.be/
5. Doedel, E.J.: AUTO: Continuation and Bifurcation Software for Ordinary Differential Equations. With major contributions from A.R. Champneys, T. F. Fairgrieve, Yu. A. Kuznetsov, B. E. Oldeman, R. C. Paffenroth, B. Sandstede, X. J. Wang and C. Zhang (2007). Available via http://cmvl.cs.concordia.ca/
6. Dumortier, F., Roussarie, R., Sotomayor, J.: Generic 3-parameters families of planar vector fields, unfoldings of saddle, focus and elliptic singularities with nilpotent linear parts. In: Dumortier, F., Roussarie, R., Sotomayor, J., Zoladek, H. (eds.) Bifurcations of Planar Vector Fields: Nilpotent Singularities and Abelian Integrals. Lecture Notes in Mathematics, vol. 1480, pp. 1–164. Springer, Berlin (1991)

7. Golubitsky, M., Josić, K., Kaper, T.J.: An unfolding theory approach to bursting in fast-slow systems. In: Broer, H.W., Krauskopf, B., Vegter, G. (eds.) Global Analysis of Dynamical Systems, pp. 277–308. Institute of Physics Publishing, Bristol (2001)
8. Guckenheimer, J., Malo, S.: Computer-generated proofs of phase portraits for planar systems. Int. J. Bifurcation Chaos **6**(5), 889–892 (1996)
9. Hodgkin, A.L., Huxley, A.F.: A quantitative description of membrane current and its application to conduction and excitation in nerve. J. Physiol. **117**(4), 500–544 (1952)
10. Hoppensteadt, F.C., Izhikevich, E.M.: Weakly Connected Neural Networks. Springer, New York (1997)
11. Khibnik, A.I., Krauskopf, B., Rousseau, C.: Global study of a family of cubic Liénard equations. Nonlinearity **11**(6), 1505–1519 (1998)
12. LeBeau, A.P., Robson, A.B., McKinnon, A.E., Sneyd, J.: Analysis of a reduced model of corticotroph action potentials. J. Theor. Biol. **192**(3), 319–339 (1998)
13. Malo, S.: Rigorous computer verification of planar vector field structure. Ph.D. thesis, Cornell University (1994)
14. Osinga, H.M., Sherman, A., Tsaneva-Atanasova, K.T.: Cross-currents between biology and mathematics: the codimension of pseudo-plateau bursting. Discrete Continuous Dyn. Syst. Ser. A **32**(8), 2853–2877 (2012)
15. Rinzel, J.: Bursting oscillations in an excitable membrane model. In: Sleeman, B.D., Jarvis, R.D. (eds.) Ordinary and Partial Differential Equations. Lecture Notes in Mathematics, vol. 1151, pp. 304–316. Springer, New York (1985)
16. Rinzel, J.: A formal classification of bursting mechanisms in excitable systems. In: Gleason, A.M. (ed.) Proceedings of the International Congress of Mathematicians, vols. 1, 2, pp. 1578–1593. American Mathematical Society, Providence RI (1987); also (with slight differences) In: Teramoto, E., Yamaguti, M. (eds.) Mathematical Topics in Population Biology, Morphogenesis and Neuroscience. Lecture Notes in Biomathematics, vol. 71, pp. 267–281. Springer, Berlin (1987)
17. Stern, J.V., Osinga, H.M., LeBeau, A., Sherman, A.: Resetting behavior in a model of bursting in secretory pituitary cells: distinguishing plateaus from pseudo-plateaus. Bull. Math. Biol. **70**(1), 68–88 (2008)
18. Tsaneva-Atanasova, K., Sherman, A., van Goor, F., Stojilkovic, S.: Mechanism of spontaneous and receptor-controlled electrical activity in pituitary somatotrophs: experiments and theory. J. Neurophysiol. **98**(1), 131–144 (2007)
19. van der Pol, B.: A theory of the amplitude of free and forced triode vibrations. Radio Rev. **1**, 701–710 (1920)
20. van der Pol, B.: On relaxation oscillations. Lond. Edinb. Dublin Philos. Mag. Ser. 7 **2**, 978–992 (1926)
21. van Goor, F., Li, Y., Stojilkovic, S.: Paradoxical role of large-conductance calcium-activated K^+ BK channels in controlling action potential-driven Ca^{2+} entry in anterior pituitary cells. J. Neurosci. **21**(16), 5902–5915 (2001)
22. Wang, X., Kooij, R.E.: Limit cycles in a cubic system with a cusp. SIAM J. Math. Anal. **23**(6), 1609–1622 (1992)

Chapter 13
Towards the General Theory of Global Planar Bifurcations

Y. Ilyashenko

To Christiane Rousseau, a wonderful mathematician and organizer of scientific life, and a dear friend.

Abstract This is an outline of a theory to be created, as it was seen in April 2015. An addendnum to the proofs at the end of the chapter describes the recent developments.

Keywords Global bifurcations • Generic families • Large bifurcation supports • Polycycles

Theory of planar bifurcations has a long and glorious history. It may be split into two parts: local and nonlocal bifurcations. Local bifurcations appeared first in the works of Poincaré. The most famous of them is the Poincaré–Andronov–Hopf bifurcation. The second part deals with the bifurcations of separatrix polygons, the polycycles. The simplest ones are separatrix loops of hyperbolic saddles and homoclinic curves of saddle-nodes. This part may be also called "semilocal bifurcations" because the perestroikas occur in arbitrary narrow neighborhoods of the polycyles. After the first founding works of Andronov and his school, this part started to develop intensively since 1980s. We plan to show that there is a third part, not yet developed, that may be called "global bifurcations." The main new effect in this theory may be called "sparking saddle connections." They were discovered by Malta–Palis in the early 1980s and described below.

This survey is aimed to outline the first steps in the development of this theory. All the new theorems below are "theorems": the proofs are not yet written.

Y. Ilyashenko (✉)
National Research University Higher School of Economics, Moscow, Russia

Department of Mathematics, College of Arts and Sciences, Cornell University, Ithaca, NY, USA

Independent University of Moscow
e-mail: yulijs@gmail.com

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_13

13.1 Global Bifurcations in Generic One-Parameter Families

13.1.1 Basic Definitions

In 1985 Arnold suggested a program of development of the global bifurcation theory in the plane. Begin with some classical notions necessary to understand his text quoted below.

Definition 1. Let M be a manifold, not necessary closed, and B be a parameter space, a ball in $\mathbb{R}^k$. Two families of vector fields $\{v_\varepsilon\}, \{w_\varepsilon\}$ on M with the parameter space $B \ni \varepsilon$ are topologically equivalent provided that there exists a skew product homeomorphism

$$H : B \times M \to B \times M,$$
$$(\varepsilon, x) \mapsto (h(\varepsilon), H(\varepsilon, x)),$$

where h is a homeomorphism $B \to B$, such that for any $\varepsilon \in B$ the homeomorphism $H(\varepsilon, \cdot)$ is an orbital topological equivalence between the vector fields v_ε and $w_{h(\varepsilon)}$.

Definition 2. Two families above are weakly equivalent if in the previous definition H is no more a homeomorphism . Namely, H is no more continuous in ε, remaining a homeomorphism of M to M for any fixed ε.

Definition 3. A local family of vector fields on M is a germ on $M \times \{0\}$ of families $\{v_\varepsilon(x)\}$, $x \in M$, $\varepsilon \in (B, 0)$. That is, the base B is replaced by a germ of a base $(B, 0)$. Local topological equivalence and weak equivalence of local families is a correspondent equivalence of some representatives of these families provided that the corresponding homeomorphism of the bases brings the critical parameter value zero of one base to that of another.

Definition 4. A local family of vector fields on M is globally (weakly) structurally stable provided that it is (weakly) topologically equivalent to all the nearby families. The term *globally* may be omitted. It recalls that the family is considered in the whole phase space.

13.1.2 Arnold's Program

The text in this subsection, except for the last sentence is a quotation from the survey [1].

Although even local bifurcations is high codimensions (at least three) on a disc are not fully investigated, it is natural to discuss nonlocal bifurcations in multiparameter families of vector fields on a two-dimensional sphere. For their

description, it is necessary to single out the set of trajectories defining perestroikas in these families.

Definitions and Examples (V.I. Arnol'd 1985)

Definition 5. A finite subset of the phase space is said to *support a bifurcation* if there exists an arbitrarily small neighborhood of this subset and a neighborhood of the bifurcation values of the parameter (depending on it) such that, outside this neighborhood of the subset, the deformation (at values of the parameter from the second neighborhood) is topologically trivial.

Example 1. Any point of a saddle connection (including both saddles) supports a bifurcation, even if one adds to it any other points. In a system with two saddle connections an interior point on one connection supports a bifurcation only with a point on the other connection.

Definition 6. The *bifurcation support* of a bifurcation is the union of all minimal sets supporting a bifurcation ("minimal" means not containing a proper subset that supports a bifurcation).

Example 2. In a system with one saddle connection (bifurcating in a standard way), the support coincides with the saddle connection, including its endpoints, the saddles.

Definition 7. Two deformations of vector fields with bifurcation supports Σ_1 and Σ_2 are said to be *equivalent on their supports* if there exist arbitrarily small neighborhoods of the supports, and neighborhoods of the bifurcation values of the parameters depending on them, such that the restrictions of the families to these neighborhoods of the supports are topologically equivalent, or weakly equivalent, over these neighborhoods of bifurcation values.

Example 3. All deformations of vector fields with a simple saddle connection are equivalent to each other, independent of the number of hyperbolic equilibria or cycles in the system as a whole.

Example 4. Four-parameter deformations of a vector field close to a cycle of multiplicity four are weakly topologically equivalent, but, generally, not equivalent: the classification of such deformations with respect to topologically equivalence involves functional invariants.

Conjecture (V.I. Arnol'd 1985). For a generic l-parameter family of vector fields on S^2:

1) *On their supports, all deformations are equivalent to a finite number of deformations (the number depends only upon l).*
2) *Any bifurcation diagram is (locally) homeomorphic to one of a finite number (depending only upon l) of generic examples.*
3) *There exist versal and weakly structurally stable deformations.*
4) *The family is globally weakly structurally stable.*

5) *The bifurcation supports consist of a finite number (depending only upon l) of (singular) trajectories.*
6) *The number of points in a minimal supporting set is bounded by a constant depending only on l.*

Certainly proofs or counterexamples to the above conjectures are necessary for investigating nonlocal bifurcations in generic l-parameter families.

The bifurcation supports defined in this subsection will be sometimes called *small supports*, because *large supports* will be defined below.

13.1.3 Sparking Saddle Connections

The key feature of the global planar bifurcations are the connections named in the title.

The simplest example of sparking saddle connections occurs for one semistable cycle, with two hyperbolic saddles: one inside and one outside the cycle. The following theorem appears first in [16]; we quote it from [7].

Theorem 1. *Suppose that a vector field X in the plane contained in a generic one-parameter family $X_\varepsilon (X_0 = X)$ has a semistable limit cycle L. Let this field has two hyperbolic saddles: one inside and the other outside the cycle. Suppose that the separatrix of one saddle winds to the cycle as $t \to +\infty$ and the separatrix of the other one does not the same as $t \to -\infty$. Then on one side of $\varepsilon = 0$, there exist two limit cycles that tend to L as $\varepsilon \to 0$: one is stable and the other is unstable. For ε on the other side of $\varepsilon = 0$, there exist no limit cycles near L. Moreover, there exists a sequence of parameter values of the form*

$$\varepsilon_n = \frac{1}{n^2}(c + o(1)),\ c \neq 0,$$

such that the field X_{ε_n} has a saddle connection for any n large enough.

The bifurcation diagram is not an isolated point (as it happens in all the classical examples of generic one-parameter families), but rather a sequence of points converging to the critical parameter values. Subsequent points in this sequence are marked by subsequent natural numbers: $\varepsilon_m, \varepsilon_{m+1}, \ldots$. The vector field corresponding to ε_n has a saddle connection that makes n circuits around the interior saddle, before closure.

13.1.4 Another Kind of Sparking Saddle Connections

Breaking of a homoclinic loop of a saddle may also generate sparking saddle connections.

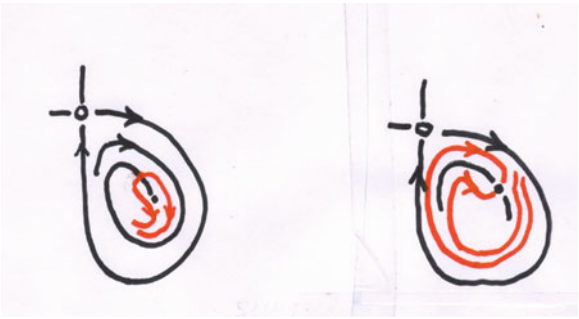


Fig. 13.1 Two phase portraits of a separatrix loop with a saddle inside corresponding to the critical parameter values; the sinks inside the loops are not shown

Theorem 2. *Suppose that a planar vector field X met in a generic one-parameter family has a separatrix loop, and the saddle value (trace of the linearization) is negative. Suppose that this vector field has a unique saddle inside the separatrix loop, with one or two incoming separatrixes winding towards the separatrix loop of the first saddle in the negative time, and exactly two other singular points, a sink and a source, inside the loop, see Fig. 13.1. Then on one side of the critical parameter value the field has a stable hyperbolic cycle with one saddle inside and one outside. On the other side of the critical value there is a countable number of bifurcation points related to saddle connections between the two saddles mentioned above. The number of circuits of these connections around the interior saddle tends to infinity as the parameter tends to the critical value.*

13.1.5 The Definitions Revisited

Let us describe the bifurcation support of the Malta–Palis bifurcation. A minimal set supporting the bifurcation consists of one point on the semistable cycle. Indeed, for any neighborhood of this point there exists a neighborhood of the critical parameter value such that for any non-zero bifurcation parameter value from the second neighborhood the corresponding sparkling saddle connections cross the first neighborhood . The representative of the local Malta–Palis family having the second neighborhood as the base, with the first neighborhood deleted from the phase space, is topologically trivial.

The union of all the minimal sets supporting the bifurcation, that is, the bifurcation support, is the semistable cycle only. In its small neighborhood sparkling saddle connections are not visible at all. So the bifurcations in this neighborhood do not describe the bifurcations in the Malta–Palis family.

Consider another example, namely, a vector field with a homoclinic trajectory of a saddle-node of multiplicity two in assumption that this field occurs in a typical one-parameter family. The minimal set supporting the bifurcation is unique in this case. It is the saddle-node singular point itself. Indeed, the vector field in any

domain that contains this point is structurally unstable. On the other hand, the same vector field in the domain with a neighborhood of the saddle-node deleted is structurally stable because it is met in a typical one-parameter family and thus has no more degeneracies. Hence, the bifurcation support consists of one point, the saddle-node itself. Yet, the bifurcation, the generation of a limit cycle, happens in a neighborhood of the whole homoclinic curve. The bifurcation support is not relevant to this bifurcation.

This is a reason to introduce new definitions.

Below we give a definition of a large bifurcation support. It is motivated by the following natural question: what does it mean that two bifurcations in two local families of vector fields on the 2-sphere are essentially the same? The answer: *the local families are weakly equivalent* makes no sense. We may extend two phase portraits with the same bifurcation by different structurally stable elements, and the local families would become nonequivalent.

The definition of the *large bifurcation support* below is aimed to answer the question above. It is adjusted, in particular, to the bifurcations of sparking saddle connections and homoclinic curves of saddle-nodes. We deal first with nonlocal families, then with local ones.

Definition 8. Consider a family of vector fields on a sphere S^2 with a parameter base B. Let $D \subset B$ be the bifurcation diagram of the family. A closed set $C \subset D \times S^2$ is called *a bifurcation carrier* if for any neighborhood U of C and for any point $b \in D$ the corresponding local family on $(B, b) \times S^2 \setminus U$ is topologically trivial; moreover, the carrier is the minimal closed set with this property.

Example 5. A bifurcation carrier in the Malta–Palis bifurcation is a countable set having exactly one point on each of the sparkling separatrixes, and one point of the semistable cycle.

Definition 9. A large bifurcation support of type one for a nonlocal family is a minimal closed set that contains all the carriers of the bifurcations in the family.

Example 6. The large bifurcation support of type one in the Malta–Palis family consists of all the sparkling saddle connections with the saddles included, plus the semistable cycle, all located in $D \times S^2$ over the corresponding bifurcation values. This support consists of an infinite number of the phase curves in the family.

We do not define the large bifurcation support of type two for nonlocal families, and pass to local families instead.

Definition 10. A large bifurcation support of type one for a local family with zero critical value of the parameter is a minimal closed set $\Sigma \subset \{0\} \times S^2$ with the following properties. For any neighborhood U of Σ in S^2 there exists a neighborhood V of 0 in B such that the large bifurcation support of the representative of the local family with the base V belongs to $V \times U$.

Example 7. The large bifurcation support of type one in the local Malta–Palis family consists of the semistable cycle and the separatrixes that wind to and from

it, the corresponding saddles included. This support consists of a finite number of orbits, namely, five: two saddles, two separatrixes, and a cycle.

Example 8. There are two different large bifurcation supports of type one for the local family described in Theorem 2. They are shown in Fig. 13.1.

This is the first part of the definition of the large bifurcation support. The second part is the following:

Definition 11. A singular curve of a vector field on a two sphere is a curve for which the phase portrait of the field in any neighborhood of the closure of the curve is topologically nonequivalent to that for any other phase curve with a nearby initial condition.

Example 9. A separatrix of a hyperbolic saddle, or a boundary curve of a parabolic sector of a saddle-node, is a singular curve.

Definition 12. Consider a local family and take all the non-hyperbolic singular points of the critical vector field. Consider all the sequences of cycles and singular curves in the product $B \times S^2$ that correspond to the parameter values converging to zero, and whose distance to at least one of these singular points tends to zero. The upper topological limit of these sequences (the set of all points whose arbitrary neighborhoods intersect infinitely many terms of the sequence) constitutes the large bifurcation support of type two.

Example 10. The large bifurcation support of type two for the polycycles *apple* and *halfapple* shown in Fig. 13.7 below consists of these polycycles.

Definition 13. A large bifurcation support for a local family is the union of the corresponding large bifurcation supports of type one and type two.

Definition 14. The bifurcations in two local families are called *equivalent*, if these families are weakly equivalent in some neighborhoods of their large bifurcation supports, and the linking homeomorphism over each base point is an isotopy, that is, may be extended to the homeomorphism of the whole sphere.

Problem 1. *Prove that for any k there is an open and dense set in the space of k-parameter local families of vector fields in the two sphere such that for any fixed family from this set the following holds. There exists a neighborhood of the fixed family such that for any two local families from this neighborhood the week topological equivalence of these two families in some neighborhoods of their large supports implies the same equivalence of the families on the whole sphere, provided that the vector fields corresponding to the critical parameter values are orbitally topologically equivalent.*

Theorem 3. *All the bifurcations that occur in generic nonlocal one-parameter families of vector fields on the two sphere have at most countable bifurcation carriers.*

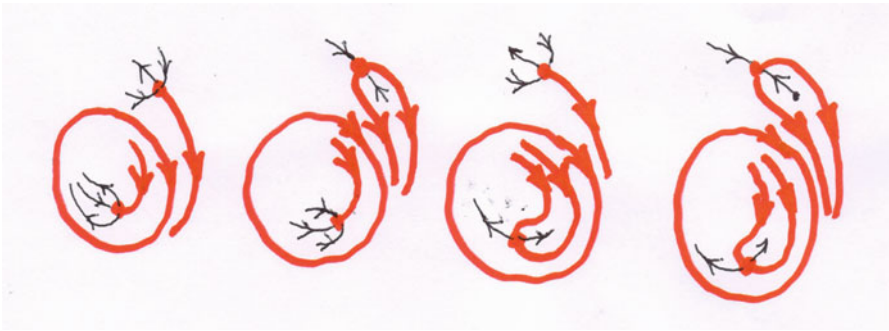


Fig. 13.2 Different locations of Cherry cells

Theorem 4. *All the generic local one-parameter families of vector fields in S^2 have large bifurcation supports that consist of a finite number of phase curves.*

Theorem 5. *There are exactly two classes of topological equivalence of bifurcations in the local families described in Theorem 2. Their large bifurcation supports are shown in Fig. 13.1.*

13.1.6 Classification of Global Bifurcations in the Local One-Parameter Families on the Sphere

In all the classification theorems below the bifurcations in local families are considered in some neighborhood of their large supports.

Theorem 6. *There are exactly six generic one-parameter families in the plane, up to topological equivalence, whose "small" representatives, that is nonlocal families corresponding to sufficiently small neighborhoods of the critical parameter value, have finite carriers. These carriers consist of exactly one point.*

These six bifurcations are all classical:
- breaking of a saddle-node singular point having no homoclinic loop;
- breaking of a saddle-node singular point having a homoclinic loop;
- Andronov–Hopf bifurcations ;
- vanishing or splitting of a semistable limit cycle;
- breaking of a separatrix loop of a hyperbolic saddle;
- breaking of a saddle connection of two different saddles.

All other bifurcations occur due to sparkling saddle connections.

Note that an arbitrary finite number of saddles may be involved in the formation of sparking saddle connections related both to semistable cycles and to saddle connections. So, an infinite number of pairwise topologically nonequivalent generic 1-parameter families occurs, see Fig. 13.3.

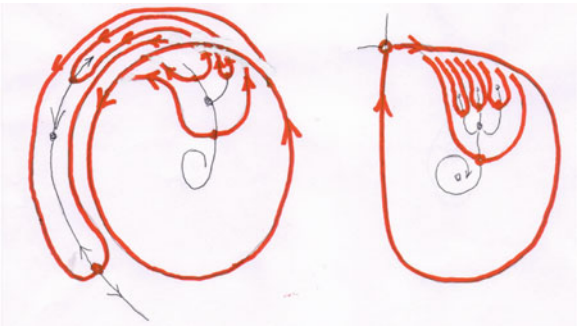


Fig. 13.3 Complicated large bifurcation supports

Theorem 7. *There is an infinite number of local one-parameter families of vector fields on the two sphere pairwise topologically nonequivalent on their large supports. There are two classes of them having an infinite bifurcation diagram: those that correspond to semistable cycle and to a separatrix loop with sparkling saddle connections. The large supports of the corresponding bifurcation consist of the cycle (in the first case), the separatrix loop (in the second case), and the separatrixes of the hyperbolic saddles that wind onto them either in the positive, or in the negative time, in both cases. Two local one-parameter families of vector fields on a two sphere whose large bifurcation support contains a separatrix loop are topologically equivalent if the large supports of the corresponding bifurcations are isotopic: may be transformed one into another by a homeomorphism of the ambient sphere, and the vector fields corresponding to the critical parameter values are orbitally topologically equivalent.*

There is but a finite number of classes of topological equivalence of local one-parameter families whose large bifurcation supports contain a semistable limit cycle and are isotopic, provided that the vector fields corresponding to the critical parameter values are orbitally topologically equivalent.

The large supports in the theorem above always consist of a finite number of phase curves. Examples are shown in Fig. 13.3. The combinatorics of these supports may be very complicated. May be, it might be described by some oriented graphs.

Example 11. The large support of the bifurcation in the Malta–Palis family consists of the semistable cycle and the separatrixes of two saddles that wind onto it from inside and from outside. Four different cases occur, see Fig. 13.2.

Note that Cherry cells enter the game, and their different location corresponds to topologically nonequivalent families.

We conclude this section by the following:

Conjecture 1. If two generic one-parameter local families of vector fields on the two sphere are equivalent in the neighborhoods of their large supports, then they are also equivalent in the basins of the attraction/repulsion of their (small) supports.

13.2 Bifurcations in Two-Parameter Families

13.2.1 Local Bifurcations

There is one (and only one!) local bifurcation in the two-parameter families that is quite new in comparison with those previously studied. This is the Bogdanov–Takens bifurcation. Its investigation was a revolution in the bifurcation theory, and opened a new period of its development.

Note that other famous two-parameter families in the plane:

the families investigated by Zoladek;

the so-called resonances 1:2, 1:3 investigated by Horozov;

the resonance 1:4, investigated by many authors but not yet fully studied,

are factorizations of higher dimensional problems, and do not belong to the subject of our survey.

More traditional are problems that occur in codimension 1, but have supplementary degeneracies. These are bifurcations of saddle-nodes and Andronov–Hopf bifurcations. Nothing interesting occurs in these families with two parameters. Three singular points may be generated in the first family instead of two. Two limit cycles occur in the second family instead of one.

Let us say a few words about the multiparameter case. Floris Takens investigated generation of limit cycles in the unfolding of a vector fields

$$v(z) = iz + az^{k+1}\bar{z}^k + \dots, \ a \neq 0 \tag{13.1}$$

written in its resonant normal form. Such a field occurs in generic k-parameter families. Its unfolding generates no more than k limit cycles.

An unfolding of a vector field on a line

$$v(x) = ax^{k+1} + \dots, \ a \neq 0$$

that occurs in generic k-parameter families can generate no more that $k + 1$ singular points.

The phase portraits of these families in both cases are easily investigated.

A striking discovery was made by Roussarie: topological classification of these families has functional moduli provided that the number of parameters is high enough: three for the Andronov–Hopf family and four for the bifurcation of a multiple limit cycle.

Theorem 8 (Roussarie [18]). *For a generic three-parameter unfolding of a vector field* (13.1) *with* $k = 3$ *there exists a functional invariant of topological classification. This invariant is a one-parameter family of diffeomorphisms of a cycle.*

On the other hand, any generic unfolding of a germ (13.1) is *weakly* topologically equivalent to one of a *finite* number of "standard" local families. So in what follows we speak about weak equivalence only.

Fig. 13.4 A rigged loop: a separatix loop of a hyperbolic saddle with a zero saddle value (trace of eigenvalues)

Fig. 13.5 A lune (*left*); a heart (*right*)

Fig. 13.6 An eight shaped figure (collection of three polycyles: two separatrix loops and their union)

Fig. 13.7 An "apple and halfapple"

13.2.2 Semilocal Bifurcations in Two-Parameter Families

At the end of 1980s a Moscow graduate student Anna Kotova collected a "zoo" of all polycycles that may occur in generic two- and three-parameter families [14]. There are "individual polycycles" (separatrix polygons homeomorphic to a circle), "collections of polycycles" (finite unions of individual polycycles), and one "ensemble": a continuous family of polycycles that occur in a generic three-parameter family. We postpone the description of this ensemble to the next section.

Later on, S. Trifonov investigated the cyclicity of all individual polycyles in the "Kotova zoo," [22].

Here we present these results for codimension two. All the polycycles below have there own names. They are shown in Figs. 13.4, 13.5, 13.6, 13.7, 13.8, 13.9 and listed in the figure captions.

Trifonov proved that no polycycle in this list has cyclicity larger than 2. On the other hand, my former student Grozovski investigated the bifurcations of the collection "apple,"and found that three limit cycles may be generated by a two-

Fig. 13.8 Boundary homoclinic loop of a saddle-node

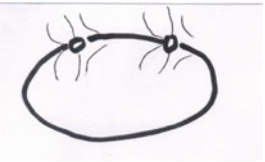


Fig. 13.9 Twin saddle-nodes

parameter unfolding of this collection. This is the simplest case when the number of cycles generated is larger than the number of parameters.

Let us say a few words about the bifurcations of a separatrix loop. It is investigated now in full generality for the families with an arbitrary number of parameters. In modern terms the result is the following:

Theorem 9. *A separatrix loop that occurs in a generic k-parameter family may generate no more than k limit cycles.*

This result was obtained by Andronova–Leontovich in the late 1940s; the sketch of the proof with the main ideas was published in [15]. Unfortunately, she never published the full proof. A complete proof of this result was obtained by Roussarie (an upper estimate) [17]. In [8] its sharpness was proved.

13.2.3 Polycycles and Sparkling Separatrixes

Conjecture 2. Sparkling separatrixes may occur for all the polycyles listed above, except for the twin saddle-node.

Problem 2. *Describe the corresponding global bifurcations. In particular, are Arnold's Conjectures 3 and 4 true for two-parameter families?*

13.2.4 Synchronized Sparkling Saddle Connections

Consider a semistable limit cycle with two saddle separatrixes winding on it from outside and two winding from inside, Fig. 13.10.

The question is: when two saddle connections may occur simultaneously under the bifurcation in this family?

As proved in [16], this cannot happen in generic one-parameter families. But it can happen in two-parameter ones [7, 9]. Let us describe this bifurcation in more detail.

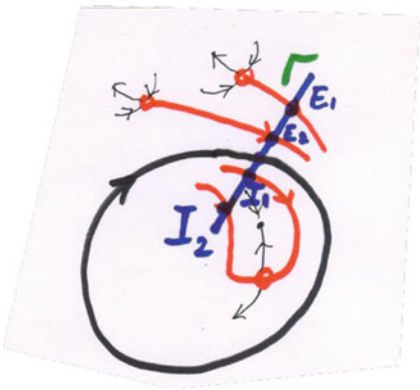


Fig. 13.10 Large bifurcation support for a family with synchronized saddle connections

Consider the parameter depending Poincaré map P of the semistable cycle corresponding to some transversal Γ. Denote the two parameters by ε, δ. Then the Poincaré map will be $P(x, \varepsilon, \delta),\ x \in \Gamma$. Suppose that ε is "responsible" for the breaking of the semistable cycle: the semistable cycle corresponds to $\varepsilon = 0$ and vanishes for $\varepsilon > 0$. By Ilyashenko and Yakovenko [10], there exists a vector field $w_{\varepsilon,\delta}$ on Γ that generates $P(\cdot, \varepsilon, \delta)$ as a time one phase flow transformation in the domain $\varepsilon \geq 0$, where the cycle vanishes. Moreover, the coordinate x on Γ and the parameters may be so chosen that

$$w_{\varepsilon,\delta}(x) = \frac{x^2 + \varepsilon}{1 + a(\varepsilon, \delta)x}.$$

Let $E_j(\varepsilon, \delta)$ and $I_j(\varepsilon, \delta)$ be the x-coordinates of the intersections of the separatrixes with Γ, continuous in ε, δ. Separatrixes passing through $E_j(\varepsilon, \delta)$ and$I_j(\varepsilon, \delta)$ coincide iff for some natural k,

$$P^k(E_j(\varepsilon, \delta),\ \varepsilon, \delta) = I_j(\varepsilon, \delta).$$

This is equivalent to

$$g^k_{w_{\varepsilon,\delta}}(E_j(\varepsilon, \delta)) = I_j(\varepsilon, \delta). \tag{13.2}$$

If this happens simultaneously for $j = 1$ and 2, then the two separatrixes coincide simultaneously, they "meet" after k turns from E_j to I_j. Consider a "time function" corresponding to the field $w_{\varepsilon,\delta}$ ($x_0 \in \Gamma$ is arbitrary):

$$T(x, \varepsilon, \delta) = \int_{x_0}^{x} \frac{d(\xi)}{w_{\varepsilon,\delta}(\xi)} \tag{13.3}$$

Equation (13.2) is equivalent to

Fig. 13.11 The bifurcation diagram in a family with two synchronized saddle connections

$$T(I_j(\varepsilon,\delta),\varepsilon,\delta) - T(E_j(\varepsilon,\delta),\varepsilon,\delta) = k. \tag{13.4}$$

If these equalities hold for a sequence $(\varepsilon_k,\delta_k) \to 0$ as $k \to \infty$, we have a sequence of simultaneous (twin) saddle connections corresponding to (ε_k,δ_k); the number of winds of this connections near the cycle that have vanished is k. Equation (13.4) for $j = 1$ and 2, imply

$$T(E_1(\varepsilon_k,\delta_k),\varepsilon_k,\delta_k) - T(E_2(\varepsilon_k,\delta_k),\varepsilon_k,\delta_k) = T(I_1(\varepsilon_k,\delta_k),\varepsilon_k,\delta_k) - T(I_2(\varepsilon_k,\delta_k),\varepsilon_k,\delta_k).$$

Passing to the limit, as $\varepsilon \to 0,\ \delta \to 0$, we get

$$T(E_1,0,0) - T(E_2,0,0) = T(I_1,0,0) - T(I_2,0,0).$$

Consider a function

$$S(\varepsilon,\delta) = [T(E_1,(\varepsilon,\delta),\varepsilon,\delta) - T(E_2(\varepsilon,\delta),\varepsilon,\delta)] - [T(I_1(\varepsilon,\delta),\varepsilon,\delta) - T(I_2(\varepsilon,\delta),\varepsilon,\delta)]. \tag{13.5}$$

The previous equality implies

$$S(0,0) = 0.$$

Let $\frac{\partial S}{\partial \varepsilon}(0,0) \neq 0$. Then the "synchronization curve" $S = 0$ is transversal to $\varepsilon = 0$. Hence, the synchronization curve intersects transversally the curves of saddle connections between E_1 and I_1 making k turns, $k \to \infty$. The intersection points correspond to synchronized connections between E_2, I_2 and E_1, I_2. The bifurcation diagram is shown in Fig. 13.11.

We will turn back to this bifurcation in the study of quasigeneric families.

13.2.5 Sparkling Saddle Connections for Two Semistable Cycles

In two-parameter families two semistable cycles may occur. Any finite number of saddles may be added "for free." Consider first the bifurcation with two separate semistable cycles and four saddles involved, see Fig. 13.12.

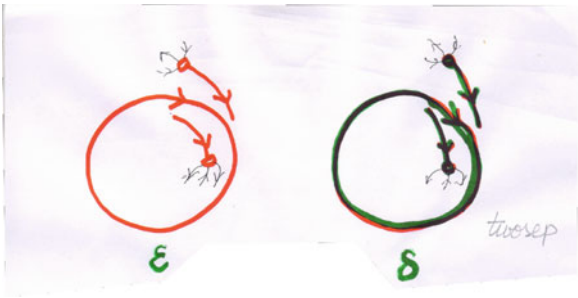


Fig. 13.12 Two separate semistable cycles

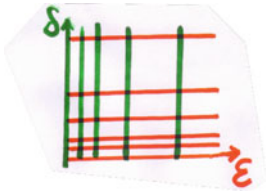


Fig. 13.13 Bifurcation diagram in a family with two separate semistable cycles, domain $\varepsilon \geq 0,\ \delta \geq 0$

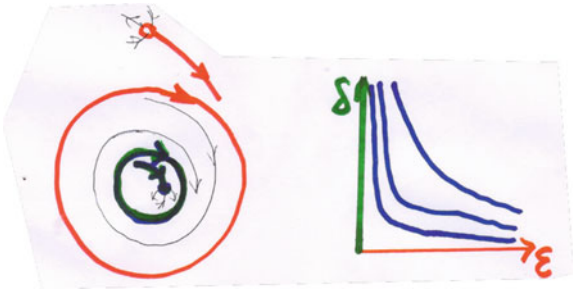


Fig. 13.14 Two semistable cycles, one inside another

Let $\varepsilon = 0$ correspond to the left semistable cycle, and $\delta = 0$ to the right one. The bifurcation diagram in the domain $\varepsilon \geq 0,\ \delta \geq 0$ is shown in Fig. 13.13.

Second, consider two semistable cycles one inside another, with two saddles, one outside the larger one, another inside the smaller one. Sparkling saddle connections will occur when both cycles disappear, Fig. 13.14.

13.2.6 Synchronized Connections and Complicated Bifurcation Diagrams in Two-Parameter Families

Consider now a more complicated case: two semistable cycles one inside another, with a saddle I inside, E outside, and B between them.

The large bifurcation support of this family is schematically shown in Fig. 13.15. The bifurcation diagram is presented in Fig. 13.15 too.

The "horizontal" curves correspond to connections between the saddles I and B. Black vertical curves correspond to connections between B and E that involve

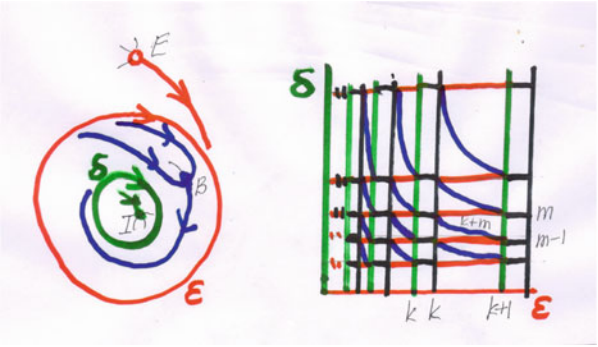
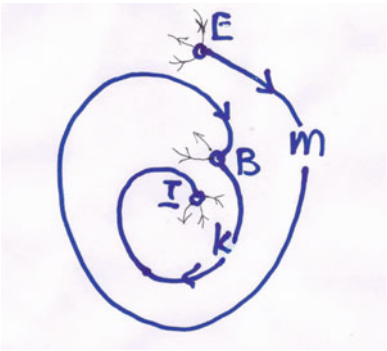


Fig. 13.15 More complicated location of saddles and bifurcation diagram in the domain $\varepsilon \geq 0,\ \delta \geq 0$

Fig. 13.16 Simultaneous saddle connections in the family considered



the lower separatrix L of B; the red ones correspond to those that involve the upper separatrix U.

Any of these lines is marked by an integer number: a number of full turns made by the connection around the interior saddle I. The intersections between vertical and horizontal curves correspond to simultaneous saddle connections, see Fig. 13.16.

There are three connections in Fig. 13.16: one between I and B, one between B and E, and the third is a compound connection between I and E, the union of the previous two.

There are also hyperbola-shaped arcs in the bifurcation diagram that correspond to the connections between E and I. They are marked by the number n of full circuits that they make around I. These curves pass through the points of synchronized connections with indexes k and m; in this case, $n = k + m$. We call them "arcs of long connection."

There are alternating thick and thin arcs on the horizontal curves. Thin ones correspond to the case when the unstable separatrix of E enters the interior domain of the (vanished) small semistable cycle; thick ones correspond to the case when this separatix enters the Cherry cell of B, see Fig. 13.17.

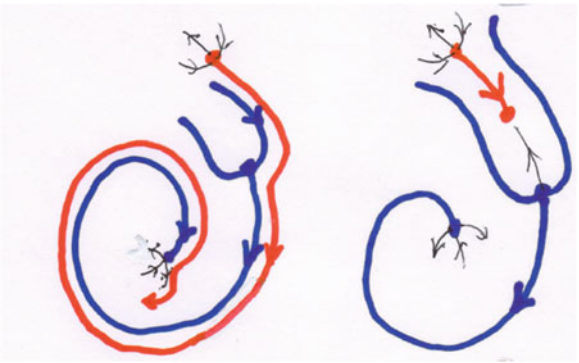


Fig. 13.17 Role of Cherry cells

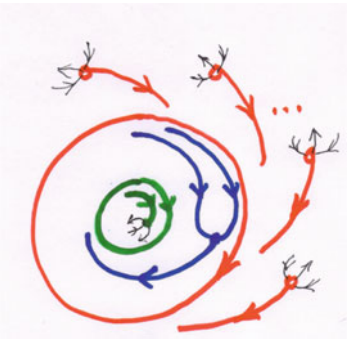


Fig. 13.18 Infinite series of large supports that correspond to an infinite set of topologically nonequivalent germs of bifurcation diagrams

13.2.7 An Infinite Number of Samples of the Bifurcation Diagrams

Theorem 10. *There exists an infinite number of topologically nonequivalent germs of bifurcation diagrams in generic two-parameter families of vector fields in the two sphere.*

Sketch of the Proof. As an example one may suggest a series of local families with the large bifurcation supports consisting of two semistable cycle, one saddle I inside both, one in between, and j saddles outside both cycles. The large support of the corresponding bifurcation is shown in Fig. 13.18.

We claim that the germs of the bifurcation diagrams for these families are topologically nonequivalent for different values of j. In more detail, denote the saddle inside the inner cycle by I, the one between the cycles by B, and the saddles outside both cycles, by $E_1, \ldots, E_j$. Let ε, δ be the parameters of the family such that $\varepsilon = 0$ ($\delta = 0$) corresponds to the presence of the larger (respectively, smaller) semistable cycle. Then, in the domain $\varepsilon > 0$, $\delta > 0$, there are two sequences of pairwise disjoint "vertical" and "horizontal" curves. The first sequence tends to $\varepsilon = 0$ and corresponds to the sparkling saddle connections between B and E_j

that occur when the larger semistable cycle vanishes. The second sequence tends to $\delta = 0$ and corresponds to sparkling saddle connections between I and B.

The two sequences together form sort of a grid Σ, with the rectangle-like cells. It is similar to the one shown in Fig. 13.13, but a bit more complicated. The nodes of this grid correspond to simultaneous saddle connection, see Fig. 13.16.

A third family Λ of bifurcation curves occurs. It corresponds to saddle connections between I and E_j. The bifurcation curves of this family are similar to those shown in Fig. 13.14. But they are dashed because of the presence of the Cherry cells between two cycles, see Fig. 13.15, left.

The whole bifurcation diagram is similar to the one shown in Fig. 13.15, right, but more complicated. There is an infinite number of cells of the grid Σ that tend to zero and intersect exactly j arcs of the third family Λ. No cell intersects more than j arcs.

This number j is a topological invariant of the bifurcation diagram constructed. Thus a countable number of germs of pairwise topologically nonequivalent germs of bifurcation diagrams in generic two-parameter families on the sphere occurs.

This completes the sketch of the proof of Theorem 10. □

13.2.8 *Quasigeneric Families with a Continuum of Topologically Nonequivalent Bifurcation Diagrams*

Intuitively speaking, a quasigeneric family is a "corrupted generic family": one (and exactly one) genericity condition is violated.

Theorem 11. *There exists a class of quasigeneric two-parameter families whose bifurcation diagram has a numeric modulus of the topological classification. Consequently, for this class of local families there exists a continual set of pairwise topologically nonequivalent germs of bifurcation diagrams .*

Sketch of the Proof. Consider a two-parameter family with two separate semistable cycles and six saddles involved, see Fig. 13.19.

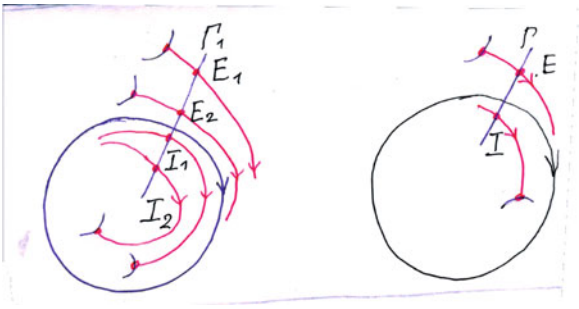


Fig. 13.19 Large support of a quasigeneric family whose bifurcation diagram has numeric modulus of topological classification

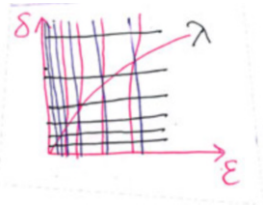


Fig. 13.20 Bifurcation diagram with a numeric modulus of topological classification

Let the saddles depend on the parameters ε, δ, as well as the "first" intersection points of the separatrixes winding to and from the first cycle. Denote these points by $E_1(\varepsilon, \delta)$ and $E_2(\varepsilon, \delta)$ for exterior saddles, $I_1(\varepsilon, \delta)$ and $I_2(\varepsilon, \delta)$ for interior saddles. Let $w_{\varepsilon,\delta}$ be the same as in Sect. 13.2.4, and S be the synchronization function (13.5). One of the genericity assumptions for the family is

$$S(0, 0) \notin \mathbb{Z}.$$

In this case, the bifurcation diagram in the family is like the diagram in Fig. 13.13.

Suppose now that the genericity assumption above fails; for instance,

$$S(0, 0) = 0.$$

In a quasigeneric family the curve λ:

$$S(\varepsilon, \delta) = 0$$

is transversal both to $\varepsilon = 0$ and $\delta = 0$ at zero. The bifurcation diagram for this family is a combination of two: the one shown in Fig. 13.11, and the other from Fig. 13.13. It is plotted in Fig. 13.20.

Roughly speaking, *the slope of the curve λ at zero is a topological invariant of the bifurcation diagram described above*.

Let us make a precise statement. Let $T(x, \varepsilon, \delta)$ be the time function corresponding to the Poincaré map of the first cycle, see (13.3), and $R(x, \varepsilon, \delta)$ be the similar function for the second cycle. Let us make a parameter change:

$$\Phi : (\mathbb{R}^+, 0) \times (\mathbb{R}^+, 0) \to (\mathbb{R}^+, \infty) \times (\mathbb{R}^+, \infty)$$

defined as follows:

$$(\varepsilon, \delta) \mapsto \tau(\varepsilon, \delta),\ \sigma(\varepsilon, \delta),$$

where

$$\tau(\varepsilon, \delta) = T(E_1(\varepsilon, \delta), \varepsilon, \delta) - T(I_1(\varepsilon, \delta),\ \varepsilon, \delta),$$
$$\sigma(\varepsilon, \delta) = R(E(\varepsilon, \delta), \varepsilon, \delta) - R(I(\varepsilon, \delta),\ \varepsilon, \delta),$$

The saddle connections between E_1 and I_1 that intersect Γ_1 $k+1$ times correspond to the line $\tau = k$. Similar connections between E and I correspond to $\sigma = k$. So the

bifurcation diagram of the family contains the grid:

$$\tau \in \mathbb{Z}^+ + k_0,\ \sigma \in \mathbb{Z}^+ + m_0$$

for some $k_0, m_0 \in \mathbb{Z}^+$.

The numbering of the lines of the grid is defined up to adding $O(1)$. So the curve $\Phi(\lambda)$ may be given by a function

$$\sigma = \varphi(\tau).$$

It is easy to prove that there exists a limit

$$\omega = \lim_{\tau \to \infty} \frac{\varphi(\tau)}{\tau}.$$

We claim that this limit is an invariant of the topological classification of the bifurcation diagrams of the class considered. Indeed, for any integral value $\tau = k > 0$, the integer part of $\varphi(k)$ is the number m of the horizontal line such that $\Phi(\lambda)$ intersects the segment

$$\{(\tau, \sigma) | \tau = k,\ \sigma \in [m, m+1]\}.$$

This number is topologically well-defined modulo an additional term $O(1)$ non depending on m. Then $\varphi + O(1)$ is well defined topologically. Hence, the limit ω is topologically well defined. In other words, ω is a topological invariant of the bifurcation diagram of the family. This completes the sketch of the proof of the theorem. □

13.3 Global Bifurcations in Generic Three-Parameter Families

As of now, this subject is almost untouched. An exclusion is the so-called ensemble "lips," a continual set of polycycles that occurs in generic three-parameter families.

13.3.1 Ensemble "Saddle Lips"

Consider a vector field v_0 with the following three degeneracies: v_0 has two saddle-nodes whose parabolic sectors are turned "face to face": a continuum of the phase curves that emerge one sector enter the other one; moreover, the separatrixes of the hyperbolic sectors of the saddle-nodes coincide, see Fig. 13.21.

The field v_0 has a continual family of polycycles: they all contain a mutual separatrix of the two saddle-nodes , the saddle-nodes included, and the phase curves that emerge one saddle-node and enter the other one, one curve for each polycycle.

This family of polycycles is bounded by the separatrixes of two saddles: E lying outside the polycycles described above, and I lying inside. The large bifurcation support for the unfolding of the ensemble "saddle lips" is the union of all the polycycles of the ensemble.

The bifurcation diagrams of this family may be unboundedly complicated. Namely, for a graph Γ of any generic monotonic function $[0, 1] \to [0, 1]$ there exists a vector field v_0 in the class described above, with the following property. The bifurcation diagram of a generic unfolding of v_0 is a surface with singularities. It contains a surface homeomorphic to a cone over the Legendre transformation of Γ, see Fig. 13.22.

This surface may have an arbitrary large number of self intersections. Thus, there exists an infinite number of bifurcation diagrams for such families; these diagrams are pairwise topologically nonequivalent [14].

Bifurcations in the ensemble "lips" without two saddles E and I is studied in [14]. Bifurcations in the same ensemble with the saddle E included and I deleted is studied in [20]. The global bifurcations in the ensemble described are not yet studied. In particular, it is unclear, whether this ensemble admits sparkling saddle connections, whatever it means.

In the English translation of Arnold et al. [1] there is a remark made by Arnold that follows the text quoted above:

> Recently A. Kotova and V. Stanzo found a counterexample to Conjecture 2. Little is now known: even for families of structurally stable and quasigeneric vector fields, Conjectures 3 and 4 (the only nontrivial in this case) are unproved.

In what follows, we will discuss these conjectures in full generality, not only for families of structurally stable and quasigeneric vector fields.

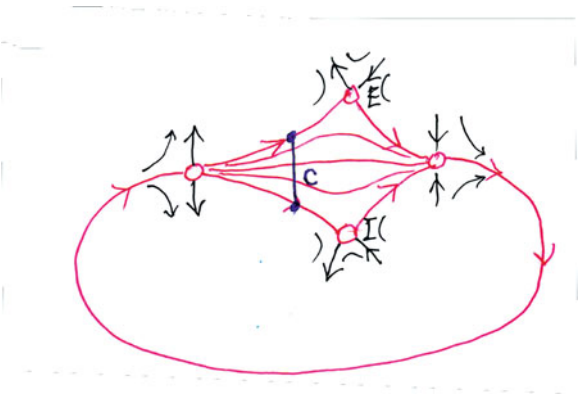


Fig. 13.21 Ensemble saddle—lips

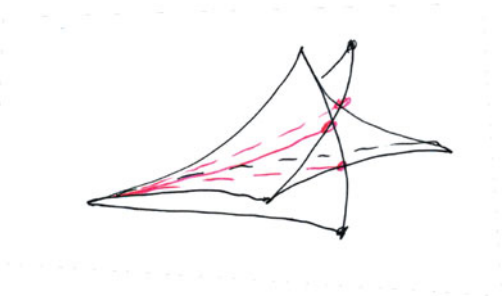


Fig. 13.22 A piece of a bifurcation diagram for the ensemble "saddle—lips"

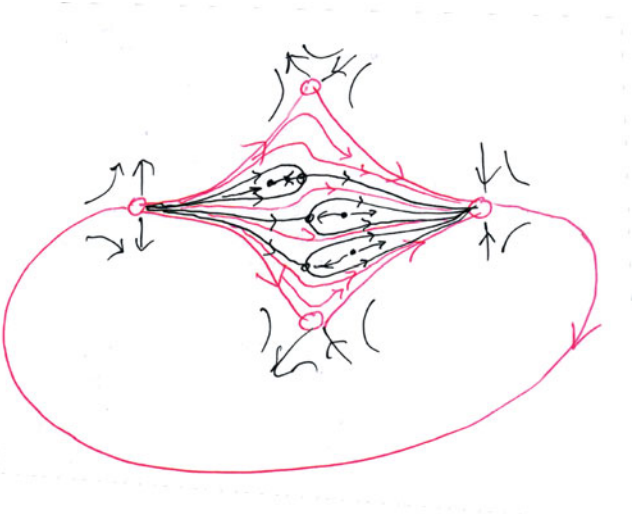


Fig. 13.23 An ensemble "shark"

13.3.2 Ensemble "Shark"

In three-parameter families a new kind of sparkling saddle connections may occur. Consider a polycycle with at least one saddle-node singular point on it. Suppose that a separatrix of some saddle enters the parabolic sector of this saddle-node. Then, after the saddle-node disappears, the separatrix may start to wind in a neighborhood of a polycycle that have vanished, and produce a variety of saddle connections with the other separatrixes.

As an illustration consider an ensemble "shark," see Fig. 13.23. The name comes from the figure that resembles the mouth full of teeth. When both saddle-nodes disappear, an enormous variety of saddle connections may occur. It is unclear whether or not the generic unfolding of such ensembles is structurally stable.

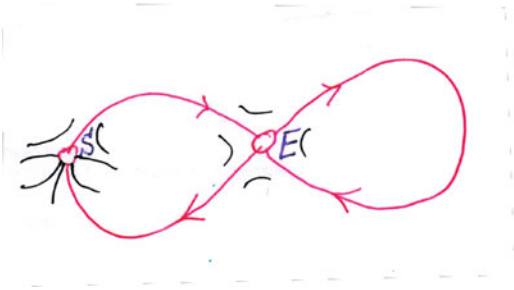


Fig. 13.24 A collection "apple-loop"

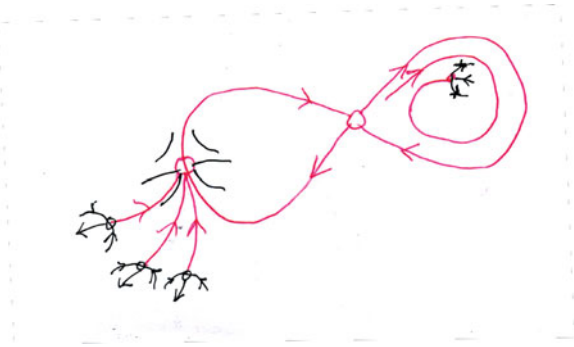


Fig. 13.25 Large bifurcation support with the "apple-loop" collection as a subset: vector field v_0

13.3.3 Extended Collection "Half-Apple and Loop"

A collection of polycycles named in the title may occur in a generic three-parameter family. It is shown in Fig. 13.24. There are three degeneracies for the field v_0 with such a collection:

- a separatrix loop of a hyperbolic saddle E;
- a saddle-node S;
- a connection between S and E: the separatrix of the hyperbolic sectors of the saddle-node coincides with the incoming separatrix of S.

Note that the outcoming separatrix of E may enter the parabolic sector of the saddle-node S without increasing the rate of degeneracy.

Separatrixes of hyperbolic saddles winding from the saddle loop inside it do not increase the rate of degeneracy, see Fig. 13.25.

The same holds true for the separatrixes that enter the saddle-node S from outside the polycycle . When the saddle-node disappears, and the connections are broken, a lot of sparkling saddle connections may occur, see Fig. 13.26.

Problem 3. *Are the generic unfoldings of the ensemble "shark" or a polycycle "halfapple and loop" with extra saddles structurally stable?*

The abundance of saddle connections that may occur makes the positive answer very plausible.

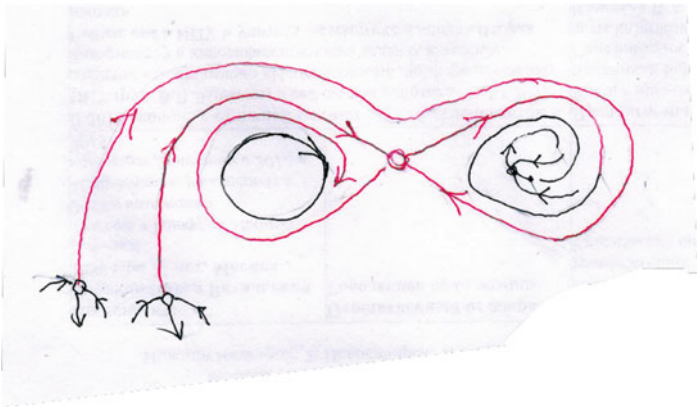


Fig. 13.26 A possible phase portrait in the unfolding of the field v_0 from the previous picture

13.3.4 Kotova Zoo Revisited

A polycycle that may occur in a generic k-parameter family is defined not only by its geometry. For instance, the first polycycle in the Kotova zoo for the two-parameter families is a separatrix loop with a zero saddle value, see Fig. 13.4.

Definition 15. A rigged polycycle is a polycycle with some additional restrictions on the jets of the corresponding vector field at the vertexes or at the edges of the polycycle.

We mention here some rigged polycycles from the Kotova zoo for three-parameter families.

– A separatrix loop with a zero saddle value and zero Melnikov integral:

$$I = \int_\gamma \mathrm{d}iv\, v\, dt,$$

where γ is the separatrix loop, t its time parametrization.

– An eight shaped figure, see Fig. 13.6, with a zero saddle value.

Definition 16. A large support that contains rigged polycycles is called *rigged large supports*

Definition 17. Two rigged large supports are equivalent if they are isotopic, and the isotopy respects the rigging relations: these relations are the same for the phase curves of two supports that are mapped to each other by the isotopy.

These definitions will be used in the next section.

13.4 Global Bifurcations with Many Parameters

No results are known to the author in general theory of global planar bifurcations with the number of parameters greater than three. Here we state some problems only.

13.4.1 Supports and Their Basins

Definition 18. A basin of an invariant set A of a planar vector field is the set of all points whose α- or ω-limit sets belong to A.

Example 12. All the phase curves that wind to or from a semistable cycle belong to the basin of this cycle. Their union equals this basin.

Problem 4. *Is it true that the large bifurcation support of a local finite-parameter family belongs to the basin of the small bifurcation support?*

Conjecture 3. The answer is "yes" for generic one-parameter families.

By definition, two local families are weakly equivalent when they are equivalent in some neighborhoods of their large bifurcation supports. The following problem is aimed to increase the domain where two families are equivalent.

Problem 5. *Is it correct that two weakly equivalent local families are in fact weakly equivalent in the basins of their large bifurcation supports?*

Again, the answer seems to be affirmative for the generic one-parameter families.

13.4.2 Bifurcational Stability

Recall our main result about generic one-parameter families.

Two local one-parameter families are topologically equivalent iff the large supports of the corresponding bifurcations are isotopic and do not contain semistable cycles; moreover, vector fields corresponding to zero parameter value are orbitally topologically equivalent, see Theorem 7.

This property gives rise to the following definition:

Definition 19. A class of local families is (strongly) *bifurcationally stable* provided that the following holds. Two local families of this class are topologically equivalent iff their large supports of the corresponding bifurcations are isotopic, and vector fields corresponding to zero parameter values are orbitally topologically equivalent. The latter assumption is required in two definitions to follow. This class is *bifurcationally stable* if the equivalence above follows from the isotopy of the rigged large supports. This class is *weakly bifurcationally stable* if the isotopy class of the

rigged large support corresponds to a finite number of topological types of local families.

Problem 6. *Describe (strongly and weakly) bifurcationally stable classes of local families.*

We expect that the "majority" of local families are not bifurcationally stable in any sense. The more interesting are the classes that are bifurcationally stable.

Conjecture 4. The following local families are bifurcationally stable:
k-parameter families with a rigged separatrix loop that may be met in generic k-parameter but not in $(k-1)$-parameter families;
k-parameter families with a homoclinic curve of a saddle-node of the multiplicity k.

Problem 7. *What may be said about the bifurcational stability of local families whose small supports are rigged polycycles met in generic two-parameter families?*

In contrast to the global bifurcation theory of k-parameter families, the semilocal one, related to bifurcations of polycycles, is more elaborated.

13.5 Bifurcations of Polycycles

This is a rich theory, and we discuss here only a few results from it. We start with the polynomial case.

13.5.1 Bifurcational Approach to the Hilbert's 16th Problem

This approach is due to Roussarie. It is related to the following form of Hilbert's 16th problem:

Problem 8. *Prove that for any n there exists $H(n)$, a Hilbert number, such that a planar polynomial vector field of degree no greater than n can have no more than $H(n)$ limit cycles.*

For what follows, we recall some well-known definitions.

An oriented polycycle is a finite union of cyclically enumerated singular points (the vertexes) and phase curves that connect the vertexes (the edges) with the following properties:

- the vertexes with the different numbers may coincide; the edges may not;
- the edge number j connects the vertexes O_j and O_{j+1};
- the time orientation of the edge number j is from O_j to O_{j+1};
- the first vertex (edge) follows the last one (cyclicity of the enumeration).

Semilocal bifurcations are considered in the neighborhoods of the polycycles. Two semilocal bifurcations of polycycles are equivalent if there exists two neighborhoods of the polycycle where two local families are weakly topologically equivalent.

Definition 20. A limit cycle is *generated by an unfolding* of a polycycle if there exists a family of limit cycles depending on the parameter of the unfolding such that the limit cycles of the family tend to the polycycle in sense of the Hausdorff distance as the parameter tends to the critical value.

Definition 21. The *cyclicity* of a polycycle in a family is the maximal number of limit cycles that may be generated by this polycycle in this family.

Conjecture 5 (Roussarie, [19]). Any polycycle that occurs in a family of planar polynomial vector fields of degree no greater than given n has a finite cyclicity.

Theorem 12 ([19]). *The above conjecture implies the existence of Hilbert number $H(n)$ for any n.*

13.5.2 The Dumortier–Roussarie–Rousseau Program

In [2] the authors listed all the rigged polycycles that may occur in the family of the quadratic vector fields. Their number appeared to be 121. It is sufficient to prove the finite cyclicity of any of them in the family of quadratic vector fields , in order to prove the existence of $H(2)$. Up to now more than 80 polycycles are studied and their finite cyclicity is proved. This partial success is a strong indication that $H(2)$ really exists.

13.5.3 The Arnold's Program for the Polycycles

It seems that the Arnold's program is perfectly adjusted to the study of bifurcations of polycycles . But the problem seems to be very difficult. Indeed, the Conjecture 2: *Any bifurcation diagram is (locally) homeomorphic to one of a finite number (depending only upon l) of generic examples* is closely related to the following:

Conjecture 6 (Hilbert–Arnold Conjecture, [6]). For any k, a polycycle met in a typical k-parameter family, has but a finite cyclicity.

13.5.4 Present Status of the Hilbert–Arnold Conjecture

The conjecture is proved for the so-called elementary polycycles: those that have vertexes as singular points whose eigenvalues are not simultaneously zero.

Theorem 13 ([10]). *Elementary polycycles met in a typical k-parameter families have but a finite cyclicity.*

Later on this cyclicity was estimated by V. Kaloshin, [12, 13]

Theorem 14 ([11]). *The cyclicity of an elementary polycycle met in a typical k-parameter family is no greater than* 2^{25k^2}.

Kaloshin suggested to estimate the cyclicity of polycycles with a fixed number n of vertices that may be met in generic k-parameter families. Kaleda and Schurov obtained this estimate.

Theorem 15 ([11]). *The cyclicity of an elementary polycycle with n vertexes met in a typical k-parameter family is no greater than* $C(n)k^{3n}$.

13.5.5 Finiteness Theorem for Generic k-Parameter Families

There's no doubt that the following theorem is true:

Theorem 16. *A vector field met in typical k-parameter family has but a finite number of limit cycles .*

Yet the theorem remains unproved.

13.5.6 Back to the Hilbert–Arnold Problem

One of the equivalent forms of the finiteness theorem for limit cycles is the following *nonaccumulation theorem*:

Theorem 17 ([3, 5]). *Limit cycles of an analytic vector field cannot accumulate to a polycycle of this field.*

Classical Seidenberg–Lefshez–Bendixson–Dumortier theorem reduces this statement to the case when the polycycle in the theorem is elementary.

The question arises: may the general Hilbert–Arnold problem be reduced to Theorem 13 by sort of desingularization in the families?

13.5.7 Desingularization in the Families

The question above was investigated by S. Trifonov.

Definition 22. Say that a family of vector fields is quasielementary if any field of the family either has but a finite number of singular points and they are all

elementary (such fields are called elementary) or has a whole curve of singular points, and becomes elementary after a division of its components by a common non-invertible analytic factor.

Theorem 18 ([21]). *Any analytic finite-parameter family of vector fields, after a special desingularization process, may be transformed to a quasielementary family of vector fields.*

13.5.8 Trifonov Phenomenon

Note that a quasielementary family may not be equivalent to a family of elementary vector fields. Indeed, a vector field with a curve of singular points may correspond to an isolated parameter value. This particular vector field may be transformed into an elementary one by the division by a non-invertible function. But the nearby vector fields have no common factor of their components. Thus the whole quasielementary family cannot be transformed into an elementary one. This effect called the *Trifonov phenomenon* prevents the reduction of the general Hilbert–Arnold problem to Theorem 13.

13.5.9 Back to the Arnold's Program

Definition 23. A nest of a planar vector field with a finite number of singular points in an open subset Z in the phase plane such that
- Z is homeomorphic to an annulus;
- the boundary curves of the annulus Z are limit cycles;
- Z contains no singular points of the field.

A nest is said to be *maximal* if it is not a proper subset of another nest.

Definition 24. Consider vector fields v_1 and v_2. Let Z_j be the union of maximal nests of $v_j, j = 1, 2$. We say that v_1 and v_2 are equivalent *modulo limit cycles* if the restriction of v_1 to $\overline{\mathbb{R} \setminus Z_1}$ is orbitally topologically equivalent to the restriction of v_2 to $\overline{\mathbb{R}^2 \setminus Z_2}$ (the bar denotes the closure of the set).

Definition 25. Two families $\{v_\varepsilon\}$ and $\{w_\varepsilon\}$ of vector fields in the total spaces $B \times M$ and $B' \times M'$ are weakly equivalent modulo limit cycles , if there exists a map

$$H\colon B \times M \to B' \times M, \quad (\varepsilon, x) \mapsto (\varphi(\varepsilon), h_\varepsilon(x)),$$

where h_ε is a homeomorphism $M \to M'$ not necessary continuous in ε, such that h_ε is a topological equivalence of v_ε and $w_\varphi(\varepsilon)$, modulo limit cycles .

The following problem is inspired by Arnold's conjectures from Sect. 13.1.2:

Problem 9. *Is it correct that for any k there is but a finite number of pairwise topologically nonequivalent generic k-parameter unfoldings of polycycles in their neighborhoods modulo limit cycles?*

Addendum. In a recent preprint: Yu. Ilyashenko, Yu. Kudryashov , I. Schurov, An open set of structurally unstable families of vector fields in the two-sphere, arXiv:1506.06797 [math.DS], an open set of three parameter families named in the title was constructed.

Acknowledgements The article was prepared within the framework of the Academic Fund Program at the National Research University Higher School of Economic (HSE) in (2016–17) (grant # 16-05-0066) and supported within the framework of a subsidy granted to the HSE by the Government of Russian Federation for the implementation of the Global Competitiveness program.

References

1. Arnold, V.I, Afrajmovich, V.S., Ilyashenko, Y.S., Shilnikov, L.P.: Bifurcation Theory and Catastrophe Theory. Translated from the 1986 Russian original by N. D. Kazarinoff, Reprint of the 1994 English edition from the series Encyclopedia of Mathematical Sciences [Dynamical systems. V, Encyclopedia Mathematical Science, vol. 5, viii+271 pp. Springer, Berlin (1994); Springer, Berlin (1999)
2. Dumortier, F., Roussarie, R., Rousseau, C.: Hilbert's 16th problem for quadratic vector fields. J. Differ. Equ. **110**(1), 86–133 (1994)
3. Écalle, J.: Introduction aux fonctions analysables et preuve constructive de la conjecture de Dulac (French). Hermann, Paris (1992)
4. Fedorov, R.M.: Upper bounds for the number of orbital topological types of polynomial vector fields on the plane "modulo limit cycles" (Russian). Uspekhi Mat. Nauk **59**(3)(357), 183–184 (2004). Translation in Russian Math. Surveys **59**(3), 569–570 (2004)
5. Ilyashenko, Y.S.: Finiteness Theorems for Limit Cycles. Translated from the Russian by H. H. McFaden. Translations of Mathematical Monographs, vol. 94, x+288 pp. American Mathematical Society, Providence, RI (1991)
6. Ilyashenko, Y.S.: Local dynamics and nonlocal bifurcations. In: Bifurcations and Periodic Orbits of Vector Fields (Montreal, PQ, 1992). NATO Advanced Science Institute Series C: Mathematical Physical Sciences, vol. 408, pp. 279–319. Kluwer Academic Publisher, Dordrecht (1993)
7. Ilyashenko, Y., Weigu, L.: Nonlocal Bifurcations. Mathematical Surveys and Monographs, vol. 66, xiv+286 pp. American Mathematical Society, Providence, RI (1999)
8. Ilyashenko, Y., Yakovenko, S.: Smooth normal forms for local families of diffeomorphisms and vector fields. Russ. Math. Surv. **46**(1), 3–39 (1991)
9. Ilyashenko, Y., Yakovenko, S.: Nonlinear Stokes Phenomena in smooth classification problems. In: Nonlinear Stokes Phenomena. Advances in Soviet Mathematics, vol. 14, pp. 235–287. American Mathematical Society, Providence, RI (1993)
10. Ilyashenko, Y., Yakovenko, S.: Finite cyclicity of elementary polycycles in generic families. In: Concerning the Hilbert 16th Problem. American Mathematical Society Translation Series 2, vol. 165, pp. 21–95. American Mathematical Society, Providence, RI (1995)
11. Kaleda, P.I., Schurov, I.V.: Cyclicity of elementary polycycles with a fixed number of singular points in generic *k*-parametric families (Russian). Algebra i Analiz **22**(4), 57–75 (2010); Translation in St. Petersburg Math. J. **22**(4), 557–571 (2011)

12. Kaloshin, V.: The existential Hilbert 16-th problem and an estimate for cyclicity of elementary polycycles. Invent. Math. **151**(3), 451–512 (2003)

13. Kaloshin, V.: Around the Hilbert-Arnold problem. In: On Finiteness in Differential Equations and Diophantine Geometry. CRM Monograph Series, vol. 24, pp. 111–162. American Mathematical Society, Providence, RI (2005)

14. Kotova, A., Stanzo, V.: On few-parameter generic families of vector fields on the two-dimensional sphere. In: Concerning the Hilbert 16th Problem, pp. 155–201. American Mathematical Society Translation Series 2, vol. 165. American Mathematical Society, Providence, RI (1995)

15. Leontovich, E.: On the generation of limit cycles from separatrices (Russian). Doklady Akad. Nauk SSSR (N.S.) **78**, 641–644 (1951)

16. Malta, I.P., Palis, J.: Families of vector fields with finite modulus of stability. In: Dynamical systems and turbulence, Lecture Notes in Mathematics, vol. 898, pp. 212–229 (1980)

17. Roussarie, R.: On the number of limit cycles which appear by perturbation of separatrix loop of planar vector fields. Bol. Soc. Brasil Mat. **17**(2), 67–101 (1986)

18. Roussarie, R.: Weak and continuous equivalences for families on line diffeomorphisms. In: Dynamical Systems and Bifurcation Theory (Rio de Janeiro, 1985). Pitman Research Notes in Mathematics Series, vol. 160, pp. 377–385. Longman Sci. Tech, Harlow (1987)

19. Roussarie, R.: A note on finite cyclicity property and Hilbert's 16th problem. In: Dynamical Systems Valparaiso 1986. Lecture Notes in Mathematics, vol. 1331, pp. 161–168. Springer, Berlin (1988)

20. Stantzo, V.: Bifurcations of the polycycle "Saddle Lip". Tr. Mat. Inst. Steklova **213** (1997), Differ. Uravn. s Veshchestv. i Kompleks. Vrem., 152–212. Translation in Proc. Steklov Inst. Math. **213**(2) , 141–199 (1996)

21. Trifonov, S.: Desingularization in families of analytic differential equations. In: Concerning the Hilbert 16th Problem. American Mathematical Society Translation Series 2, vol. 165, pp. 97–129. American Mathematical Society, Providence, RI (1995)

22. Trifonov, S.I.: Cyclicity of elementary polycycles of generic smooth vector fields (Russian). Tr. Mat. Inst. Steklova 213 (1997). Differ. Uravn. s Veshchestv. i Kompleks. Vrem, 152–212; Translation in Proc. Steklov Inst. Math. **213**(2), 141–199 (1996)

Chapter 14
Slow-Fast Dynamics and Its Application to a Biological Model

Chengzhi Li

Abstract In this article we introduce some basic concepts about slow-fast dynamics and its application to a biological model, that is a predator–prey system with response functions of Holling type. The relevant studies were collaborated with Kening Lu in Li and Lu (J Differ Equ 257:4437–4469, 2014) and with Huiping Zhu in Li and Zhu (J Differ Equ 254:879–910, 2013). Another application to a medical model, especially a SIS epidemic model with nonlinear incidence, was published in Li et al. (J Math Anal Appl 420:987–1004, 2014), collaborated with Jiaquan Li, Zhien Ma, and Huiping Zhu. The studies are based on singular perturbation theory developed by F. Dumortier, R. Roussarie, and P. De Maesschalck, see, for example, Dumortier and Roussarie (Mem Am Math Soc 121(577):1–100, 1996), Dumortier and Roussarie (J Differ Equ 174:1–29, 2001), Dumortier and Roussarie (Discrete Continuous Dyn Syst Ser S 2:723–781, 2009), De Maesschalck and Dumortier (Trans Am Math Soc 358(5):2291–2334, 2006), De Maesschalck and Dumortier (Proc R Soc Edinb A 138(2):265–299, 2008), De Maesschalck et al. (Indag Math 22:165–206, 2011), and De Maesschalck et al. (C R Math Acad Sci Paris 352(4):317–320, 2014).

Keywords Singular perturbation • Slow-fast cycle and its cyclicity • Slow divergence integral • Predator–prey system

14.1 Singular Perturbation

Slow-fast dynamics appears in many problems and models with different time-scales or different space-scales. Usually, the problems are expressed by differential equations with a critical set of singularities, forming a curve, a surface, or a manifold, and the most singularities on the critical set can be removed by perturbation with a small parameter and some complex phenomenon happens. This is called the *singular perturbation*.

C. Li (✉)
School of Mathematical Sciences, Peking University, Beijing 100871, China
e-mail: licz@math.pku.edu.cn

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_14

We consider the planar system

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= f(x, y, \varepsilon), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= \varepsilon g(x, y, \varepsilon), \end{aligned} \tag{14.1}$$

where $(x, y) \in \mathbb{R}^2$, $0 < \varepsilon \ll 1$, f and g are C^k-functions with $k \geq 3$. If we divide the second equation by ε, and let $\tau = \varepsilon t$, then the time t in the first equation and the time τ in the second equation have very different scales, since ε is small. Hence, t is called the *fast time* while τ is called the *slow time*. We can expect that the movement is very fast along x-direction and quite slow along y-direction. Using time τ Eq. (14.1) can be changed to

$$\begin{aligned} \varepsilon \frac{\mathrm{d}x}{\mathrm{d}\tau} &= f(x, y, \varepsilon), \\ \frac{\mathrm{d}y}{\mathrm{d}\tau} &= g(x, y, \varepsilon). \end{aligned} \tag{14.2}$$

If we let $u = \varepsilon x$ then system (14.2) becomes

$$\begin{aligned} \frac{\mathrm{d}u}{\mathrm{d}\tau} &= f(\tfrac{u}{\varepsilon}, y, \varepsilon), \\ \frac{\mathrm{d}y}{\mathrm{d}\tau} &= g(\tfrac{u}{\varepsilon}, y, \varepsilon). \end{aligned} \tag{14.3}$$

The scales of the space variables on the right-hand side are very different.

To study the behavior of the orbits for $0 < \varepsilon \ll 1$, we first consider the limiting case $\varepsilon = 0$. Then, (14.1) becomes the *fast subsystem* (the so-called *layer equation*)

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= f(x, y, 0), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= 0, \end{aligned} \tag{14.4}$$

and (14.2) becomes the *low subsystem* (the so-called *reduced equation*)

$$\begin{aligned} 0 &= f(x, y, 0), \\ \frac{\mathrm{d}y}{\mathrm{d}\tau} &= g(x, y, 0). \end{aligned} \tag{14.5}$$

The set of singularities of system (14.4) is formed by

$$S = \{(x, y) \,|\, f(x, y, 0) = 0\},$$

which is the phase space of (14.5). Usually S is a curve or a manifold in $\mathbb{R}^2$, and is called a *slow curve* or *slow manifold*. We suppose that the critical curve S can be expressed by $y = \varphi(x)$ and has one or two *non-degenerate* fold point(s), that is, $\varphi' = 0$ and $\varphi'' \neq 0$ at the point(s). In this case S is called U-shaped or S-shaped, shown in Fig. 14.1a and b, respectively, where the fast orbits of (14.4) are sketched by solid lines and the slow curve is shown by dotted line.

By Fenichel theory [17], along the compact part of S where (14.4) is *normally hyperbolic*, it is perturbed to a nearby invariant manifold S^ε of (14.1) for $0 < \varepsilon \ll 1$. However, if S has a non-normally hyperbolic point (this happens at a *fold point* of S) then the geometric singular perturbation theory does not apply. In a pioneering

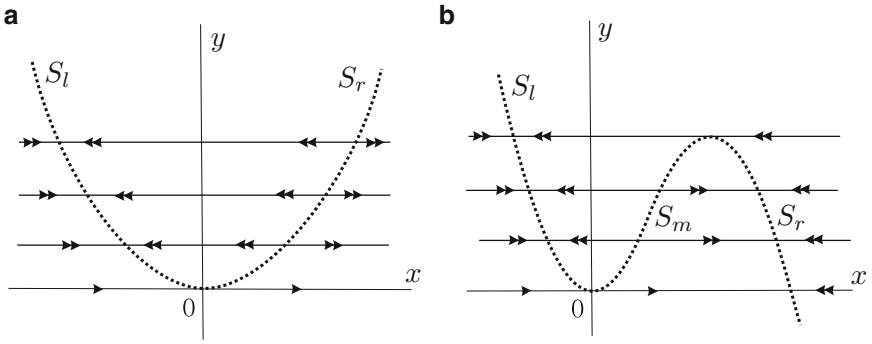


Fig. 14.1 The slow curve S. (**a**) U-shaped; (**b**) S-shaped

work [14], Dumortier and Roussarie first used the blow-up technique at fold points, combined with the center manifold theory globally, successfully studied the singular perturbations of van der Pol equation. They (also with P. De Maesschalck) generalized their method and results to wide classes of systems, see, for example, [10–13, 15, 16]. Krupa and Szmolyan [21, 22] used the blow-up method, provided in [14], combined with Fenichel theory, also studied the singular perturbations of planar systems, and give a standard form near a non-degenerate canard point. Let us briefly introduce a part of their results, which will be used below.

We suppose that a fold point of S is located at $p = (0, \varphi(0)) = (0, 0)$, and $\varphi'(0) = 0$, $\varphi''(0) > 0$. To see the slow movement, restricted to the slow curve S, from the second equation of (14.5) we have

$$\varphi'(x)\frac{\mathrm{d}x}{\mathrm{d}\tau} = g(x, \varphi(x), 0). \tag{14.6}$$

In fact, the point $(x, y, \varepsilon) = (x, \varphi(x), 0)$, $x \neq 0$, is a normally hyperbolic singular point of the vector field (14.1)+$0\frac{\partial}{\partial\varepsilon}$. Hence, near the critical curve, outside a small neighbourhood of the fold point $(0, 0)$, center manifolds are given by

$$y = \varphi(x) + \varepsilon h(x) + O(\varepsilon^2),$$

where the function h can be easily found. Now the equations in system (14.1) imply that the dynamics inside such center manifolds can be given by

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \varepsilon\left(\frac{g(x, \varphi(x), 0)}{\varphi'(x)} + O(\varepsilon)\right).$$

Dividing this equation by ε, using $\tau = \varepsilon t$, and letting ε go to 0, we can find the slow dynamics (14.6) for $x \neq 0$.

If $g(0, 0, 0) \neq 0$, for example, $g(0, 0, 0) < 0$, since $\varphi'(x) = x\varphi''(0) + O(x^2)$, the movement on slow curve for $\varepsilon = 0$ is shown in Fig. 14.2a, and the perturbation of

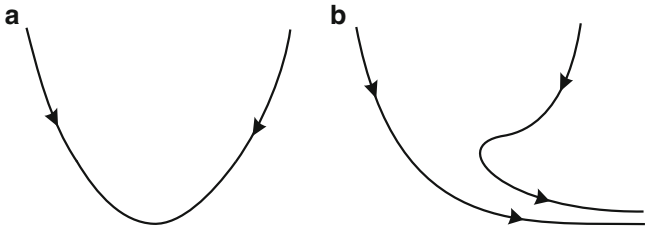


Fig. 14.2 Behavior of the slow curve near a jump point for: (**a**) $\varepsilon = 0$; (**b**) $0 < \varepsilon \ll 1$

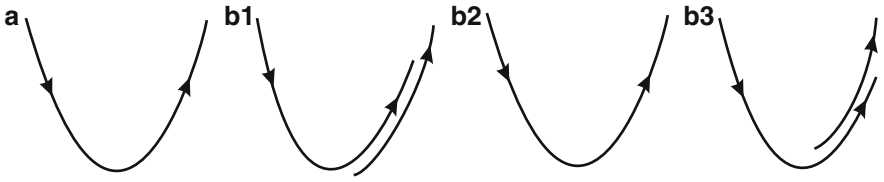


Fig. 14.3 Behavior of the slow curve near a canard point for: (**a**) $\varepsilon = 0$; (**b**) $0 < \varepsilon \ll 1$

the slow curve near the fold point for $0 < \varepsilon \ll 1$ is shown in Fig. 14.2b, see [14] or [21, 22]. In this case the fold point is called a *jump point*.

If $g(0, 0, 0) = 0$, for example, $g(x, \varphi(x), 0) = x + O(x^2)$, then along the slow curve we have

$$\frac{dx}{d\tau} = \frac{1 + O(x)}{\varphi''(0) + O(x)},$$

the movement on slow curve for $\varepsilon = 0$ is shown in Fig. 14.3a, and the perturbation of the slow curve near the fold point for $0 < \varepsilon \ll 1$ is shown in Fig. 14.3b1–b3, see [14] or [21, 22]. In this case, the fold point is called a *canard point*, and it is necessary to introduce one more parameter $\lambda = \lambda(\varepsilon) \to 0$ as $\varepsilon \to 0$, the middle case (b2) of Fig. 14.3 corresponds to some function $\lambda = \lambda_c(\sqrt{\varepsilon})$, see Theorem 3.2 of [22].

Now we consider a *limit periodic set* or *slow-fast cycle*, consisting of compact pieces of slow curve and some compact parts of fast orbits for $\varepsilon = 0$, that forms a loop Γ, the orientation of the flow on Γ is counterclockwise (or clockwise if we change some signs in above conditions), uniformly for all parts. We want to study for $0 < \varepsilon \ll 1$ is there a periodic orbit γ_ε of the system (14.1) such that $\gamma_\varepsilon \to \Gamma$ (in Hausdorff distance) as $\varepsilon \to 0$? and how many such γ_ε for a given Γ ? Roughly speaking, the cyclicity of Γ is the maximal number of such γ_ε (see Definition 1.1).

It is clear from above analysis that if the slow curve is U-shaped and the unique fold point is a jump point, then there is no such a slow-fast cycle in a neighborhood of the fold point. In S-shaped case, if the two fold points are both jump points, the only possible slow-fast cycle Γ is shown in Fig. 14.4a, and after perturbation the

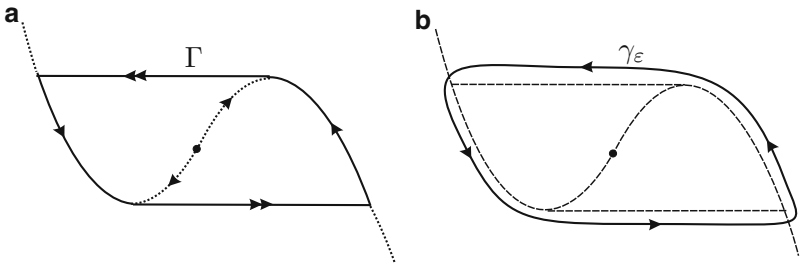


Fig. 14.4 The relaxation oscillation: (**a**) $\varepsilon = 0$; (**b**) $0 < \varepsilon \ll 1$

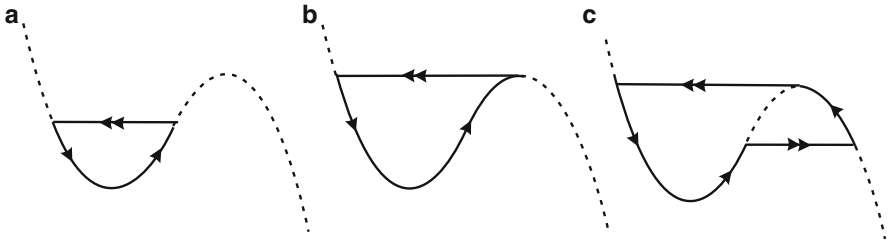


Fig. 14.5 The slow-fast cycle. (**a**) without a head; (**b**) in transitory case; (**c**) with a head

only possible periodic orbit γ_ε is shown in Fig. 14.4b, and it is called *relaxation oscillation*. Note that in this case, a singularity of the slow movement appears on the middle part of the slow curve, denoted in Fig. 14.4 by a black point.

Appearance of a canard point is the most interesting case. We consider the *S*-shaped case with a canard point (at the origin) and with a jump point. From the cases (b1), (b2), and (b3) of Fig. 14.3, we may regard the slow-fast cycles as Fig. 14.5a, b, and c, respectively. Of course, we suppose that there are no any singularities of the slow movement on Γ (the canard point becomes a singularity of system (14.1) after perturbation, surrounded by a possible limit cycles γ_ε). The authors of [11] discussed the case if a singularity appears at a "corner" of Γ.

If the slow arcs, contained in a slow-fast cycle Γ, are all normally attracting or all normally repelling, like in Fig. 14.4a, then Γ is called a *common slow-fast cycle*. Otherwise, like anyone in Fig. 14.5, Γ is called a *canard slow-fast cycle*, see Definition 5 of [12]. The cases in Fig. 14.5a and c are called *canard slow-fast cycle without a head* and *canard slow-fast cycle with a head*, respectively, and Fig. 14.5b is the *critical case*, it is called a *transitory canard cycle*, see [13].

For simplicity of notations we rewrite systems (14.1) as

$$X_{\varepsilon,\mu}: \quad \begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = f(x, y, \mu, \varepsilon), \\ \dfrac{\mathrm{d}y}{\mathrm{d}t} = \varepsilon g(x, y, \mu, \varepsilon), \end{cases} \tag{14.7}$$

where $\varepsilon \geq 0$ is a small parameter, $\mu = (\lambda, \bar{\mu})$ is a multi-dimensional parameter in a compact subset of $\mathbb{R}^1 \times \mathbb{R}^p$, $\mu_0 = (0, \bar{\mu}_0)$, f and g are C^k functions with $k \geq 3$. Besides, for $\varepsilon = 0$ the system has a U-shaped or S-shaped slow curve $\{(x, y) | f(x, y, \mu_0, 0) = 0\}$ with an equation $y = \varphi(x)$, and the point $(0, \varphi(0)) = (0, 0)$ is a canard point. In S-shaped case, one more fold point at the maximum $(x_2, \varphi(x_2))$ is a jump point ($x_2 > 0$ and $\varphi(x_2) > 0$). In U-shaped case $S = S_l \cup \{(0, 0)\} \cup S_r$ and in S-shaped case $S = S_l \cup \{(0, 0)\} \cup S_m \cup \{(x_2, \varphi(x_2))\} \cup S_r$, see Fig. 14.1a and b.

To keep the normally hyperbolic property outside the fold points we assume that for $(\mu, \varepsilon) = (\mu_0, 0)$

(**A1**) In U-shaped case $\frac{\partial f}{\partial x} < 0$ on S_l and $\frac{\partial f}{\partial x} > 0$ on S_r; In S-shaped case $\frac{\partial f}{\partial x} < 0$ on $S_l \cup S_r$ and $\frac{\partial f}{\partial x} > 0$ on S_m.

To keep the generic property of the canard point $(0, 0)$ and the jump point $(x_2, \varphi(x_2))$, we assume

(**A2**) $\frac{\partial^2 f}{\partial x^2} \neq 0$ and $\frac{\partial f}{\partial y} \neq 0$ at both $(0, 0, \mu_0, 0)$ and $(x_2, \varphi(x_2), \mu_0, 0)$; $g(0, 0, \mu_0, 0) = 0$, $\frac{\partial g}{\partial x}(0, 0, \mu_0, 0) \neq 0$ and $\frac{\partial g}{\partial \lambda}(0, 0, \mu_0, 0) \neq 0$; $g(x_2, \varphi(x_2), \mu_0, 0) \neq 0$.

Note that paper [10] discussed some non-generic cases. As we mentioned before, to keep the fold points to be non-degenerate, we assume

(**A3**) $\varphi(0) = \varphi'(0) = 0, \varphi''(0) > 0$; $\varphi(x_2) = \varphi'(x_2) = 0, \varphi''(x_2) < 0$.

Near the non-degenerate canard point $(x, y) = (0, 0)$, $X_{\varepsilon,\mu}$ can be transformed to the form (see Sect. 3.2 of [22]):

$$\begin{aligned} \dot{x} &= -yh_1 + x^2 h_2 + \varepsilon h_3, \\ \dot{y} &= \varepsilon[xh_4 - \lambda h_5 + yh_6], \end{aligned} \tag{14.8}$$

where $h_j = h_j(x, y, \mu, \varepsilon) = 1 + O(x, y, \mu, \varepsilon)$, for $j = 1, 2, 4, 5$, $h_3 = h_3(x, y, \mu, \varepsilon) = O(x, y, \mu, \varepsilon)$, and $h_6 = h_6(x, y, \mu, \varepsilon)$, and the above-mentioned critical function $\lambda_c(\sqrt{\varepsilon})$ has the expansion

$$\lambda_c(\sqrt{\varepsilon}) = \lambda^* \varepsilon + O(\varepsilon^{3/2}), \tag{14.9}$$

which is C^k-smooth in $\sqrt{\varepsilon}$ and where

$$\lambda^* = \left[\frac{1}{8}\frac{\partial(h_1 - 3h_2 - 4h_3 + 2h_4)}{\partial x} - \frac{h_6}{4}\right]\Bigg|_{(x,y,\lambda,\varepsilon)=(0,0,0,0)}. \tag{14.10}$$

Now we consider the two types of slow-fast cycles, shown in Fig. 14.6.

Definition 1.1. For fixed $\mu_0 = (0, \bar{\mu}_0)$ and $s > 0$, if there are $\sigma > 0$ and $\varepsilon_0 > 0$, such that for each $\varepsilon \in (0, \varepsilon_0)$, the system (14.7) with $\mu = (\lambda, \bar{\mu})$ has a limit cycle γ_ε^μ in the σ-neighborhood of the slow-fast cycle $\Gamma(s)$, corresponding to $(\varepsilon, \mu) =$

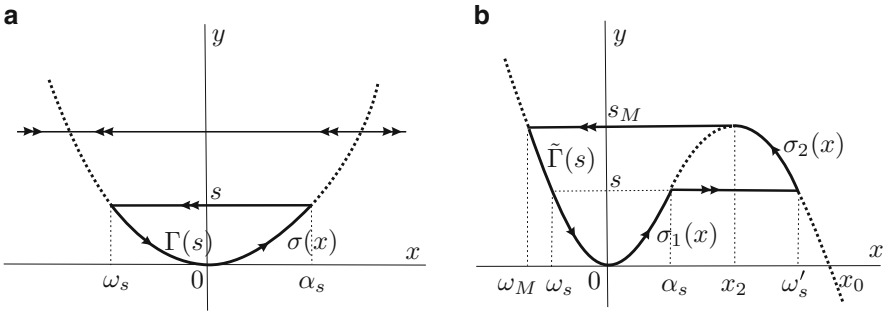


Fig. 14.6 The limit period set (slow-fast cycle) (**a**) $\Gamma(s)$; (**b**) $\tilde{\Gamma}(s)$

$(0, \mu_0)$, and $\gamma_\varepsilon^\mu \to \Gamma(s)$ (in Hausdorff distance) as $\varepsilon \to 0$, then γ_ε^μ is called a *canard cycle*, bifurcating from $\Gamma(s)$. The maximal number of such canard cycles, taking into account of their multiplicities, is called the cyclicity of $\Gamma(s)$ for system (14.7) at $(\varepsilon, \mu) = (0, \mu_0)$ and is denoted by $Cycl(X_{\varepsilon,\mu}, \Gamma(s), (0, \mu_0))$.

Remark 1.1. The above definition is about the cyclicity of the slow-fast cycle $\Gamma(s)$ without a head. If we change $\Gamma(s)$ to $\tilde{\Gamma}(s)$, then we obtain the definition of cyclicity for the slow-fast cycle with a head.

Remark 1.2. It was proved in Theorems 3.3 and 3.5 of [22] that for the existence of canard cycles if

$$A = \left[\frac{\partial(-h_1 + 3h_2 - 2h_4)}{\partial x} - 2h_6\right]\Big|_{(x,y,\lambda,\varepsilon)=(0,0,0,0)} \neq 0, \tag{14.11}$$

then the parameter $\lambda = \lambda(s, \sqrt{\varepsilon})$ is C^k-smooth in $(s, \sqrt{\varepsilon})$ and satisfies

$$|\lambda(s, \sqrt{\varepsilon}) - \lambda_c(\sqrt{\varepsilon})| \leq e^{-K/\varepsilon}, \tag{14.12}$$

for a constant $K > 0$, where $\lambda_c(\sqrt{\varepsilon})$ is given in (14.9).

An important problem is how to determine the cyclicity of a given slow-fast cycle $\Gamma(s)$ or $\tilde{\Gamma}(s)$. The following *slow divergence integral* is a crucial tool for this purpose (see, for example, [11, 14, 15, 22]).

$$\text{For } \Gamma(s): \ I(s, \mu_0) = \int_{\omega_s}^{\alpha_s} \frac{\partial f}{\partial x}(x, \varphi(x), \mu_0, 0) \frac{\varphi'(x)}{g(x, \varphi(x), \mu_0, 0)} \mathrm{d}x; \tag{14.13}$$

$$\begin{aligned} \text{For } \tilde{\Gamma}(s): \ \tilde{I}(s, \mu_0) = &\int_{\omega_M}^{\alpha_s} \frac{\partial f}{\partial x}(x, \varphi(x), \mu_0, 0) \frac{\varphi'(x)}{g(x, \varphi(x), \mu_0, 0)} \mathrm{d}x \\ &+ \int_{\omega'_s}^{x_2} \frac{\partial f}{\partial x}(x, \varphi(x), \mu_0, 0) \frac{\varphi'(x)}{g(x, \varphi(x), \mu_0, 0)} \mathrm{d}x. \end{aligned} \tag{14.14}$$

Definition 1.2. A slow-fast cycle Γ is called non-degenerate, if all of its extreme (fold) points are non-degenerate, and it is not a transitory canard cycle.

For more explanation about transitory canard cycles, see the recent work [13].

Theorem 1.1. *If a slow-fast cycle is non-degenerate, then the following statements hold (see [11, 15] for example):*

(I) If $I(s,\mu_0)\neq 0$, *then* $Cycl(X_{\varepsilon,\mu},\Gamma(s),(0,\mu_0))\leq 1$. *Besides, if* $I(s,\mu_0)<0$ *(or* >0*) then the perturbed canard limit cycle from* $\Gamma(s)$ *is stable (or unstable).*

(II) If $I(s,\mu_0)=0$ *and* $\frac{\partial I}{\partial s}(s,\mu_0)\neq 0$, *then* $Cycl(X_{\varepsilon,\mu},\Gamma(s),(0,\mu_0))\leq 2$.

(III) If $I(s,\mu_0)=0$ *and* (s,μ_0) *is a zero point of* $\frac{\partial I}{\partial s}$ *with multiplicity* m, *then* $Cycl(X_{\varepsilon,\mu},\Gamma(s),(0,\mu_0))\leq 2+m$.

Remark 1.3. If the slow-fast cycle contains two extreme points, and we change $I(s,\mu_0)$ and $\Gamma(s)$ in above statements to $\tilde{I}(s,\mu_0)$ and $\tilde{\Gamma}(s)$, respectively, then Theorem 1.1 is also true.

Remark 1.4. It was proved in [13] that for the transitory slow-fast cycle of case I, shown in Fig. 14.5b, its cyclicity is at most 1 if the slow divergence integral along it is non-zero, and its cyclicity is at most 2 if the integral is zero.

14.2 A New Formula of Slow Divergence Integral

If the slow curve is U-shaped, for each $x\in[\omega_s,0]$ we define $\sigma(x)\in[0,\alpha_s]$ by

$$\varphi(x)=\varphi(\sigma(x)), \tag{14.15}$$

Hence for $x\in[\omega_s,0)$ we have that

$$\sigma'(x)=\frac{\varphi'(x)}{\varphi'(\sigma(x))}<0. \tag{14.16}$$

Similarly, if the slow curve is S-shaped (see Fig. 14.6b), for each $x\in[\omega_M,0]$ we define $\sigma_1(x)\in[0,x_2]$ and $\sigma_2(x)\in[x_2,x_0]$ by

$$\varphi(x)=\varphi(\sigma_j(x)),\ j=1,2, \tag{14.17}$$

and for $x\neq\omega_M$, $x\neq 0$ we have that

$$\sigma_1'(x)=\frac{\varphi'(x)}{\varphi'(\sigma_1(x))}<0,\ \sigma_2'(x)=\frac{\varphi'(x)}{\varphi'(\sigma_2(x))}>0. \tag{14.18}$$

Let

$$h(x)=\frac{\frac{\partial f}{\partial x}(x,\varphi(x),\mu_0,0)}{g(x,\varphi(x),\mu_0,0)}, \tag{14.19}$$

and let $x = \varphi^{-1}(y)$ be the single-valued inverse function of $y = \varphi(x)$ for $x \in [\omega_s, 0]$ in U-shaped case or for $x \in [\omega_M, 0]$ for S-shaped case. Then we have

Theorem 2.1. *The slow divergence integrals* (14.13) *and* (14.14) *can be changed, respectively, to*

$$I(s, \mu_0) = \int_0^s [h(\sigma(x)) - h(x)]|_{x=\varphi^{-1}(y)} \mathrm{d}y, \tag{14.20}$$

and

$$\tilde{I}(s, \mu_0) = \int_0^s [h(\sigma_1(x)) - h(x)]|_{x=\varphi^{-1}(y)} \mathrm{d}y + \int_s^{s_M} [h(\sigma_2(x)) - h(x)]|_{x=\varphi^{-1}(y)} \mathrm{d}y, \tag{14.21}$$

where $s_M = \varphi(x_2)$ *is the local maximum value at the jump point* $(x_2, \varphi(x_2))$.

Remark 2.1. The proof of Theorem 2.1 can be found in [26] or [24], where the results were generalized to the case when the slow curve has more than two fold points.

Remark 2.2. The formulas (14.20) and (14.21) for system (14.7) need certain conditions, especially the orientation of the vector field along the slow-fast cycle is counterclockwise. If we change some signs on the conditions such that the orientation is clockwise, then for the same function h the formulas (14.20) and (14.21) should be changed, respectively, to

$$I(s, \mu_0) = \int_0^s [h(x) - h(\sigma(x))]|_{x=\varphi^{-1}(y)} \mathrm{d}y, \tag{14.22}$$

and

$$\tilde{I}(s, \mu_0) = \int_0^s [h(x) - h(\sigma_1(x))]|_{x=\varphi^{-1}(y)} \mathrm{d}y + \int_s^{s_M} [h(x) - h(\sigma_2(x))]|_{x=\varphi^{-1}(y)} \mathrm{d}y. \tag{14.23}$$

See system (2.1), formulas (2.7) and (2.10) of [25].

14.3 Predator–Prey Systems with Response Functions of Holling Type

14.3.1 Introduction

The classical predator–prey systems with a response function $p(x)$ of Holling type can be written in the form

$$\begin{cases} \dot{x} = rx(1 - \dfrac{x}{K}) - yp(x) = p(x)(F(x) - y), \\ \dot{y} = y(-d + cp(x)), \\ \quad x(0) \geq 0, y(0) \geq 0, \end{cases} \tag{14.24}$$

where $x \geq 0$ and $y \geq 0$ denote the number or density of the prey and predator populations, respectively, r, K, d, and c are positive constants, and $F(x) = rx(1 - \frac{x}{K})/p(x)$.

For predator and prey, a functional response is the intake rate of a predator as a function of prey density. Following Holling [19, 20], the functional responses were classified into three types, which are called Holling type I, II, and III, respectively, with $p(x) = mx$, $p(x) = \frac{mx}{b+x}$, and $p(x) = \frac{mx^2}{a+x^2}$, where $m, a, b > 0$. The following response function:

$$p(x) = \frac{mx^2}{ax^2 + bx + 1}, \quad b > -2\sqrt{a} \tag{14.25}$$

is called generalized Holling type III, and

$$p(x) = \frac{mx}{ax^2 + bx + 1}, \quad b > -2\sqrt{a} \tag{14.26}$$

is called Holling type IV response function by Colling [9], and it is also called Monod–Haldane function (see Andrews [2]).

In all cases $p(x) > 0$ for $x > 0$, and the behavior of $p(x)$ is shown in Fig. 14.7 for different Holling types.

With Holling type I functional response the system (14.24) is the well-known Lotka–Volterra model. In the Holling type II functional response, the function is increasing and saturates, i.e., has a finite positive limit as x approaches infinity.

For the generalized response function of Holling type III, for x sufficiently large, $p(x)$ resembles the Holling type II model, the effect of inhibition is seen, although $p(x)$ has a local maximum if $-2\sqrt{a} < b < 0$. However, for x small the behavior of $p(x)$ is different from Holling type II model.

The response function of Holling type IV increases to a maximum and then decreases, approaching zero as x approaches infinity, is used to model the situation

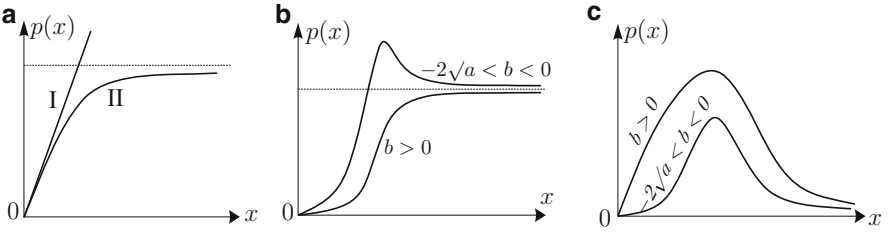


Fig. 14.7 Holling type functional responses. (**a**) Types I and II; (**b**) generalized type III; (**c**) type IV

where the prey can better defend or disguise themselves when their population becomes large enough, a phenomenon called group defence [18]. For x small $p(x)$ resembles the Holling type II model if $b > 0$ and resembles the generalized Holling type III model if $-2\sqrt{a} < b < 0$.

The predator–prey system (14.24) with Holling type of functional responses has been extensively studied by many authors, including the studies by May [27], Cheng [8], Chen and Jing [6], Cheng and Zhang [7], Wolkowicz [31], Wrzosek [32], Rothe and Shafer [29], Ruan and Xiao [30], Zhu, Campbell, and Wolkowicz [34], Xiao and Zhu [33], Broer et al. [4, 5], Lamontagne et al. [23], and recent work by Li and Zhu [25]. The readers can find an extended list of references in the papers [33, 34], and in the book [3] by Bazykin, where contains the description of some models accounting the Allee effect. In all of these studies, the existence and number of limit cycles are important topics in the bifurcation study of the predator–prey systems for a better understanding of many real world oscillatory phenomena in nature [1, 27, 28].

It was proved that the system (14.24) has at most one limit cycle for Holling type II functional response by Cheng [8] and Chen and Jing [6], and for original Holling type III functional response by Chen and Zhang [7]; the system (14.24) can have codimension 2 (and at most codimension 2) Hopf bifurcation for generalized Holling type III by Lamontagne et al. [23] and for Holling type IV by Xiao and Zhu [33]; the system (14.24) can have codimension 3 Bogdanov–Takens bifurcation for generalized Holling type III by Lamontagne et al. [23] and Holling type IV by Broer et al. [4, 5].

Note that in system (14.24) the parameter d is the death rate of the predator while c is the efficiency of the predator to convert prey into predators. In biology and ecology, it is also interesting to investigate the case when d and c are small. For example, the larger animals like lions or wolves do not need to prey so frequently. In this case the slow-fast dynamics occur. Li and Zhu [25] studied this case for response functions of all Holling types II, generalized III and IV. It was shown that at most two limit cycles bifurcate from the slow-fast cycles of U-shaped and at most one from S-shaped. Unlike the Hopf bifurcation or Bogdanov–Takens bifurcations, the limit cycles may appear in a large region in the phase space. Also, the codimension two Hopf bifurcation can be seen as a limit case of the slow-fast cycle bifurcation of order two when the slow-fast cycle shrinks to a canard point.

The study for system (14.24) with response function of Holling types II is relatively simple, the study in [25] was mainly for Holling type IV, and briefly for generalized Holling type III. As an example of the application of slow-fast dynamics, in this article we will introduce the study of system (14.24) with response functions of generalized Holling type III. The treatment is a little different from [25], especially the orientation of slow-fast cycles is counterclockwise, in order to use the formulas (14.20) and (14.21) directly. Of course, the results are the same as in [25].

Thus, we suppose that in system (14.24) the parameters d and c are small, and the function $p(x)$ is given in (14.25), i.e.,

$$p(x) = \frac{m x^2}{ax^2 + bx + 1}, \quad F(x) = \frac{r}{mx}(ax^2 + bx + 1)(1 - \frac{x}{K}), \quad b > -2\sqrt{a}, K > 0.$$

We eliminate a, m, and r by scaling of phase variables, time, and parameters. For this purpose, we let

$$(x, y, t) = \left(\frac{1}{\sqrt{a}}\bar{x}, \bar{r}\bar{y}, \frac{\sqrt{a}}{m\bar{r}}\,\bar{t}\right),$$

and

$$(r, K, b, d, c) = \left(\frac{m}{\sqrt{a}}\,\bar{r}, \frac{1}{\sqrt{a}}\,\bar{K}, \sqrt{a}\,\bar{b}, \frac{m\bar{r}}{\sqrt{a}}\,\bar{d}, \sqrt{a}\,\bar{r}\,\bar{c}\right).$$

Removing all bars, we obtain the same form (14.24) with $a = m = r = 1$, i.e., system (14.24) becomes

$$\dot{x} = p(x)(F(x) - y), \quad \dot{y} = y(-d + c\,p(x)), \tag{14.27}$$

where

$$p(x) = \frac{x^2}{x^2 + bx + 1}, \; F(x) = \frac{1}{x}(x^2 + bx + 1)\left(1 - \frac{x}{K}\right), \quad b > -2, K > 0. \tag{14.28}$$

Since $p(x) > 0$ for $x > 0$, we divide the two equations in (14.27) by $p(x)$, and still use $\mathrm{d}t$ for $p(x)\mathrm{d}t$. Besides let $d = \varepsilon, c = \alpha\varepsilon$, where $\varepsilon \geq 0$ small, then system (14.27) becomes

$$\frac{\mathrm{d}x}{\mathrm{d}t} = F(x) - y, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = \varepsilon y\left(\alpha - \frac{1}{p(x)}\right), \tag{14.29}$$

where $\alpha = \alpha(\varepsilon)$ will be chosen later. Hence, the slow curve is given by $y = F(x)$.

14.3.2 The Existence of Fold (Extreme) Points on Slow Curve

Let $\lambda = (b, K)$, then the non-degenerate fold points are given by $F_x = 0$ and $F_{xx} \neq 0$. Since $Kx > 0$ it is easy to find that $F_x = 0$ is equivalent to

$$\chi(x, \lambda) = 2x^3 + (b - K)x^2 + K = 0, \tag{14.30}$$

which has always a negative root, because $K > 0$. On the other hand, $F_{xx} = 0$ is equivalent to $x^3 - K = 0$. Removing x from these two equations we find that $F_x = F_{xx} = 0$ is given by the curve

$$C_1: \quad C_1(\lambda) = (K - b)^3 - 27K = 0, \tag{14.31}$$

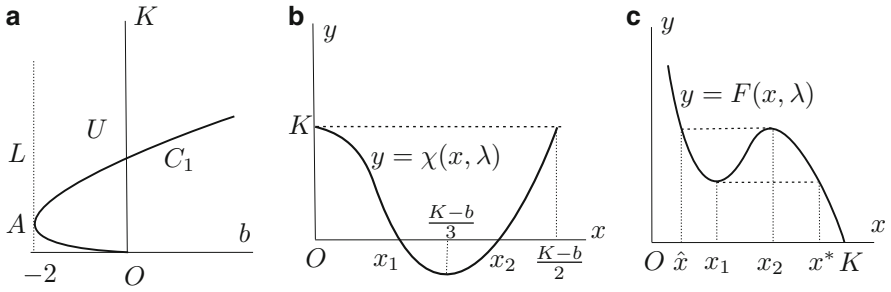


Fig. 14.8 (**a**) The behavior of C_1; (**b**) The values $x_1 < x_2$; (**c**) The behavior of $F(x, \lambda)$ for $\lambda \in U$

which is a third order parabola and is tangent to the straight line $L : \{b = -2\}$ at the point $A(-2, 1)$, see Fig. 14.8a. If $\lambda = (b, K)$ is located right to C_1, then (14.30) has only a negative root, hence $F(x, \lambda)$ is monotone for $x > 0$, there is no slow-fast cycle. If λ is located in the narrow region below C_1, right to L and above $\{K = 0\}$, then both of two positive zeros of (14.30) are greater than K, hence F is negative at these values, we do not need to consider. It is clear that only in the region above C_1, that is,

$$U : \quad \{\lambda = (b, K) \,|\, C_1(\lambda) > 0, b + 2 > 0 \text{ and } K > 1\},$$

Equation (14.30) has two zeros x_1 and x_2 satisfying $0 < x_1 < x_2 < K$ and $F''(x_j) \neq 0$ for $j = 1, 2$.

Note that $\chi_x(x, \lambda) = 2[3x - (K - b)]x$ and $\chi(0, \lambda) = \chi(\frac{K-b}{2}, \lambda) = K > 0$ (see Fig. 14.8b), and from (14.31) it is obvious that $K > b$ for $\lambda \in U$ because $K > 0$, we have that

$$0 < x_1 < \frac{K - b}{3} < x_2 < \frac{K - b}{2}, \quad \lambda \in U. \tag{14.32}$$

Lemma 3.1. *Only for $\lambda = (b, K) \in U$ the slow curve $y = F(x)$ has a unique local (non-degenerate) minimum point at $(x_1, F(x_1))$ and a unique local (non-degenerate) maximum point at $(x_2, F(x_2))$, shown in Fig. 14.8c.*

14.3.3 The Existence of Slow-Fast Cycles for $\varepsilon = 0$

We consider that $(x_j, F(x_j))$ is a canard point, where $j = 1$ or $j = 2$, then, from the condition for Eq. (14.8), we need choose $\alpha = \frac{1}{p(x_j)} - \bar{\lambda}$, where $\bar{\lambda} = \bar{\lambda}(\varepsilon) \to 0$ as $\varepsilon \to 0$, and system (14.29) takes the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = F(x) - y, \quad \frac{\mathrm{d}y}{\mathrm{d}t} = \varepsilon y \left[\phi(x)\frac{x - x_j}{x_j^2 x^2} - \bar{\lambda} \right] = \varepsilon g(x, y, \bar{\lambda}), \tag{14.33}$$

where $\phi(x) = (1 + bx_j)x + x_j$. Since $F''(x_1) > 0$ and $F''(x_2) < 0$, to guarantee the existence of a slow-fast cycle around the point $(x_j, F(x_j))$ for $j = 1, 2$, we need a condition

$$\phi(x_j) = (1 + bx_j)x_j + x_j = (2 + bx_j)x_j > 0, \text{ for } \lambda \in U.$$

If $b \geq 0$ we certainly have $(2 + bx_j) > 0$; if $b \in (-2, 0)$, we first consider the critical case $(2 + bx_j) = 0$, which gives by (14.30) that

$$\chi\left(-\frac{2}{b}, \lambda\right) = \frac{(b-2)(b+2)(Kb+4)}{b^3}, \quad b \in (-2, 0).$$

From (14.31) we find

$$C_1(\lambda)|_{K=-\frac{4}{b}} = \frac{(b^2+16)(b^2-2)^2}{(-b)^3} > 0, \quad b \in (-2, 0).$$

Hence in (b, K)-plane $(2 + bx_j) = 0$ is given by the following curve:

$$C_2: \quad C_2(\lambda) = bK + 4 = 0 \text{ for } b \in (-\sqrt{2}, 0).$$

Besides, C_2 is tangent to C_1 at the point $P(-\sqrt{2}, 2\sqrt{2})$, and divide U into U_1, U_2, and U_3, see Fig. 14.9a.

We denote the part of C_2 below P by C_2' and the part above P still by C_2. It is easy to check that $(2 + bx_j) > 0$ for $j = 1, 2$ and $\lambda \in U_2$, and $(2 + bx_1) > 0$ but $(2 + bx_2) < 0$ when λ crosses C_2 into U_1, and $(2 + bx_j) < 0$ for $j = 1, 2$ when λ crosses C_2' into U_3. Now let

$$V_1 = U_1 \cup C_2 \cup U_2, \quad V_2 = U_2 \subset V_1. \tag{14.34}$$

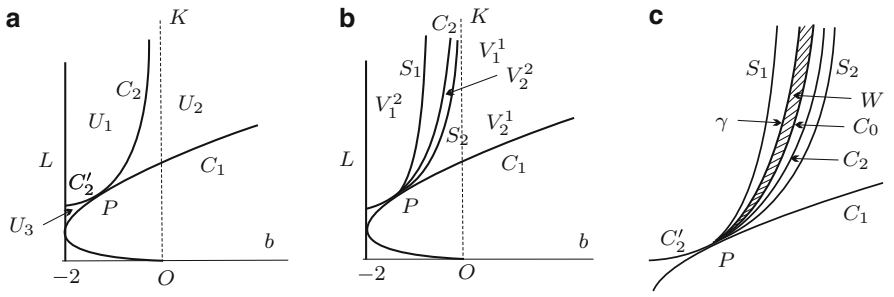


Fig. 14.9 The regions V_1^j, V_2^j, and W. (**a**) $V_1 = U_1 \cup C_2 \cup U_2$, $V_2 = U_2$. (**b**) S_1 divides V_1, S_2 divides V_2. (**c**) Shaded region W

Lemma 3.2. *A slow-fast cycles appear around* $(x_1, F(x_1))$ *only if* $\lambda \in V_1$*, and a slow-fast cycles appear around* $(x_2, F(x_2))$ *only if* $\lambda \in V_2$.

14.3.4 The Existence of a Saddle Point on a Slow-Fast Cycle

We consider the possibility that a saddle point appears on a slow-fast cycle. For a slow-fast cycle Γ_1 around $(x_1, F(x_1))$, since $\phi(x)$ is linear in x and $\phi(0) > 0, \phi(x_1) > 0$, a saddle point on Γ_1 must appear from the upper-right corner, that is $\phi(x_2) \leq 0$, and the critical case is $\phi(x_2) = 0$. Similarly, a critical case for a saddle point appearing on a slow-fast cycle Γ_2 around $(x_2, F(x_2))$ is $\phi(x^*) = 0$, for the positions of x_1, x_2, and x^*, see Fig. 14.8c. Thus the critical conditions for appearing of a saddle point on Γ_1 and Γ_2 are, respectively,

$$(1 + bx_1)x_2 + x_1 = 0, \quad b \in (-2, 0) \tag{14.35}$$

and

$$(1 + bx_2)x^* + x_2 = 0, \quad b \in (-2, 0). \tag{14.36}$$

From $F(x) - F(x_1) = \frac{(x-x_1)^2(x^*-x)}{Kx}$ and $F'(x_2) = 0$ we find x^*, and similarly to find $\hat{x}$, see Fig. 14.8c:

$$x^* = \frac{2(x_2)^2}{x_1 + x_2}, \quad \hat{x} = \frac{2(x_1)^2}{x_1 + x_2}. \tag{14.37}$$

Substituting the expression of x^* in (14.36), we find

$$(3 + 2bx_2)x_2 + x_1 = 0, \quad b \in (-2, 0). \tag{14.38}$$

Eliminating x_1 from (14.35) and (14.30) with $x = x_1$, and eliminating x_2 from the resulting equation and (14.30) with $x = x_2$, we obtain a curve

$$S_1 : \{\lambda \in V_1 \mid (b^2 - 1)bK + b^2 + 2 = 0,\ b \in (-\sqrt{2}, -1), K > 2\sqrt{2}\},$$

which is strictly monotone. Since $S_1(\lambda)|_{K=-\frac{4}{b}} = 3(2 - b^2)$, S_1 is located entirely left to C_2 for $b > -\sqrt{2}$, and as $b \to -\sqrt{2}$ the two curves S_1 and C_2 have the same end-point $P \in C_1$, shown in Fig. 14.9b.

Similarly, eliminate x_1 from (14.38) and (14.30) with $x = x_1$, then eliminate x_2 from the resulting equality and (14.30) with $x = x_2$, we obtain

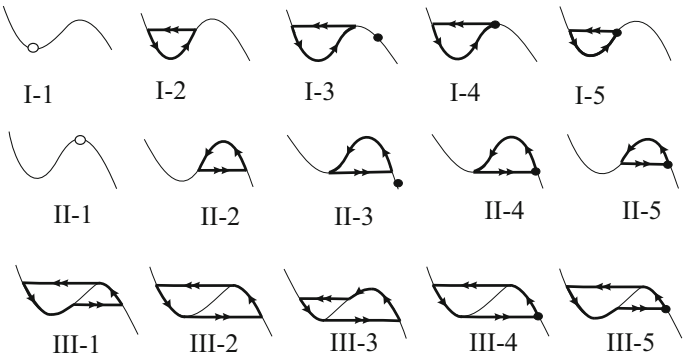


Fig. 14.10 The different types of slow-fast cycles for generalized Holling type III response function

$$b^2(b^2-1)K^5 + b(4b^4+3b^2-6)K^4 + (6b^6-12b^4+32b^2-9)K^3$$
$$+b(4b^6+12b^4-132b^2+209)K^2 + (b^8-3b^6+32b^4-209b^2+343)K$$
$$+b^3(b^2-3)^2 = 0,$$

which defines a monotone curve S_2 for $\lambda \in [-\sqrt{2}, 0)$ (as well as some extra curves). S_2 is right to C_2 for $b > -\sqrt{2}$, and the two curves have the same end-point $P \in C_1$. For $j = 1$ or $j = 2$ the curve S_j divides the region V_j into two disjoint parts: V_j^1 on the right and V_j^2 on the left, see Fig. 14.9b. Note that $(V_2^1 \cup S_2 \cup V_2^2) \subset V_1^1$.

Lemma 3.3. *For $j = 1$ or 2 if $\lambda \in V_j^1$ then there is no saddle point on a slow-fast cycle around the fold point $(x_j, F(x_j))$; if $\lambda \in V_j^2 \cup S_j$ then there is a saddle point on any slow-fast cycle around the fold point $(x_j, F(x_j))$, and in the critical case $\lambda \in S_j$ the saddle point is located at the upper-right corner of the slow-fast cycle for $j = 1$ or at the lower-right corner of the slow-fast cycle for $j = 2$, see I-4 and II-4 of Fig. 14.10, respectively.*

14.3.5 Main Results

We will show that there are two monotone curves $C_0 = \{\lambda \in V_1 \,|\, b^2 - bK - 6 = 0\}$ and γ, both are located between C_2 and S_1, having the common end-point at P, and γ is located between C_0 and S_1. Denote the region between C_0 and γ by W, see Fig. 14.9c. We list some possible slow-fast cycles in Fig. 14.10.

Theorem 3.1. *For system (14.33) the following statements hold, where $\lambda = \lambda(\varepsilon, b, K)$ for small $\varepsilon > 0$, and the different types of slow-fast cycles are shown in Fig. 14.10:*

(A) The cyclicity is at most one for a slow-fast cycle of type

(a) *I-2, if* $(b, K) \in V_1$, *especially I-3, I-4, or I-5, if* $(b, k) \in V_1^1, S_1$, *or* V_1^2, *respectively;*
(b) *II-2, if* $(b, K) \in V_2$, *especially II-3, II-4, or II-5, if* $(b, k) \in V_2^1, S_2$, *or* V_2^2, *respectively;*
(c) *III-1, if* $(b, K) \in V_1^1 \supset V_2$;
(d) *III-2 or III-3, if* $(b, K) \in V_2^1$;
(e) *III-4, if* $(b, K) \in S_2$;
(f) *III-5, if* $(b, K) \in V_2^2$.

(B) *The cyclicity is at most two for a slow-fast cycle of type*

(a) *I-2, if* $(b, K) \in W$;
(b) *I-3, if* $(b, K) \in \gamma$.

(C) *There is no slow-fast cycle with cyclicity three or higher than three, except for the cases I-1 and II-1.*

Remark 3.1. From this theorem we see that for system (14.33) with $0 < \varepsilon \ll 1$ at most two limit cycles can be bifurcated from a U-shaped slow-fast cycle around the fold point $(x_1, F(x_1))$, and at most one limit cycle can be bifurcated from a U-shaped slow-fast cycle around the fold point $(x_2, F(x_2))$, or from a S-shaped slow-fast cycle related to these two fold points.

Remark 3.2. For the cases I-1 and II-1 in Fig. 14.10 we need to find the maximal number of small limit cycles born from a singular point by perturbations, and this cannot be studied only by slow divergence integral, see, for example, [16]. In the paper [23] by Lamontagne et al. (Propositions 6.8, 6.9 and Theorem 6.1), it was shown that for the general system (14.27) the Hopf bifurcation has codimension at most two, and the codimension two happens exactly when $(b, K) \in C_0$. In fact, by a change of variables $(x, y, t) \mapsto (Kx, cKy, \frac{t}{cK^2})$, the system (14.27) is transformed to system (1.3) of [23] with $\alpha = K^2, \beta = bK, \rho = \frac{1}{cK^2}$, and $\delta = \frac{d}{cK^2}$, and the equation of $C_0 : b^2 = bK + 6$ becomes $\beta^2 = \alpha(\beta + 6)$, which is the expression (6.15) of [23], corresponding to Hopf bifurcation of codimension two. This hints a fact that for system (14.33) at most two small limit cycles can be bifurcated from a singular point for $(b, K) \in C_0$. The problem is that we need to prove the uniformity with respect to small ε. We also remark that the point P in Fig. 14.9, with $(b, K) = (-\sqrt{2}, 2\sqrt{2})$, corresponds to the point S_d with $(\alpha, \beta) = (8, -4)$ of [23], at this point an attracting Bogdanov–Takens bifurcation of codimension 3 occurs (Theorem 5.2 of [23]).

Remark 3.3. The cases I-3 and II-3 in Fig. 14.10 correspond to a transitory slow-fast cycle of Type I, defined in [13]. It was proved in [13] that for this type of slow-fast cycle the cyclicity is one if $I \neq 0$ and is two if $I = 0$, where I is the corresponding slow divergence integral. We will prove that $I = 0$ in case I-3 if $(b, K) \in \gamma$, and $I \neq 0$ for other cases. This implies the results of Theorem 3.1 for these cases.

Remark 3.4. In cases I-4, I-5, II-4, II-5, III-4, and III-5 in Fig. 14.10, a saddle point appears on a corner of the slow-fast cycle. We will prove that in these cases the

corresponding slow divergence integral is not equal to zero, hence by Theorem 2.22 of [11], we obtain that the cyclicity is also at most one.

14.3.6 *Proof of Theorem 3.1*

We will study the slow divergence integral in form (14.20) or (14.21), then use Theorem 1.1 to give a proof, except for cases I-1, II-1, and I-3, II-3 in Statements (A) and (B), as we explained in Remarks 3.2 and 3.3.

We first consider the cyclicity of a U-shaped slow-fast cycle around the canard point $(x_1, F(x_1))$. Since $F'(x_1) = 0$, we have

$$F'(x) = F'(x) - F'(x_1) = \frac{K(x + x_1) - 2(xx_1)^2}{K(xx_1)^2}(x - x_1).$$

Then by the formula (14.19) from system (14.33) with $j = 1$ and $\bar{\lambda}(0) = 0$ we obtain

$$h(x) = \frac{K(x + x_1) - 2(xx_1)^2}{KF(x)[(1 + bx_1)x + x_1]}.$$

By formula (14.20) we need to study the slow divergence integral

$$I(s, \lambda) = \int_{Y_1}^{s} \varphi(x(y), \lambda)\, \psi(x(y), \lambda)\, dy, \quad \lambda \in V_1, \tag{14.39}$$

where $s \in (Y_1, Y_2)$, $Y_1 = F(x_1(\lambda))$ and $Y_2 = F(x_2(\lambda))$, $x(y) = F^{-1}(y) \in (\hat{x}, x_1)$ for $y \in (Y_1, s)$, and

$$\begin{aligned} \varphi(x, \lambda) &= \frac{x_1^2(\sigma(x) - x)}{KF(x)[(1 + bx_1)x + x_1][(1 + bx_1)\sigma(x) + x_1]} > 0, \\ \psi(x, \lambda) &= -2(bx_1 + 1)x\sigma(x) - 2x_1(x + \sigma(x)) - bK. \end{aligned} \tag{14.40}$$

where $\sigma(x) \in (x_1, x_2)$ is defined by $F(x) = F(\sigma(x))$ for $x \in (\hat{x}, x_1)$ as in Sect. 14.2.

Remark 3.5. In [25] we consider system in form (2.11) of that paper, the orientation of a slow-fast cycle is clockwise, but here we use the form (14.33) directly, and the orientation of a slow-fast cycle is counterclockwise, so the slow divergence integral has a different sign, see Remark 2.2. Especially, the function $\varphi(x, \lambda)$ above has a different sign from [25].

We need to study the number of zero of $I(s, \lambda)$ in $s \in (Y_1, Y_2)$ for $\lambda \in V_1$. For this purpose, a crucial step is to study the number of zeros of $\psi(x, \lambda)$ for $x \in (\hat{x}, x_1)$ and $\lambda \in V_1$. We make a change of $(x, \sigma(x))$ to (ξ, η) as follows

$$\xi = \sigma(x) + x, \quad \eta = \sigma(x)x, \tag{14.41}$$

It is obviously that for $x \in (\hat{x}, x_1)$ the minimum and maximum values of ξ are

$$\xi_{1m} = 2x_1, \quad \xi_{1M} = \hat{x} + x_2 = \frac{2x_1^2 + x_1x_2 + x_2^2}{x_1 + x_2}. \tag{14.42}$$

From $F(\sigma(x)) = F(x)$ we find

$$[(K - b) - (x + \sigma(x))]x\sigma(x) = K. \tag{14.43}$$

Lemma 3.4. *For $\lambda \in V_1$ and $x \in (\hat{x}, x_1)$ we have that (1) $K - b - \xi > 0$ and (2) $\frac{d\xi}{dx} < 0$.*

In fact, if $\lambda \in V_1$ and λ near the curve C_1 then by (14.32) ξ near $2x_1 \sim \frac{2(K-b)}{3}$, hence $K - b - \xi > 0$ for ξ near ξ_{1m}; from (14.43) we see that $[(K - b) - \xi]$ keeps a same sign for all $\lambda \in V_1$, which proves the statement (1) of Lemma 3.4. We note that

$$\frac{d\xi}{dx} = 1 + \sigma'(x) = \frac{F'(\sigma(x)) + F'(x)}{F'(\sigma(x))}.$$

It is obviously $F'(\sigma(x)) > 0$ and computation gives

$$F'(x) + F'(\sigma(x)) = \frac{2[(K - b) - (x + \sigma(x))](x\sigma(x))^2 - K(x^2 + \sigma^2(x))}{K(x\sigma(x))^2}.$$

By using (14.43) we immediately get

$$F'(x) + F'(\sigma(x)) = -\frac{(\sigma(x) - x)^2}{(x\sigma(x))^2} < 0.$$

The statement (2) of Lemma 3.4 is verified.

By using (14.41), (14.43) and Lemma 3.4 (1) we have

$$\eta = \frac{K}{K - b - \xi}. \tag{14.44}$$

Substituting (14.41) and (14.44) in the second equality of (14.40), we find

$$\psi(x, \lambda) = \frac{\psi_1(\xi, \lambda)}{K - b - \xi}, \tag{14.45}$$

where

$$\psi_1(\xi, \lambda) = 2x_1\xi^2 + [bK - 2x_1(K - b)]\xi - K[b(K - b) + 2(1 + bx_1)]. \tag{14.46}$$

By Lemma 3.4 (2), instead to study the number of zeros of $\psi(x,\lambda)$ for $x \in (\hat{x}, x_1)$ it is enough to study the number of zeros of $\psi_1(\xi,\lambda)$ for $\xi \in (\xi_{1m}, \xi_{1M})$.

From (14.46), (14.42) and (14.30) with $x = x_1$ we obtain

$$\psi_1(\xi_{1m},\lambda) = 8(x_1)^3 + 4(b-K)(x_1)^2 + K(b^2 - Kb - 2) = K(b^2 - bK - 6). \quad (14.47)$$

We define a curve

$$C_0 : \{\lambda \in V_1 \,|\, b^2 - bK - 6 = 0,\ b \in (-\sqrt{2}, 0), K > 2\sqrt{2}\}. \quad (14.48)$$

It is not hard to check that the curve C_0 is located strictly between S_1 and C_2 and has the same end-point at P as $b \to -\sqrt{2}$, see Fig. 14.9c. Thus, we have the following result:

Lemma 3.5. *$\psi_1(\xi_{1m},\lambda) > 0, = 0$ or < 0 if $\lambda \in V_1$ is left, on or right to C_0, respectively.*

From (14.46) and (14.42) we find $\psi_1(\xi_{1M},\lambda) = \frac{\zeta(\lambda)}{(x_2+x_1)^2}$, where

$$\begin{aligned}\zeta(\lambda) = {} & 2x_1[4(x_1)^4 + 4(x_1)^3x_2 + 5(x_1)^2(x_2)^2 + 2x_1(x_2)^3 + (x_2)^4] \\ & +2(b-K)[2(x_1)^4 + 3(x_1)^3x_2 + 2(x_1)^2(x_2)^2 + x_1(x_2)^3] \\ & +K(b^2 - bK - 2)(x_1 + x_2)^2 - bK(x_1)^2x_2.\end{aligned} \quad (14.49)$$

Eliminating x_1 from (14.49) and (14.30) with $x = x_1$, then eliminating x_2 from the resulting equality and (14.30) with $x = x_2$, we obtain

$$(b^2 - bK - 6)\,[(K-b)^3 - 27K]\,[(b^3 - b)K + b^2 + 2] = 0.$$

This means that a necessary condition for $\psi_1(\lambda) = 0$ is $\lambda \in C_0 \cup C_1 \cup S_1$. It can be checked that $\psi_1(\lambda) < 0$ if $\lambda \in C_0 \cup C_1$ and $\psi_1(\lambda) = 0$ only if $\lambda \in S_1$, so we have the following result:

Lemma 3.6. *$\psi_1(\xi_{1M},\lambda) > 0, = 0$ or < 0 if $\lambda \in V_1$ is left, on or right to S_1 respectively.*

Since $\psi_1(\xi,\lambda)$ is a quadratic polynomial in ξ, by Lemmas 3.5 and 3.6 it is easy to obtain the next Lemma.

Lemma 3.7. *For $\lambda = (b,K) \in V_1$, the following statements hold:*

(1) $\psi_1(\xi,\lambda) > 0$ for all $\xi \in (\xi_{1m}, \xi_{1M})$ if λ is located left to the curve S_1;
(2) $\psi_1(\xi,\lambda) < 0$ for all $\xi \in (\xi_{1m}, \xi_{1M})$ if λ is located right to the curve C_0;
(3) For each λ between S_1 and C_0, there is a unique $\xi^*_\lambda \in (\xi_{1m}(\lambda), \xi_{1M}(\lambda))$, continuous in λ, such that $\psi_1(\xi^*_\lambda,\lambda) = 0$, and

$$(\xi - \xi^*_\lambda)\psi_1(\xi,\lambda) < 0, \text{ for } \xi \in (\xi_{1m}(\lambda), \xi_{1M}(\lambda)) \setminus \{\xi^*_\lambda\}.$$

Moreover, $\lim_{\lambda\to C_0}\xi^*_\lambda = \xi_{1m}$, $\lim_{\lambda\to S_1}\xi^*_\lambda = \xi_{1M}$.

Since $\frac{\mathrm{d}\xi}{\mathrm{d}x} < 0$ and $\frac{\mathrm{d}y}{\mathrm{d}x} = F'(x) < 0$ for $x \in (\hat{x}, x_1)$, by (14.40) and (14.45) we can transform Lemma 3.7 to the following form:

Lemma 3.8. *For* $\lambda = (b, K) \in V_1$, *the following statements hold:*

(1) $\psi(x(y), \lambda) > 0$ *for all* $y \in (Y_1, Y_2)$ *if* λ *is located left to the curve* S_1*;*
(2) $\psi(x(y), \lambda) < 0$ *for all* $y \in (Y_1, Y_2)$ *if* λ *is located right to the curve* C_0*;*
(3) For each λ *between* S_1 *and* C_0, *there is a unique* $y^*_\lambda \in (Y_1, Y_2)$, *continuous in* λ, *such that* $\psi(x(y^*_\lambda), \lambda) = 0$, *and*

$$(y - y^*_\lambda)\psi(x(y), \lambda) < 0, \text{ for } y \in (Y_1, Y_2) \setminus \{y^*_\lambda\}. \tag{14.50}$$

Moreover, $\lim_{\lambda\to C_0} y^*_\lambda = Y_1$, $\lim_{\lambda\to S_1} y^*_\lambda = Y_2$.

By (14.39), (14.40) and statements (1) and (2) of Lemma 3.8 we obtain that $I(s, \lambda) > 0$ for all $s \in (Y_1, Y_2)$ if $\lambda \in V_1$ is located left to the curve S_1; and $I(s, \lambda) < 0$ for all $s \in (Y_1, Y_2)$ if $\lambda \in V_1$ is located right to the curve C_0. Hence in these cases, by Theorem 1.1, the cyclicity of possible slow-fast cycles around the canard point $(x_1, F(x_1))$ is at most one.

It remains to consider the case $\lambda \in G$, which is the region between the curves S_1 and C_0, expressed, respectively, by $b = b_s(K)$ and $b = b_0(K)$ for $K > 2\sqrt{2}$, see Fig. 14.9c.

Let $s_\ell(\lambda) = Y_1 + \ell\,(Y_2 - Y_1)$, $\ell \in (0, 1)$. Recall that $Y_1 = Y_1(\lambda)$ and $Y_2 = Y_2(\lambda)$, and the slow-fast cycles $\Gamma(Y_1)$, $\Gamma(s_\ell)$, and $\Gamma(Y_2)$ look like, respectively, I-1, I-2, and I-3 of Fig. 14.10.

Lemma 3.9. *The following statements hold:*

(a) For each $K > 2\sqrt{2}$ *and* $\ell \in (0, 1)$, *there is a unique* $b = b_\ell(K) \in (b_s(K), b_0(K))$ *such that* $I(s_\ell(\lambda_{K,\ell}), \lambda_{K,\ell}) = 0$ *and* $\frac{\mathrm{d}I}{\mathrm{d}s}(s_\ell(\lambda_{K,\ell}), \lambda_{K,\ell}) \neq 0$ *with* $\lambda_{K,\ell} = (b_\ell(K), K)$. *That is, the cyclicity of* $\Gamma(s_\ell(\lambda_{K,\ell}))$ *is at most two.*
(b) For any fixed $\ell \in (0, 1)$, *the set* $\gamma_\ell = \{\lambda = \lambda_{K,\ell},\ K > 2\sqrt{2}\}$ *forms a curve in* G, *and* $\lim_{K\to 2\sqrt{2}}(b_\ell(K), K) = P$ *for all* $\ell \in (0, 1)$. *Moreover,* $b_0(K) = \lim_{\ell\to 0+0} b_\ell(K)$, *i.e.,* $\gamma_0 = C_0$, *and* $b_1(K) = \lim_{\ell\to 1-0} b_\ell(K) \in (b_s(K), b_0(K))$, *i.e.,* $\gamma = \{(b_1(K), K) : K > 2\sqrt{2}\}$ *is located strictly between* C_0 *and* S_1, *see Fig. 14.9c.*
(c) The curves $\{\gamma_\ell : 0 \le \ell \le 1\}$ *give a foliation of the region* $W \subset G$, *where the corresponding slow-fast cycle* $\Gamma(s_\ell(\lambda_{K,\ell}))$ *has cyclicity at most two.*

To verify statement (a), we first choose $b \in (b_s(K), b_0(K))$ to be very close to $b_s(K)$, such that $y^*_\lambda > s_\ell(\lambda)$. By Lemma 3.8(3) this is possible, and by formulas (14.39), (14.40), and (14.50), we have $I(s_\ell(\lambda), \lambda) > 0$, since $\varphi > 0$ and $\psi > 0$ for $y \in (Y_1, s_\ell(\lambda))$. We next increases b along the segment $(b_s(K), b_0(K)$, such that $y^*_\lambda < s_\ell(\lambda)$ (by Lemma 3.8(3), this is also possible), then

$$I(s_\ell(\lambda),\lambda)=\int_{Y_1(\lambda)}^{y_\lambda^*}\varphi(x(y),\lambda)\,\psi(x(y),\lambda)dy+\int_{y_\lambda^*}^{s_\ell(\lambda)}\varphi(x(y),\lambda)\,\psi(x(y),\lambda)dy.$$

The first integral is positive while the second negative, see (14.50). If we continuously increases b along the segment $(b_s(K),b_0(K))$, then $y_\lambda^*\to Y_1(\lambda)$ as $b\to b_0(K)$, hence the first integral goes to zero. On the other hand, $(s_\ell(\lambda)-y_\lambda^*)\to(s_\ell(\lambda)-Y_1)>0$ as $b\to b_0(K)$. Therefore, $I(s_\ell(\lambda),\lambda)<0$ for $0<b_0(K)-b\ll 1$, hence $I(s_\ell(\lambda),\lambda)$ has a odd number of zeros for $b\in(b_s(K),b_0(K))$.

Suppose that $b_\ell(K)\in(b_s(K),b_0(K))$ is such a zero point, i.e., $I(s_\ell(\lambda_{K,\ell}),\lambda_{K,\ell})=0$ with $\lambda_{K,\ell}=(b_\ell(K),K)$. Then by the uniqueness of zero of $\psi(x(y),\lambda)$ in y we immediately get $\psi(s_\ell(\lambda_{K,\ell}))\neq 0$, hence $\frac{\mathrm{d}I}{\mathrm{d}s}(s_\ell(\lambda_{K,\ell}),\lambda_{K,\ell})\neq 0$, and this also implies that $I(s_\ell(\lambda),\lambda)$ has a unique zero for $b\in(b_s(K),b_0(K))$.

The proofs of statements (b) and (c) are simple, we omit them here. From statement (b) we see that for $\lambda\in G\setminus\{W\cup\gamma\}$ the cyclicity of any slow-fast cycles is at most one.

We next consider the U-shaped slow-fast cycle around $(x_2,F(x_2))$. Since the discussion is very similar to above, we only indicate the differences. Instead of (14.39) we need to consider the slow divergence integral

$$I(s,\lambda)=\int_s^{Y_2}\varphi(x(y),\lambda)\,\psi(x(y),\lambda)\,dy,\quad \lambda\in V_2, \tag{14.51}$$

where the functions φ and ψ have similar expression (14.40) but change x_1, x, and $\sigma(x)$ in the right-hand sides to x_2, $\sigma_1(x)$, and $\sigma_2(x)$, respectively, with $\hat{x}<x<x_1<\sigma_1(x)<x_2<\sigma_2(x)<x^*$, see Fig. 14.8c.

Instead of (14.41) we let

$$\xi=\sigma_1(x)+\sigma_2(x),\quad \eta=\sigma_1(x)\sigma_2(x).$$

Note that $V_2\subset V_1$, $F(\sigma_j(x))=F(x)$ for $j=1,2$, hance $\frac{\mathrm{d}\xi}{\mathrm{d}x}=\frac{F'(x)[F'(\sigma_1(x))+F'(\sigma_2(x))]}{F'(\sigma_1(x))F'(\sigma_2(x))}$, $F'(x)<0, F'(\sigma_1(x))>0, F'(\sigma_2(x))<0$, and

$$F'(\sigma_1(x))+F'(\sigma_2(x))=-\frac{(\sigma_2(x)-\sigma_1(x))^2}{\sigma_1(x)\sigma_2(x)}<0.$$

Hence, Lemma 3.4 is still true for $\xi=\sigma_1(x)+\sigma_2(x)$, $\eta=\sigma_1(x)\sigma_2(x)$, and $\lambda\in V_2$.

Thus, if we replace x_1 by x_2 then obtain the same expression (14.46), and denote it by $\psi_2(\xi,\lambda)$. We also have $\xi_x<0$ for $x\in(\hat{x},x_1)$ as in Lemma 3.4 (2), but instead of (14.42) we have that the minimum and maximum values for ξ are, respectively,

$$\xi_{2m}=x_1+x^*=\frac{(x_1)^2+x_1x_2+2(x_2)^2}{x_2+x_1},\quad \xi_{2M}=2x_2. \tag{14.52}$$

Similarly to obtain (14.47) we find

$$\psi_2(\xi_{2M}, \lambda) = \psi_2(2x_2, \lambda) = K(b^2 - bK - 6). \tag{14.53}$$

Since V_2 is located entirely right to the curve C_0, we have $\psi_2(\xi_{2M}, \lambda) < 0$ for all $\lambda \in V_2$. Let us show that $\psi_2(\xi_{2m}, \lambda) < 0$ for $\lambda \in V_2$, hence $\psi_2(\xi, \lambda) < 0$ for all $\xi \in (\xi_{2m}, \xi_{2M})$ and $\lambda \in V_2$, because ψ_2 is a quadratic polynomial in ξ and the coefficient of ξ^2 is positive, and this implies that $I(s, \lambda) < 0$ for all $s \in (Y_1(\lambda), Y_2(\lambda))$ and $\lambda \in V_2$, i.e., the cyclicity of the corresponding slow-fast cycles is at most one.

The expression of $\psi_2(\xi_{2m}, \lambda)$ is the same as $\psi_1(\xi_{1M}, \lambda)$, exchanging x_1 and x_2. Hence, by the same procedure we obtain that a necessary condition for $\psi_2(\xi_{2m}, \lambda)$=0 is $\lambda \in C_0 \cup C_1 \cup S_1$, but $V_2 \cap \{C_0 \cup C_1 \cup S_1\} = \emptyset$, hence $\psi_2(\xi_{2m}, \lambda)$ has a fixed sign for all $\lambda \in V_2$. Choosing a special value $\lambda \in V_2$, we can find that $\psi_1(\xi_m, \lambda) < 0$ for $\lambda \in V_2$.

At last we consider the S-shaped slow-fast cycle $\tilde{\Gamma}$ with $(x_1, F(x_1))$ and $(x_2, F(x_2))$ as two fold points. As we discussed above that this is possible only for $\lambda \in V_2 \subset V_1$. We only study the types III-1 and III-2 in Fig. 14.10, because the proof for type III-3 is completely the same as for type III-1, and types III-4 and III-5 are similar to III-2 and III-1, respectively, having a saddle point at the lower-right corner of $\tilde{\Gamma}$, by Theorem 2.22 of [11] the conclusion is the same, because we will prove that the slow divergence integral is non-zero.

Now we use $\tilde{\Gamma}(s)$ and $\tilde{I}(s, \lambda)$ to denote the slow-fast cycle and the corresponding slow divergence integral. If $s = 0$ then it is type III-2 and if $s \in (0, Y_2)$ it is type III-1, see Fig. 14.6b. The type III-2 is a common slow-fast cycle, and its cyclicity is at most one, see [12]. In fact it is easily to check directly by the formula (14.21) that $\tilde{I}(0, \lambda) < 0$ for all $\lambda \in V_2$. When $s = Y_2$, we have the critical case-transitory slow-fast cycles, see Fig. 14.5b and [13]. Since the region V_2 is right to the curve C_0, by Lemma 3.8(2), (14.39) and $\varphi > 0$ we have $\tilde{I}(Y_2, \lambda) < 0$ for all $\lambda \in V_2$. Therefore, by the following lemma we immediately get $\tilde{I}(s, \lambda) < 0$ for all $s \in (0, Y_2)$ and $\lambda \in V_2$, which gives a proof for type III-1.

Lemma 3.10. *For all $\lambda \in V_2$ and $s \in (0, Y_2)$ we have $\frac{\mathrm{d}\tilde{I}}{\mathrm{d}s}(s, \lambda) > 0$.*

To prove this fact, we note that for a type III-1 slow-fast cycle, the point $(x_1, F(x_1))$ is a canard point while $(x_2, F(x_2))$ is a jump point, shown in Fig. 14.6b. Hence

$$g(\sigma_j(x), F(\sigma_j(x)), 0) > 0, \quad j = 1, 2, \tag{14.54}$$

where $g(x, y, \lambda)$ appears in the second equation of (14.33). By formula (14.21)

$$\tilde{I}(s, \lambda) = \int_0^s [h(\sigma_1(x)) - h(x)]|_{x=F^{-1}(y)} \mathrm{d}y + \int_s^{Y_2} [h(\sigma_2(x)) - h(x)]|_{x=F^{-1}(y)} \mathrm{d}y,$$

where $x = F^{-1}(y) \in (\hat{x}, x_1)$ is the inverse function of $y = F(x) \in (Y_1, Y_2)$, note that $s = 0$ corresponds to $y = Y_1$. Hence

$$\frac{\mathrm{d}\tilde{I}}{\mathrm{d}s}(s, \lambda) = h(\sigma_1(x)) - h(\sigma_2(x))|_{x=F^{-1}(s)}. \tag{14.55}$$

By (14.19)

$$h(\sigma_j(x)) = \frac{F'(\sigma_j(x))}{g(\sigma_j(x), F(\sigma_j(x)), 0)}, \quad j = 1, 2,$$

and it is obvious $F'(\sigma_1(x)) > 0$ and $F'(\sigma_2(x)) < 0$. Hence from (14.54) and (14.55), we obtain $\frac{\mathrm{d}\tilde{I}}{\mathrm{d}s}(s, \lambda) > 0$ for any $s \in (0, Y_2)$ and $\lambda \in V_2$.

Thus, the proof of Theorem 3.1 is finished.

Acknowledgements The author appreciates the reviewer of this paper for the valuable comments that help him to improve the earlier version of the manuscript. This work is partially supported by grant NSFC-11271027.

References

1. Albrecht, F., Gatzke, H., Wax, N.: Stable limit cycles in prey-predator populations. Science **181**, 1073–1074 (1973)
2. Andrews, J.F.: A mathematical model for the continuous culture of microorganisms utilizing inhibitory substrates. Biotechnol. Bioeng. **10**, 707–723 (1968)
3. Bazykin, A.D.: Nonlinear Dynamics of Interacting Populations. World Scientific Series on Nonlinear Science Series A, vol. 11. World Scientific, Singapore/New Jersey/London/Hong Kong (2000)
4. Broer, H.W., Naudot, V., Roussarie, R., Saleh, K.: A predator-prey model with non-monotonic response function. Regul. Chaotic Dyn. **11**, 155–165 (2006)
5. Broer, H.W., Saleh, K., Naudot, V., Roussarie, R.: Dynamics of a predator-prey model with non-monotonic response function. Discrete Continuous Dyn. Syst. **18**, 221–251 (2007)
6. Chen, L., Jing, Z.: Existence and uniqueness of limit cycles of differential equations with predator-prey interactions. Kexue Tongbao **29**(9), 521–523 (1984) (in Chinese)
7. Chen, J., Zhang, H.: The qualitative analysis of two species predator-prey model with Holling's type III functional response. Appl. Math. Mech. **7**(1), 73–80 (1986) (in Chinese)
8. Cheng, K.S.: Uniqueness of a limit cycle for a predator-prey system. SIAM J. Math. Anal. **12**(4), 541–548 (1981)
9. Collings, J.B.: The effects of the functional response on the bifurcation behavior of a mite predator-prey interaction model. J. Math. Biol. **36**, 149–168 (1997)
10. De Maesschalck, P., Dumortier, F.: Canard solutions at non-generic turning points. Trans. Am. Math. Soc. **358**(5), 2291–2334 (2006)
11. De Maesschalck, P., Dumortier, F.: Canard cycles in the presence of slow dynamics with singularities. Proc. R. Soc. Edinb. A **138**(2), 265–299 (2008)

12. De Maesschalck, P., Dumortier, F., Roussarie, R.: Cyclicity of common slow-fast cycles. Indag. Math. **22**, 165–206 (2011)

13. De Maesschalck, P., Dumortier, F., Roussarie, R.: Canard cycle transition at a slow-fast passage through a jump point. C. R. Math. Acad. Sci. Paris **352**(4), 317–320 (2014)

14. Dumortier, F., Roussarie, R.: Canard cycles and center manifolds. Mem. Am. Math. Soc. **121**(577), 1–100 (1996)

15. Dumortier, F., Roussarie, R.: Multiple canard cycles in generalized Liénard equations. J. Differ. Equ. **174**, 1–29 (2001)

16. Dumortier, F., Roussarie, R.: Birth of canard cycles. Discrete Continuous Dyn. Syst. Ser. S **2**, 723–781 (2009)

17. Fenichel, N.: Geometric singular perturbation theory. J. Differ. Equ. **31**, 53–98 (1979)

18. Freedman, H.I., Wolkowicz, G.S.K.: Predator-prey systems with group defence: the paradox of Bull. Math. Biol. **48**(5–6), 493–508 (1986)

19. Holling, C.S.: The components of predation as revealed by a study of small-mammal predation of the European pine sawfly. Can. Entomol. **91**, 293–320 (1959)

20. Holling, C.S.: The functional response of predators to prey density and its role in mimicry and population regulation. Mem. Entomol. Soc. Can. **45**, 3–60 (1965)

21. Krupa, M., Szmolyan, P.: Extending singular perturbation theory to non-hyperbolic points–fold and canard points in two dimensions. SIAM J. Math. Anal. **33**, 286–314 (2001)

22. Krupa, M., Szmolyan, P.: Relaxation oscillation and canard explosion. J. Differ. Equ. **174**, 312–368 (2001)

23. Lamontagne, Y., Coutu, C., Rousseau, C.: Bifurcation analysis of a predator-prey system with generalized Holling type III function response. J. Dyn. Differ. Equ. **20**, 535–571 (2008)

24. Li, C., Lu, K.: Slow divergence integral and its application to classical Liénard equations of degree 5. J. Differ. Equ. **257**, 4437–4469 (2014)

25. Li, C., Zhu, H.: Canard limit cycles for predator-prey systems with Holling types of functional response. J. Differ. Equ. **254**, 879–910 (2013)

26. Li, C, Li, J., Ma, Z., Zhu, H.: Canard phenomenon for an SIS epidemic model with nonlinear incidence. J. Math. Anal. Appl. **420**, 987–1004 (2014)

27. May, R.M.: Limit cycles in predator-prey communities. Science **17**, 900–902 (1972)

28. Rosenzweig, M.L.: Paradox of enrichment: destabilization of exploitation ecosystems in ecological time. Science **171**, 385–387 (1971)

29. Rothe, F., Shafer, D.S.: Multiple bifurcation in a predator-prey system with nonmonotonic predator response. Proc. R. Soc. Edinb. Sect. A **120**(3–4), 313–347 (1992)

30. Ruan, S., Xiao, D.: Global analysis in a predator-prey system with nonmonotonic functional response. SIAM J. Appl. Math. **61**(4), 1445–1472 (2001)

31. Wolkowicz, G.S.K.: Bifurcation analysis of a predator-prey system involving group defence. SIAM J. Appl. Math. **48**(3), 592–606 (1988)

32. Wrzosek, D.: Limit cycles in predator-prey models. Math. Biosci. **98**(1), 1–12 (1990)

33. Xiao, D., Zhu, H.: Multiple focus and Hopf bifurcations in a predator-prey system with nonmonotonic functional response. SIAM J. Appl. Math. **66**(3), 802–819 (2006)

34. Zhu, H., Campbell, S.A., Wolkowicz, G.S.K.: Bifurcation analysis of a predator-prey system with nonmonotonic function response. SIAM J. Appl. Math. **63**(2), 636–682 (2002)

Chapter 15
Abelian Integrals: From the Tangential 16th Hilbert Problem to the Spherical Pendulum

Pavao Mardešić, Dominique Sugny, and Léo Van Damme

Dedicated to Christiane Rousseau for her mathematical and non-mathematical impact

Abstract In this chapter we deal with abelian integrals. They play a key role in the infinitesimal version of the 16th Hilbert problem. Recall that 16th Hilbert problem and its ramifications is one of the principal research subject of Christiane Rousseau and of the first author. We recall briefly the definition and explain the role of abelian integrals in 16th Hilbert problem. We also give a simple well-known proof of a property of abelian integrals. The reason for presenting it here is that it serves as a model for more complicated and more original treatment of abelian integrals in the study of Hamiltonian monodromy of fully integrable systems, which is the main subject of this chapter. We treat in particular the simplest example presenting non-trivial Hamiltonian monodromy: the spherical pendulum.

Keywords Abelian integrals • Gauss-Mannin monodromy • 16th Hilbert problem • Hamiltonian monodromy

P. Mardešić (✉)
Institut de Mathématiques de Bourgogne, Université de Bourgogne, UMR CNRS 5584, 9 Avenue Alain Savary, BP 47 870, F-21078 Dijon Cedex, France
e-mail: pavao.mardesic@u-bourgogne.fr

D. Sugny • L. Van Damme
Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS-Université de Bourgogne, 9 Avenue Alain Savary, BP 47 870, F-21078 Dijon Cedex, France

Institute for Advanced Study, Technische Universität München, Lichtenbergstrasse 2 a, 85748 Garching, Germany
e-mail: dominique.sugny@u-bourgogne.fr

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_15

15.1 Infinitesimal 16th Hilbert Problem

15.1.1 Abelian Integrals

Consider a polynomial vector field $X = P(x,y)\frac{\partial}{\partial x} + Q(x,y)\frac{\partial}{\partial y}$ in the real plane, $P, Q \in \mathbb{R}[x,y]$. Its trajectories are solutions of the system of differential equations

$$\begin{aligned} \dot{x} &= P(x,y) \\ \dot{y} &= Q(x,y). \end{aligned} \tag{15.1}$$

Its trajectories are curves in the real plane. We will be interested in periodic trajectories. A periodic trajectory is isolated if it is not approached by a continuous family of periodic orbits. In that case, we say that it is a *limit cycle*.

The famous *16th Hilbert problem* asks to give a bound $H(n)$ for the number of limit cycles in the polynomial system (15.1) in function of the degree of the system. Very recently a full solution of this problem is announced in a preprint by Llibre and Pedregal [29], but the paper still has to undergo serious verification by the mathematical community. After this chapter was submitted the authors of [29] recognized a mistake in their proof.

A vast project for proving the existence of such a bound $H(n)$ for $n = 2$ has been launched in [14] by Dumortier, Roussarie, and Rousseau.

The difficulty of this problem led to many weakened versions of this problem. In particular the *infinitesimal version of the 16th Hilbert problem* formulated by Arnol'd [3].

Given a polynomial function $F(x,y) \in \mathbb{R}[x,y]$, its associated *Hamiltonian vector field* X_F is given by

$$X_F = -\frac{\partial F}{\partial y}\frac{\partial}{\partial x} + \frac{\partial F}{\partial x}\frac{\partial}{\partial y}. \tag{15.2}$$

Its trajectories belong to level curves of F. If some are periodic, they are non-isolated. Hence, there are no limit cycles in the system (15.2).

Consider now a small polynomial deformation

$$X_F + \epsilon Y, \tag{15.3}$$

of the Hamiltonian vector field, where Y is a polynomial vector field and $\epsilon \in \mathbb{R}$ a small parameter. Normally, the majority of periodic trajectories will be broken, for $\epsilon \neq 0$ small. However, some periodic trajectories can survive and become isolated. This means that limit cycles can be created. Infinitesimal 16th Hilbert problem asks for a bound $h(n)$ on the number of such limit cycles created in small deformations (15.3) in function of the degree n of the deformation.

Assume that there exists a family $\gamma(h) \subset F^{-1}(h)$ of periodic orbits of the Hamiltonian vector field X_F. In order to examine if some of these orbits *survive* in the deformation (15.3), we take a transversal segment Σ to the family $\gamma(h)$ parametrized by the values h of F. We consider the first return (or Poincaré) map $P_\epsilon(h)$ of the deformed family. Periodic orbits correspond to fixed points of the Poincaré map. Limit cycles correspond to isolated fixed points of the Poincaré map.

Given a vector field $Y = A(x,y)\frac{\partial}{\partial x} + B(x,y)\frac{\partial}{\partial y}$, its dual one-form ω_Y is the form $\omega_Y = B(x,y)dx - A(x,y)dy$. In particular, one-form dual to a Hamiltonian vector field X_F is the exact form $dF = \frac{\partial F}{dx}dx + \frac{\partial F}{dy}dy$. The Poincaré map of (15.3) starts with the identity. The principal part of the Poincaré map minus identity is well known since Poincaré and Pontryaguin. Indeed,

$$P_\epsilon(h) = h - \epsilon I(h) + o(\epsilon), \tag{15.4}$$

where

$$I(h) = \int_{\gamma(h)} \omega_Y. \tag{15.5}$$

Here $o(\epsilon)$ denotes a function depending on ϵ and h, such that $\frac{o(\epsilon)}{\epsilon}$ tends to zero, for ϵ tending to zero. The proof is easy. Let $\gamma_\epsilon(h)$ be the deformed (not necessarily closed) trajectory of (15.3) starting at $S \cap F^{-1}(h)$. This means that

$$\int_{\gamma_\epsilon(h)} dF + \epsilon\omega_Y \equiv 0.$$

By the definition of Poincaré map, it then follows that

$$P_\epsilon(h) - h = -\epsilon \int_{\gamma_\epsilon(h)} \omega_Y = -\epsilon \int_{\gamma(h)} + o(\epsilon).$$

The function $I(h)$ is an *abelian integral*, meaning the integral of a polynomial (or rational) differential one-form along a one-cycle belonging to a solution of a polynomial (or rational) equation. We call *displacement function* Δ_ϵ the difference between the Poincaré map and identity. In the region where the Poincaré map is differentiable, so implicit function theorem can be applied, simple zeros of the abelian integral give rise to simple zeros of the displacement function Δ_ϵ.

For that reason, the infinitesimal 16th Hilbert problem is closely related to the *tangential 16th Hilbert problem*, asking for the bound $A(n)$ for the number of zeros of abelian integrals associated to degree n deformations (15.3).

The existence of such a bound $A(n)$ has been proved for any n by Khovanskii [27] and Varchenko [34] and improved by Binyamini et al. [8]. It is trivial that $A(n) = 0$, for $n = 0, 1$. A highly non-trivial result that $A(2) = 3$ was proved in a sequence of papers considering different cases. Let us cite here only the paper of Gavrilov [22], treating the generic cases. An explicit bound is not known for any other value of n.

Let us note that the two versions (infinitesimal and tangential) of the 16th Hilbert problem are not equivalent precisely because the Poincaré map is not differentiable in a neighborhood of a polycycle (bounding a family of periodic orbits). It has been proved by Dumortier and Roussarie in [13] that there exist zeros of the displacement function (and limit cycles) of a small deformation (15.3) of a Hamiltonian system, which do not correspond to zeros of the abelian integrals. Dumortier and Roussarie call these cycles *alien cycles*. Also the study of the points of loss of differentiability of abelian integrals gives the clue to the study of their zeros (see, e.g., [26]).

15.1.2 Monodromy of Abelian Integrals

We give here a simple result, whose method of proof will be used also in the second part of this chapter.

We say that a critical point of F is a *Morse critical point*, if by a local analytic complex change of coordinates in the (x, y)-plane, the function F can be brought to the form $u^2 + v^2$. If we permit only real changes of coordinates, a Morse critical point can be brought to the form $u^2 + v^2$ (center type) of $u^2 - v^2$ (saddle type).

Proposition 1.1. *Consider a center c_0 of a Hamiltonian vector field X_F. Let $\gamma(h)$ be the family of periodic orbits of X_F surrounding the center c_0. Suppose that the basin of the center is bounded by a* polycycle *formed of saddle-points s_j, $j = 1, \ldots, k$, and their separatrices (see Fig. 15.1). We can suppose that $F(c_0) = 0$ and $F(p_j) = h_0$.*

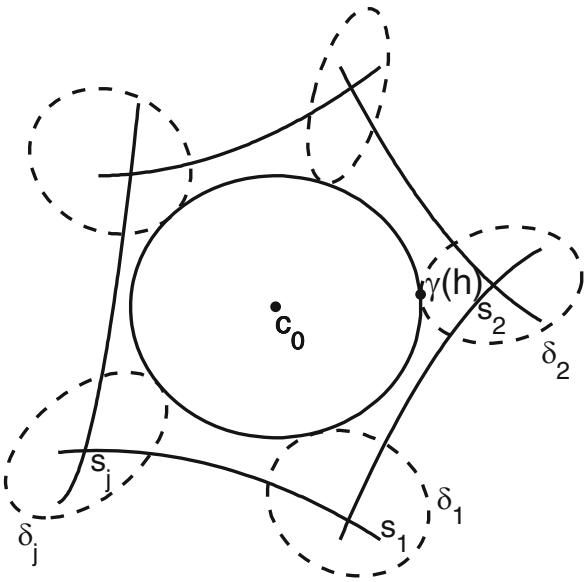


Fig. 15.1 Cycle $\gamma(h)$ surrounding a center c_0 near a polycycle and *complex* cycles δ_j surrounding saddles s_j

Consider the abelian integral $I(h)$ (15.5) associated to the deformation (15.3). Then, the abelian integral $I(h)$ can be written in the form

$$I(h) = f(h) + \log(h - h_0)g(h), \tag{15.6}$$

where $f(h)$ and $g(h)$ are analytic functions in a neighborhood of h_0. Moreover, $g(h_0) = 0$.

The proof of the proposition is geometric and is essentially the *Picard–Lefschetz formula*. We present the proof here, because it will serve as the model for the proofs of results about Hamiltonian monodromy and in particular the study of the spherical pendulum.

Let us first consider a special situation. Consider the polynomial function $f(x, y) = x^2 + y^2$ in $\mathbb{C}[x, y]$. For each $t \in \mathbb{C} \setminus \{0\}$, the Riemann surface $f^{-1}(t)$ is a tube (i.e., a sphere from which two points have been deleted). This can be seen from the equation $y = \sqrt{t - x^2}$ in the following way. If $t - x^2 \neq 0$, then there are two solutions for y (see Fig. 15.2). One should imagine that there is a cut along the curve connecting the two ramification points. Moreover, the rear side of the upper slit is glued with the front side of the lower slit and the front side of the upper slit is glued with the rear side of the lower slit. This gives us a tube, i.e., a sphere from which two points (corresponding to infinity in upper and lower branch) are deleted. This can be visualized by flipping the lower leaf before glueing.

For each one of the two points $x_j = \sqrt{t}, j = 1, 2$, there is only one corresponding value y. These two points are called ramification points, since in them the two branches of the Riemann surface meet together. Note that when performing a full turn in the x-plane, the two solutions y exchange. This is so, because any solution is multiplied by a square root of $e^{2\pi i}$, which is -1. We can identify a cycle δ_0 going around the throat of the tube. If t goes to zero, then the two ramification points merge. The *cycle δ_0 vanishes* in a singular point.

Consider now a *relative cycle* connecting the two points at infinity on the tube for $t \neq 0$. It starts on one branch and we can choose it to pass through one of the ramification points. We can follow the cycles for nearby values of t, just slightly modifying the ramification points. What happens with this cycle if the value t performs a full turn around the origin? Well, then each of the two ramification points $x_1(t)$ and $x_2(t)$ will perform half a turn and ultimately the two points will be exchanged. Looking carefully, we see that the cycle δ_1 has been transformed to itself plus the vanishing cycle δ_0 (see Fig. 15.3).

In fact, the whole phenomenon is purely local and everything happens exactly in the same way for any cycle in a neighborhood of a Morse singular point. More precisely, let δ_0 be a vanishing cycle at a Morse singular point. A relative cycle δ_1 defined in a neighborhood of the Morse singular point will be transformed by monodromy around the critical value of the Morse point to

$$M(\delta_1) = \delta_1 + (\delta_1, \delta_0)\delta_0, \tag{15.7}$$

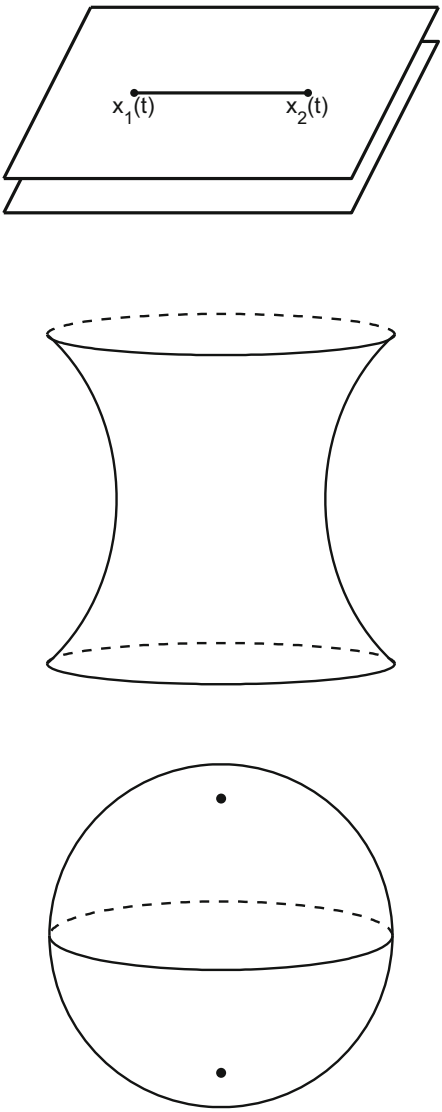


Fig. 15.2 Regular fiber $f^{-1}(t)$, for $f(x, y) = x^2 + y^2$

where (δ_1, δ_0) is the intersection number between the two cycles. This is the famous Picard–Lefschetz formula describing the action of the *Gauss–Manin monodromy M* on cycles.

Proof of the Proposition. Note that in the abelian integral the form we integrate is univalued. Hence, all the multivaluedness of the abelian integral $I(h) = \int_{\delta(h)} \omega$ comes from the multivaluedness of the cycles which we just studied. Let us complexify and follow the cycle of integration $\delta(h)$ as h performs a full turn around $h = h_0$. By the Picard–Lefschetz formula, the cycle will be deformed to itself plus a

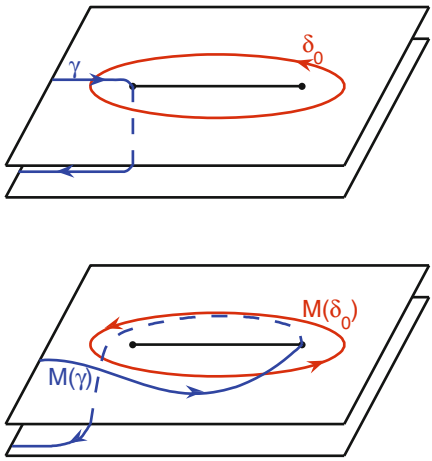


Fig. 15.3 Monodromy $M(\gamma)$ of a relative cycle γ in a neighborhood of a Morse singular point

sum of one vanishing cycle $\delta_j(h)$, for each (Morse) singular point p_j in the boundary of the polycycle. From the univaluedness of the form ω it now follows:

$$M(I)(h) = I(h) + \sum_{j=1}^{k} \int_{\delta_j(h)} \omega. \tag{15.8}$$

Note that each one of the integral functions $\int_{\delta_j(h)} \omega$ is univalued in a neighborhood of $h = h_0$ and moreover tends to zero for $h = h_0$, as the cycle $\delta_j(h)$ vanishes to the singular point p_j. Consider now the monodromy of the function

$$I(h) - \sum_{j=1}^{k} \frac{\log(h - h_0)}{2\pi i} \int_{\delta_j(h)} \omega \tag{15.9}$$

around $h = h_0$. The function is univalued, since the multivaluedness of the two terms cancels. It is not difficult to see that $h = h_0$ is a removable singularity of an analytic function $f(h)$. Putting

$$g(h) = \sum_{j=1}^{k} \frac{1}{2\pi i} \int_{\delta_j(h)} \omega,$$

which is an analytic function vanishing at $h = h_0$, proves the proposition. □

15.2 Hamiltonian Monodromy of the Spherical Pendulum

15.2.1 Introduction

Hamiltonian monodromy is nowadays a well-recognized topological property of Hamiltonian integrable systems [9] both in mathematics [10] and in physics [15]. For a dynamical system with a finite number of degrees of freedom, Hamiltonian monodromy is the simplest topological obstruction to the existence of global action-angle variables [12, 30]. A comprehensive introduction can be found in standard textbooks such as [10, 15]. To be more precise, let us consider an integrable system with two degrees of freedom defined by an energy-momentum map. Let $\mathcal{R}$ be the set of regular values of the image of this map. From the Liouville–Arnold theorem [2], the preimage of a point of $\mathcal{R}$, which is assumed to be compact, is a torus. We define the monodromy map by associating to a loop of $\mathcal{R}$ a monodromy matrix characterizing the rotation of the action-angle coordinates along this loop. If this 2 by 2 matrix is different from the identity then the monodromy is said to be non-trivial and it is not possible to define global action-angle coordinates. Due to its topological character, monodromy can be non-trivial if $\mathcal{R}$ is not simply connected. The simplest example of non-trivial monodromy is given by an energy-momentum map with an isolated singularity. The preimage of this singular point is a single or a multiple pinched torus [10, 15]. Hamiltonian monodromy has also some profound implications in quantum mechanics since it is a topological obstruction to the existence of global quantum numbers [36]. The semi-classical limit allows to establish a clear relationship between classical and quantum monodromy [1, 25, 36]. From a historical point of view, non-trivial Hamiltonian monodromy has been first exhibited in the case of a spherical pendulum [12], but is now found in a variety of physical systems [5, 11, 17, 19, 21, 24, 37]. The notion of Hamiltonian monodromy has known recently important developments with the introduction of fractional monodromy [16, 20, 23, 31] and bidromy [18, 32] phenomena. In these generalized versions of Hamiltonian monodromy, we consider loops of the energy-momentum diagram which can cross lines of weak singular values of $\mathcal{R}$. In the case of fractional monodromy, this line corresponds to a line a curled tori. For the bidromy phenomenon, this line is a set of points which lift in the original phase space to bitori. In this case, the procedure is more complex since a bipath formed of two closed paths has to be considered. This specificity is related to the two leaves of $\mathcal{R}$ which is characteristic of the bidromy topological character. In particular, when the path crosses transversally the line of bitori, each of the two paths lies in a different leaf.

Hamiltonian monodromy has been described from a real point of view in these preceding works. Some papers have tried in the past few years to describe this concept from complex geometry [4, 28, 38] and the Picard–Lefschetz theory [6, 7, 35]. Two different approaches can be roughly considered. The first one is based on a complexification of the initial Hamiltonian system and on the computation of the monodromy matrix of the complexified system. This method is in general difficult to apply due to the dimension of the dynamical complex system obtained.

In addition, all the information achieved regarding the monodromy of the homology group of the complex torus is not relevant from a real point of view. In a second approach that some of the authors introduced in [33], we assume that the momentum map of the energy-momentum diagram is defined through a global S^1-action. A quotient of the initial phase space by this action leads to the reduced phase space whose coordinates are the invariant polynomials [10]. The idea is then to complexify only the reduced phase space and to use this complexification not to compute a complex monodromy but the real one. In other words, this complexification can be viewed here as a new way to define Hamiltonian monodromy and to compute the corresponding monodromy matrix. This method has been introduced in [33] and used to define and compute fractional monodromy for $m : -n$ resonant systems. After this first initial attempt, it is clear that a lot of work remains to be done in this direction in order to construct a robust mathematical framework for this approach and to understand its limit and its domain of application. One open question is, for instance, its relation with the other complex approach. In this context, one of the goals of this chapter is to give a complete overview of this complex method by considering the historical problem for which Hamiltonian monodromy has been computed for the first time, that is the spherical pendulum.

The organization of this second part of the chapter is as follows. We first consider the spherical pendulum in the real case in Sect. 15.2 and we recall the standard way to obtain the monodromy matrix. We use this example to demonstrate the efficiency of our complex approach in Sect. 15.2.3. Concluding remarks and prospective views are given in Sect. 15.2.4.

15.2.2 Hamiltonian Monodromy of the Spherical Pendulum in the Real Case

15.2.2.1 The Spherical Pendulum

A spherical pendulum is a mechanical system which consists in a mass moving without friction on a sphere. The mass is only subject to a constant gravity field along the z-direction [10]. The Hamiltonian of the system can be written on the tangent space $T\mathbb{R}^3$ as:

$$H = \frac{1}{2}(p_x^2 + p_y^2 + p_z^2) + z, \tag{15.10}$$

where (x, y, z, p_x, p_y, p_z) are coordinates of $T\mathbb{R}^3$. The pendulum is constrained to move on a sphere, i.e., such that $x^2 + y^2 + z^2 = 1$ and $xp_x + yp_y + zp_z = 0$. The phase space of the system is the tangent bundle TS^2 of the sphere. The Hamiltonian H is Liouville integrable since it Poisson commutes with the momentum J given by:

$$J = xp_y - yp_x, \tag{15.11}$$

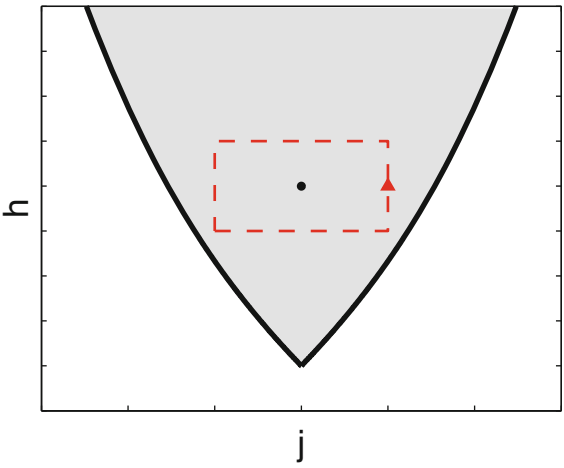


Fig. 15.4 Energy-momentum diagram of the spherical pendulum. The parameters (h, j) represent, respectively, the values of the Hamiltonian H and of the momentum J. The *black dot* displays the position of the singular point. The region in *gray* corresponds to the regular values of the energy-momentum map $\mathcal{EM}$. A loop Γ is depicted in *dashed lines*

which is the z-component of the kinetic momentum. The topological properties of this Hamiltonian can be investigated by introducing the following energy-momentum map:

$$\mathcal{EM} : \mathbf{z} \in TS^2 \rightarrow (H(\mathbf{z}), J(\mathbf{z})) \in \mathbb{R}^2. \tag{15.12}$$

The image of the energy-momentum map called the bifurcation diagram or the energy-momentum diagram is displayed in Fig. 15.4. The energy-momentum map has only one non-trivial singular point where the differentials dH and dJ are not linearly independent. This point of coordinates $(j = 0, h = 1)$ lifts in the original phase space to a two-dimensional pinched torus [15]. The topology of the singular torus can be determined from the singular reduction techniques [10].

We consider a singular reduction with respect to the S^1-action of J in order to decrease the dimension of the problem. We introduce the algebra of invariant polynomials which is generated by the following six polynomials [10]:

$$\sigma_1 = z;\ \sigma_2 = p_z;\ \sigma_3 = p_x^2 + p_y^2 + p_z^2;$$
$$\sigma_4 = xp_x + yp_y;\ \sigma_5 = x^2 + y^2;\ \sigma_6 = xp_y - yp_x.$$

These polynomials form a basis of the polynomials which Poisson-commute with J. In particular, note that $J = \sigma_6$ and $H = \sigma_3/2 + \sigma_1$, which shows that $\{J, H\} = 0$. The generators σ_i satisfy the relations:

$$\sigma_3(1 - \sigma_1^2) = \sigma_2^2 + j^2, \tag{15.13}$$

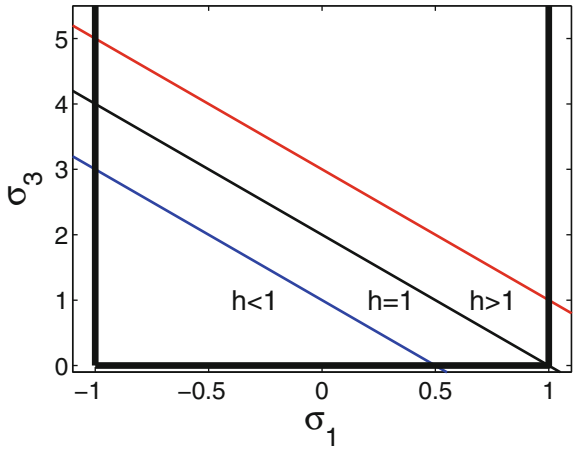


Fig. 15.5 Intersections of the reduced phase space M_0 (in *large solid lines*) with the level sets $H_0 = h$ with $h > 1$, $h = 1$ and $h < 1$

for a given value j of J and the constraints $\sigma_1^2 \leq 1$ and $\sigma_3 \geq \sigma_2^2$. The reduced phase space is a semialgebraic variety defined by $M_j = J^{-1}(j)/S^1$ and this variety is explicitly given in the space $\mathbb{R}^3 = (\sigma_1, \sigma_2, \sigma_3)$ by Eq. (15.13). The topology of a given torus characterized by a regular value of (h, j) can be obtained by considering the intersection of the reduced phase space M_j with the level set $H_j = h$. Here, we introduce the reduced Hamiltonian H_j, which is a map from M_j to $\mathbb{R}$ sending a point of M_j to $H(\sigma_1, \sigma_3)$. In a regular case, this intersection is a circle which lifts in the original phase space to a two-dimensional torus. However, the flow of J defines an S^1-action on the phase space, but this action is not regular since the points $(0, 0, \pm 1, 0, 0, 0)$ are fixed. For such points $j = 0$ and the corresponding reduced phase space M_0 has two singular points at $(\pm 1, 0, 0)$. The projections of the reduced phase space M_0 and of the level set $H = h$ onto the plane (σ_1, σ_3) are displayed in Fig. 15.5. Figure 15.5 shows that one of the singular points of the reduced phase space belongs to the level set $H_0 = 1$. Since this point does not lift to a circle but to a point in the initial phase space, we deduce that the singular torus is pinched in this point.

15.2.2.2 The Monodromy Matrix

In this paragraph, we recall some basic facts about the monodromy phenomenon and the computation of the monodromy matrix. The first step consists in defining a 2-torus bundle over the regular values of the image of the energy-momentum map, denoted $\mathcal{R}$. This bundle is locally trivial and the monodromy is the simplest obstruction for it to be globally trivial [12]. An explicit construction of the monodromy matrix can be made as follows. We consider a loop Γ along $\mathcal{R}$ and we fix a point of this loop. For this point (h, j) of Γ, we define a basis of the homology group $H_1(T^2(h, j), \mathbb{Z})$. Deforming continuously this basis along Γ, it may have

changed when we come back to the original point, which leads to the monodromy matrix. Note that this matrix only depends on the homotopy class of the loop Γ. There exists a natural and straightforward way to define the basis of the homology group by using the flow of the vector fields associated with H and J. The momentum J generates an S^1-action on the regular torus. Let θ be an angle conjugated to J. The flow of H from a point of this circle defines an orbit which intersects at a time T the flow of J. The two points of intersection of the two flows define two angles θ_i and θ_f and a twist $\Theta = \theta_f - \theta_i$. It can be shown that the two functions $T(h,j)$ and $\Theta(h,j)$ allow us to completely reconstruct the basis of the homology group [10, 15]. The two cycles of the basis are associated, respectively, with the flow of the vector fields $X_1 = 2\pi X_J$ and $X_2 = -\Theta(h,j)X_j + T(h,j)X_h$. In addition, the monodromy matrix is related to the behavior of the functions Θ and T along the considered loop of $\mathcal{R}$.

A standard result in this direction is the fact that after a counterclockwise loop around an isolated singular point (which lifts to a singular pinched torus), the rotation angle Θ increases by 2π, while T remains unchanged. The corresponding monodromy matrix M is then:

$$M = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}.$$

We finish this section with the expressions of Θ and T as a function of the invariant polynomials. We refer the interested reader to [15] for the explicit derivation of these relations. Let (h,j) be a regular value of the energy-momentum diagram, the functions Θ and T can be expressed in the invariant polynomials basis as follows:

$$\Theta = 2j \int_{\sigma_1^-}^{\sigma_1^+} \frac{d\sigma_1}{(1-\sigma_1^2)Q(\sigma_1)}, \tag{15.14}$$

and

$$T(h,j) = 2 \int_{\sigma_1^-}^{\sigma_1^+} \frac{d\sigma_1}{Q(\sigma_1)}, \tag{15.15}$$

where

$$Q(\sigma_1) = \sqrt{2(h-\sigma_1)(1-\sigma_1^2) - j^2}. \tag{15.16}$$

$\sigma_1^{\pm}$ are the two roots of Q in the interval $[-1,1]$ (see below for details).

15.2.3 Hamiltonian Monodromy of the Spherical Pendulum in the Complex Case

15.2.3.1 Complexification of the Spherical Pendulum

The goal of this section is to apply the complex approach introduced in [33] to the spherical pendulum. We complexify the coordinates of the reduced phase space $(\sigma_1, \sigma_2, \sigma_3, j)$ and we set $y = \sigma_2$ and $x = \sigma_1$. Using the fact that $h = \sigma_3/2 + x$, the relation (15.13) becomes

$$y^2 = 2(h - x)(1 - x^2) - j^2, \tag{15.17}$$

which defines a Riemann surface for fixed values of h and j, similarly as in Sect. 15.1. This surface has three ramification points, denoted x_0, x_+, and x_-. Note that $x_\pm$ are the complexifications of the roots $\sigma_1^\pm$ of Q. The Riemann surface defined by (15.17) is schematically represented in Fig. 15.9. In the reduced phase space M_j, the original torus projects to a cycle $\delta(h,j)$, $(h,j) \in \mathcal{R}$, which is delimited by σ_1^- and σ_1^+, the two roots of Q in the interval $[-1, 1]$. Returning back to the Riemann surface, the cycle $\delta(h,j)$ is represented by a real oval from x_- to x_+. To be coherent with the real approach, we assume that the cycle is oriented from x_- to x_+ in the upper leaf and from x_+ to x_- in the lower one. All these notations are displayed in Fig. 15.9.

We pursue the complex approach by introducing the complex continuation of the functions Θ and T, which are now interpreted as abelian integrals over the loop δ of the Riemann surface:

$$\Theta(h,j) = j \int_\gamma \frac{dx}{(1 - x^2)y} \tag{15.18}$$

$$T(h,j) = \int_\gamma \frac{dx}{y}. \tag{15.19}$$

Note that the positive and the negative determinations of the square root y have been chosen, respectively, for the upper and the lower leaves of the Riemann surface. We have also added a factor $1/2$ in the definition of the two integrals to coincide with the real case.

Let Γ be a small real loop around the singular point $(h = 1, j = 0)$. In the real case, we recall that the computation of the monodromy matrix associated with Γ is based on the variation of the function Θ along this loop, the variation of the function T being equal to zero. This variation is denoted $\Delta_\Gamma \Theta$. The last step of the method consists in computing $\Delta_\Gamma \Theta$ by using the extension of Θ to the complex domain. For that purpose, we introduce a Gauss–Manin connection along the loop Γ [38] and we use this connection to make the parallel transport of δ along Γ. This will lead to the monodromy of δ and to the variation of Θ. The Gauss–Manin connection is determined by following the motion of the ramification points of the Riemann surface. This motion is analyzed in the next section.

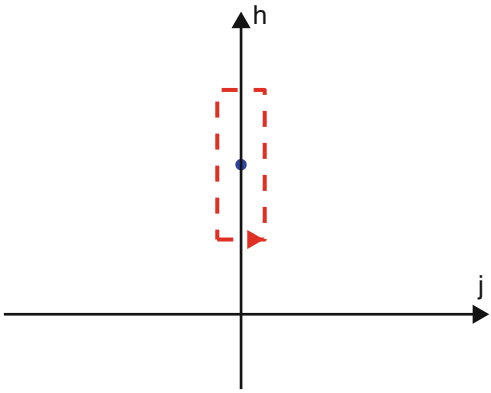


Fig. 15.6 Local bifurcation diagram in the neighborhood of the singular value $(j = 0, h = 1)$. This point is represented by a full *dot*. The *dashed line* depicts the loop Γ used to calculate the Gauss–Manin monodromy

15.2.3.2 Local Computation of the Ramification Points

The determination of the Gauss–Manin monodromy requires the knowledge of the evolution of the ramification points of the Riemann surface along the loop Γ. To simplify the computation, we consider a small loop around the singular value $(j = 0, h = 1)$.

For this singular value, the polynomial Q has three roots $x_- = -1$ and $x_0 = x_+ = 1$. Let δh, δj, and δx be the small variations with respect to the energy, momentum, and roots, respectively. A first numerical examination of the roots shows that they collide with the poles 1 and -1 of the integral Θ on the vertical line of equation $j = 0$. We have therefore to determine their behavior around this line. To simplify the computation, we will consider a loop such that $1 >> \delta h >> \delta j$. This loop is schematically represented in Fig. 15.6.

We first consider the case where $h = 1 + \delta h$ and $x = 1 + \delta x$. The roots of the polynomial Q satisfy:

$$2(\delta h - \delta x)(-2\delta x - \delta x^2) - \delta j^2 = 0. \tag{15.20}$$

A direct expansion of the left-hand side leads to:

$$-4\delta h \delta x - 2\delta h \delta x^2 + 4\delta x^2 + 2\delta x^3 - \delta j^2 = 0. \tag{15.21}$$

Using a Newton diagram, we can neglect the terms $\delta h \delta x^2$ and δx^3 with respect to δx^2. One finally arrives to the condition:

$$4\delta x^2 - 4\delta h \delta x - \delta j^2 = 0. \tag{15.22}$$

The two roots are given by the following expressions:

$$x_+ = 1 + \frac{\delta h - \sqrt{\delta h^2 + \delta j^2}}{2}$$

$$x_0 = 1 + \frac{\delta h + \sqrt{\delta h^2 + \delta j^2}}{2},$$

with the convention that $x_+ \leq 1$ and $x_0 \geq 1$. Using the relations $1 \gg \delta h \gg \delta j$, we finally get

$$x_+ = 1 + (\delta h - |\delta h| - \frac{\delta j^2}{2|\delta h|})/2$$
$$x_0 = 1 + (\delta h - |\delta h| + \frac{\delta j^2}{2|\delta h|})/2,$$

which simplifies into:

$$x_+ = 1 - \frac{\delta j^2}{4\delta h}$$
$$x_0 = 1 + \delta h + \frac{\delta j^2}{4|\delta h|},$$

for $\delta h > 0$ and

$$x_+ = 1 - |\delta h| - \frac{\delta j^2}{4|\delta h|}$$
$$x_0 = 1 + \frac{\delta j^2}{4|\delta h|}.$$

for $\delta h < 0$. The same work can be made for the x_- root and we obtain

$$x_- = -1 + \frac{\delta j^2}{8}. \tag{15.23}$$

15.2.3.3 Gauss–Manin Monodromy

In this section, following [33], we reformulate the notion of Hamiltonian monodromy in the complex domain. The starting point of this approach is to interpret the integral expression of the function Θ as a real integral on a Riemann surface defined by (15.18). This allows us to use the topological properties of this Riemann surface along a complex loop in a complexified energy-momentum diagram. This loop is depicted in Fig. 15.7 where two bypasses around the line $j = 0$ can be seen. We denote the two semi-circles by Γ_+ and Γ_-, for $h > 1$ and $h < 1$, respectively. In this work, we have considered deformations of the real loop in the half-plane $\Im[j] > 0$ but they could be equivalently done in the half-plane $\Im[j] < 0$. We then define a Gauss–Manin connection along the loop Γ by following the motion of the ramification points of the Riemann surface. By using the results of Sect. 15.2.3.2,

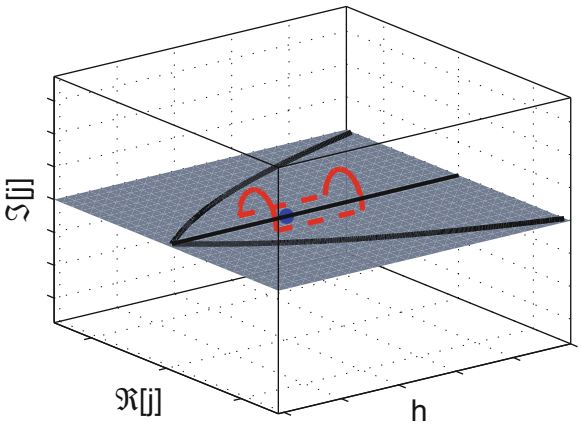


Fig. 15.7 Complex loop used to compute the Gauss–Manin monodromy in the space $(\Re[j], \Im[j], h)$. The *gray plane* indicates the position of the real bifurcation diagram. The two semi-circles of the loop Γ around the line $j = 0$ are in the half-plane $\Im[j] > 0$

it is straightforward to deduce this behavior along the two bypasses. Note that along the rest of the loop, the evolution of the three ramification points is trivial, i.e., we have $-1 < x_- < x_+ < 1$ and $x_0 > 1$. We parameterize the two arcs of circle Γ_+ and Γ_- by $h = h_0$ and $j = j_0 e^{it}$ with $t \in [0, \pi]$ if $h_0 > 1$ and $t \in [\pi, 0]$ for $h_0 < 1$. There is a qualitative difference in the motion of the ramification points according to the value of h_0. Using the asymptotic expressions of Sect. 15.2.3.2, we observe that x_- turns around the pole in -1 for the two bypasses, while x_+ and x_0 turn around the one in $x = 1$ only for $h_0 > 1$ and $h_0 < 1$, respectively. From this information, it is now straightforward to transport the cycle δ along Γ_+ and Γ_-. In accordance with the Picard–Lefschetz theory [38], we see that this cycle is transformed into itself plus some vanishing cycles around the poles in $x = -1$ and $x = 1$. The parallel transport of the cycle δ is summarized in Fig. 15.8 for $h_0 > 1$ and in Fig. 15.9 for $h_0 < 1$.

We have now all the tools in hand to transport the cycle δ along the loop Γ and to deduce the corresponding Gauss–Manin monodromy. We compute this monodromy from a base point (h_0, j_0), with $h_0 > 0$ and $j_0 > 0$, but this computation is independent of the chosen base point. We first observe that the contribution of the two vanishing cycles arising from the motion of x_- cancels each other. We then deduce that:

$$\Delta_\Gamma \Theta = \int_{\delta_0 (h_0 > 1, j_0 > 0)} j_0 \frac{dx}{(1 - x^2) y}, \tag{15.24}$$

where δ_0 is the vanishing cycle around the pole in $x = 1$. This expression can be explicitly computed from a residue:

$$\Delta_\Gamma \Theta = 2 \times 2\pi i j_0 \mathrm{Res}\left(\frac{1}{(1 - x^2) y}, x = 1\right). \tag{15.25}$$

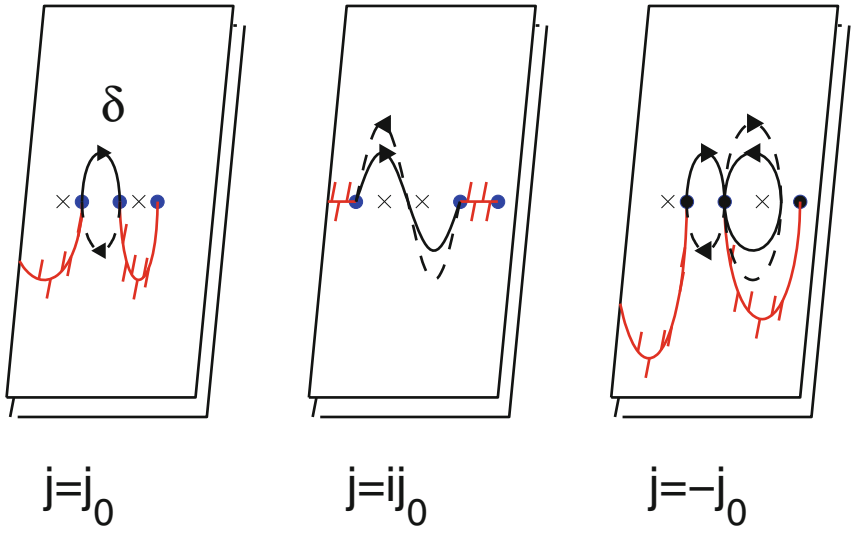


Fig. 15.8 Transport of the cycle δ along the semi-circle Γ_+. The *crosses* indicate the position of the poles in $x = \pm 1$. The *solid lines without arrows* represent arbitrary branch cuts of the Riemann surfaces and the *full dots*, the ramification points (see the text for details). The parts in *solid and dashed lines* of the loop lie, respectively, in the upper and lower leaves of the Riemann surface. For the first figure, the ramification points are from left to right, x_-, x_+, and x_0

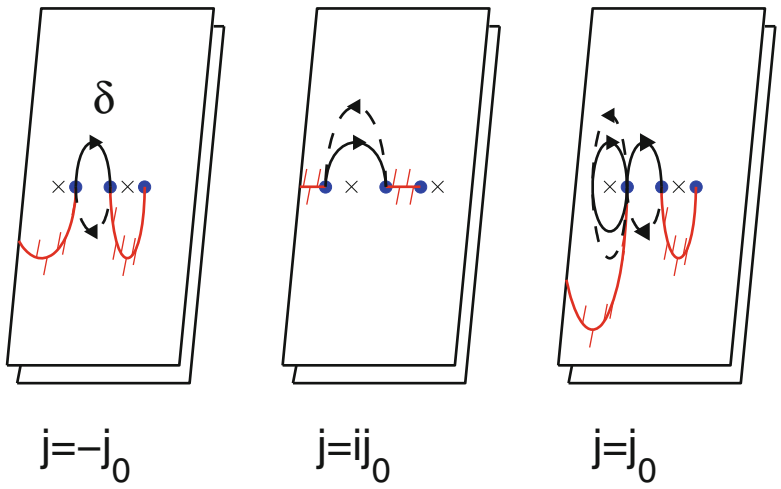


Fig. 15.9 Same as Fig. 15.8 but for the semi-circle Γ_-

Note that the first factor two comes from the fact that the contribution of the two leaves has been added. We finally arrive to:

$$\Delta_\Gamma \Theta = 2\pi, \tag{15.26}$$

which is exactly the result obtained in the real case [10, 12].

15.2.4 Conclusion and Prospective Views

In the second part of this chapter, we have reviewed in the case of a spherical pendulum a complex approach to compute the Hamiltonian monodromy by using a complex extension of the bifurcation diagram. This allows us to compute straightforwardly the monodromy matrix from the determination of a simple residue. We have also given a clear evidence of the relationship that can exist between Hamiltonian monodromy and its complexified counterpart, the Gauss–Manin monodromy. This work can also be viewed as a first step in the use of complex geometry and abelian integrals to solve standard mechanical problems. We hope that the methods presented in this chapter could stimulate some works in this direction both in physics and in mathematics.

Acknowledgements D. Sugny acknowledges support of the Technische Universität München, Institute for Advanced Study, funded by the German Excellence Initiative and the European Union Seventh Framework Programme under grant agreement number 291763. As a Hans Fisher fellow (partially financed through the Marie Curie COFUND program), D. Sugny also acknowledges the support from the European Union.

References

1. Alber, M.S., Marsden, J.E.: Semiclassical monodromy and the spherical pendulum as a complex Hamiltonian system. In: Conservative Systems and Quantum Chaos. Fields Institute Communication, no. 8. AMS and Fields Institute, Calgary (1996)
2. Arnol'd, V.I.: Mathematical Methods of Classical Mechanics. Springer, New York (1989)
3. Arnol'd, V.I.: Arnold's Problems. Springer, Berlin (2004). Translated and revised edition of the 2000 Russian original
4. Arnol'd, V.I., Goussein-Zade, S.M., Varchenko, A.N.: Singularities of Differential Mappings. Birkhauser, Boston (1988)
5. Assémat, E., Efstathiou, K., Joyeux, M., Sugny, D.: Fractional bidromy in the vibrational spectrum of HOCl. Phys. Rev. Lett. **104**, 113002 (2010)
6. Audin, M.: Hamiltonian monodromy via Picard-Lefschetz theory. Commun. Math. Phys. **229**, 459 (2002)
7. Beukers, F., Cushman, R.H.: The complex geometry of the spherical pendulum. Contemp. Math. Celest. Mech. **292**, 47 (1999)

8. Binyamini G., Novikov, D., Yakovenko, S.: On the number of zeros of Abelian integrals: a constructive solution of the infinitesimal Hilbert sixteenth problem. Invent. Math. **181**, 227 (2010)
9. Bolsinov, A.V., Fomenko, A.T.: Integrable Hamiltonian Systems: Geometry, Topology, Classification. Chapman and Hall/CRC, Boca Raton (2004)
10. Cushman, R.H., Bates, L.: Global Aspects of Classical Integrable Systems. Birkhäuser, Basel (1997)
11. Cushman, R.H., Dullin, H.R., Giacobbe, A., Holm, D.D., Joyeux, M., Lynch, P., Sadovskii, D.A., Zhilinskii, B.I.: CO2 molecule as a quantum realization of the 1:1:2 resonant swing-spring with monodromy. Phys. Rev. Lett. **93**, 024302 (2004)
12. Duistermaat, J.J.: On global action-angle coordinates. Commun. Pure Appl. Math. **33**, 687 (1980)
13. Dumortier, F., Roussarie, R.: Abelian integrals and limit cycles. J. Differ. Equ. **227**, 116 (2006)
14. Dumortier, F., Roussarie, R., Rousseau, C.: Hilbert 16-th problem for quadratic vector fields. J. Differ. Equ. **110**, 86 (1994)
15. Efstathiou, K.: Metamorphoses of Hamiltonian Systems with Symmetries. Lecture Notes in Mathematics, vol. 1864. Springer, Berlin (2005)
16. Efstathiou, K., Broer, H.W.: Uncovering fractional monodromy. Commun. Math. Phys. **324**, 549 (2013)
17. Efstathiou, K., Sadovskiĭ, D.A.: Normalization and global analysis of perturbations of the hydrogen atom. Rev. Mod. Phys. **82**, 2099 (2010)
18. Efstathiou, K., Sugny, D.: Integrable Hamiltonian systems with swallowtails. J. Phys. A **43**, 085216 (2010)
19. Efstathiou, K., Joyeux, M., Sadovskiĭ, D.A.: Global bending quantum number and the absence of monodromy in the HCN-CNH molecule. Phys. Rev. A **69**, 032504 (2004)
20. Efstathiou, K., Cushman, R.H., Sadovskiĭ, D.A.: Fractional monodromy in the 1:1:2 resonance. Adv. Math. **20**, 241 (2007)
21. Efstathiou, K., Lukina, O.V., Sadovskiĭ, D.A.: Most typical 1:2 resonant perturbation of the hydrogen atom by weak electric and magnetic fields. Phys. Rev. Lett. **101**, 253003 (2008)
22. Gavrilov, L.: The infinitesimal 16th Hilbert problem in the quadratic case. Invent. Math. **143**, 449 (2001)
23. Giacobbe, A.: Fractional monodromy: parallel transport of homology cycles. Differ. Geom. Appl. **26**, 140 (2008)
24. Grondin, L., Sadovskiĭ, D.A., Zhilinskii, B.I.: Monodromy as topological obstruction to global action-angle variables in systems with coupled angular momenta and rearrangement of bands in quantum spectra. Phys. Rev. A **65**, 012105 (2001)
25. Guillemin, V., Uribe, A.: Monodromy in the quantum spherical pendulum. Commun. Math. Phys. **122**, 563 (1989)
26. Ilyashenko, Y.S.: Appearance of limit cycles by perturbation of the equation $dw/dz = Rz/Rw$, where $R(z, w)$ is a polynomial. Mat. Sbornik (New Series) **78**, 360 (1969)
27. Khovanskii, A.: Real analytic manifolds with the property of finiteness and complex abelian integrals. Funktsional. Anal. i Prilozhen. **18**, 40 (1984)
28. Kirwan, F.: Complex Algebraic Curves. Cambridge University Press, Cambridge (1993)
29. Llibre, J., Pedregal, P.: Hilbert's 16th problem. When variational principles meet differential systems (2015). arXiv: 14116814
30. Nekhoroshev, N.N.: Action-angle variables and their generalizations. Trans. Mosc. Math. Soc. **26**, 180 (1972)

31. Nekhoroshev, N.N., Sadovskiĭ, D.A., Zhilinskiĭ, B.I.: Fractional Hamiltonian monodromy. Ann. Inst. Henri Poincaré **7**, 1099 (2006)
32. Sadovskiĭ, D.A., Zhilinskiĭ, B.I.: Hamiltonian systems with detuned 1:1:2 resonance: Manifestation of bidromy. Ann. Phys. **32**, 164 (2007)
33. Sugny, D., Mardesic, P., Pelletier, M., Jebrane, A., Jauslin, H.R.: Fractional Hamiltonian monodromy from a Gauss-Manin monodromy. J. Math. Phys. **49**, 042701 (2008)
34. Varchenko, A.N.: Estimation of the number of zeros of an abelian integral depending on a parameter and limit cycles. Funct. Anal. Appl. **18**, 14 (1984)
35. Vivolo, O.: The monodromy of the Lagrange top and the Picard-Lefschetz formula. J. Geom. Phys. **46**, 99 (2003)
36. Vu Ngoc, S.: Quantum monodromy in integrable systems. Commun. Math. Phys. **203**, 465 (1999)
37. Winnewisser, B.P., Winnewisser, M., Medvedev, I.R., Behnke, M., De Lucia, F.C., Ross, S.C., Koput, J.: Experimental confirmation of quantum monodromy: the millimeter wave spectrum of cyanogen isothiocyanate NCNCS. Phys. Rev. Lett. **95**, 243002 (2005)
38. Zoladek, H.: The Monodromy Group. Birkhauser, Boston (2006)

Chapter 16
Towards a Science of Security Games

Thanh Hong Nguyen, Debarun Kar, Matthew Brown, Arunesh Sinha, Albert Xin Jiang, and Milind Tambe

Abstract Security is a critical concern around the world. In many domains from counter-terrorism to sustainability, limited security resources prevent full security coverage at all times; instead, these limited resources must be scheduled, while simultaneously taking into account different target priorities, the responses of the adversaries to the security posture and potential uncertainty over adversary types.

Computational game theory can help design such security schedules. Indeed, casting the problem as a Bayesian Stackelberg game, we have developed new algorithms that are now deployed over multiple years in multiple applications for security scheduling. These applications are leading to real-world use-inspired research in the emerging research area of "security games"; specifically, the research challenges posed by these applications include scaling up security games to large-scale problems, handling significant adversarial uncertainty, dealing with bounded rationality of human adversaries, and other interdisciplinary challenges.

Keywords Security games • Bayesian Stackelberg games • Game theory • Scalability • Uncertainty • Bounded rationality

16.1 Introduction

Security is a critical concern around the world that arises in protecting our ports, airports, transportation and other critical national infrastructure from adversaries, in protecting our wildlife and forests from poachers and smugglers, and in curtailing the illegal flow of weapons, drugs, and money; and it arises in problems ranging from physical to cyber-physical systems. In all of these problems, we have limited security resources which prevent full security coverage at all times; instead, security

T.H. Nguyen (✉) • D. Kar • M. Brown • A. Sinha • M. Tambe
University of Southern California, Los Angeles, CA, USA
e-mail: thanhhng@usc.edu; dkar@usc.edu; mattheab@usc.edu; aruneshs@usc.edu; tambe@usc.edu

A.X. Jiang
Trinity University, San Antonio, TX, USA
e-mail: xjiang@trinity.edu

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_16

resources must be deployed intelligently taking into account differences in priorities of targets requiring security coverage, the responses of the attackers to the security posture, and potential uncertainty over the types, capabilities, knowledge, and priorities of attackers faced.

Game theory, which studies interactions among multiple self-interested agents, is well-suited to the adversarial reasoning required for security resource allocation and scheduling problems. Casting the problem as a Bayesian Stackelberg game, we have developed new algorithms for efficiently solving such games that provide randomized patrolling or inspection strategies. These algorithms have led to some initial successes in this challenging problem arena, leading to advances over previous approaches in security scheduling and allocation, e.g., by addressing key weaknesses of predictability of human schedulers. These algorithms are now deployed in multiple applications: ARMOR has been deployed at the Los Angeles International Airport (LAX) since 2007 to randomize checkpoints on the roadways entering the airport and canine patrol routes within the airport terminals [17]; IRIS, a game-theoretic scheduler for randomized deployment of the US Federal Air Marshals Service (FAMS) requiring significant scale-up in underlying algorithms, has been in use since 2009 [17]; PROTECT, which schedules the US Coast Guard's (USCG) randomized patrolling of ports using a new set of algorithms based on modeling bounded-rational human attackers, has been deployed in the port of Boston since April 2011 and is in use at the port of New York since February 2012 [39], and is headed for nationwide deployment; another application for deploying escort boats to protect ferries has been deployed by the USCG since April 2013 [10]; and TRUSTS [51] has been evaluated in field trials by the Los Angeles Sheriffs Department (LASD) in the LA Metro system and a nationwide deployment is now being evaluated at TSA. Most recently, PAWS—another game-theoretic application using a Bayesian distribution of boundedly rational attackers was tested by rangers in Uganda for protecting wildlife in Queen Elizabeth National Park (QENP) in April 2014 [49]; MIDAS which is based on modeling behaviors of attackers combined with the robust approach is in use by USCG for protecting fisheries [14]. These initial successes point the way to major future applications in a wide range of security domains; with major research challenges in scaling up our game-theoretic algorithms, in addressing human adversaries' bounded rationality and uncertainties in action execution and observation, as well as in multiagent learning.

Given many game-theoretic applications for solving real-world security problems, this book chapter will provide an overview of the models and algorithms, key research challenges and a brief description of our successful deployments with emphasis on *three key lessons*: (1) computational game theory-based decision aids are in daily use by security agencies due to their capability for optimizing limited security resources against strategic adversaries; (2) these applications provide fundamental research challenges, leading to an (emerging) science of security games, including the challenge of massive scale games which cannot fit into memory and the challenge of modeling many different forms of uncertainty in outcomes and preferences, action execution, and human decision-making; and (3) current security

game applications for solving *green security games* such as protecting wildlife and the environment are challenging for AI; these are important global problems that provide open research problems to integrate AI research (including planning and learning) in security games.

16.2 Stackelberg Security Games

Stackelberg security games (SSGs) were first introduced to model leadership and commitment [44], and are now used to study security problems ranging from "police and robbers" scenario [12], computer network security [29], missile defense systems [5], and terrorism [38]. Models for arms inspections and border patrolling have also been modeled using inspection games [3], a related family of Stackelberg games.

This section provides details on this use of Stackelberg games for modeling security domains. We first give a generic description of security domains followed by *security games*, the model by which security domains are formulated in the Stackelberg game framework.

16.2.1 Security Domain Description

In a security domain, a defender must perpetually defend a set of targets using a limited number of resources, whereas the attacker is able to surveil and learn the defender's strategy and attack after careful planning. This fits precisely into the description of a Stackelberg game if we map the defender to the leader's role and the attacker to the follower's role [3, 6]. An action, or *pure strategy*, for the defender represents deploying a set of resources on patrols or checkpoints, e.g., scheduling checkpoints at the LAX airport or assigning federal air marshals to protect flight tours. The pure strategy for an attacker represents an attack at a target, e.g., a flight. The strategy for the leader is a *mixed strategy*, a probability distribution over the pure strategies of the defender. Additionally, with each target are also associated a set of payoff values that define the utilities for both the defender and the attacker in case of a successful or a failed attack. These payoffs are represented using the *security game* model, described next.

16.2.2 Definition of SSGs

A key assumption of security games is that the payoff of an outcome depends only on the target attacked, and whether or not it is covered by the defender [25]. The payoffs do *not* depend on the remaining aspects of the defender allocation.

Table 16.1 Example of a security game with two targets

	Defender		Attacker	
Target	Covered	Uncovered	Covered	Uncovered
t_1	10	0	−1	1
t_2	0	−10	−1	1

For example, if an adversary succeeds in attacking target t_1, the penalty for the defender is the same whether the defender was guarding target t_2 or not.

This allows us to compactly represent the payoffs of a security game. Specifically, a set of four payoffs is associated with each target. These four payoffs are the rewards and penalties to both the defender and the attacker in case of a successful or an unsuccessful attack, and are sufficient to define the utilities for both players for all possible outcomes in the security domain. Table 16.1 shows an example security game with two targets: t_1 and t_2. In this example game, if the defender was *covering* (protecting) target t_1 and the attacker attacked t_1, the defender would get 10 units of reward whereas the attacker would receive −1 units. We make the assumption that in a security game it is always better for the defender to cover a target as compared to leaving it uncovered, whereas it is always better for the attacker to attack an uncovered target. This assumption is consistent with the payoff trends in the real-world. A special case is *zero-sum games*, in which for each outcome the sum of utilities for the defender and attacker is zero, although in general security games are not necessarily zero-sum.

In the above example, all payoff values are exactly known. In practice, we often have uncertainty over the payoffs and preferences of the players. Bayesian games are a well-known game-theoretic model in which such uncertainty is modeled using multiple types of players, with each associated with its own payoff values. For security games of interest, the main source of payoff uncertainty is regarding the attacker's payoffs. In the resulting *Bayesian Stackelberg game* model, there is only one leader type (e.g., only one police force), although there can be multiple follower types (e.g., multiple attacker types trying to infiltrate security) [35]. Each follower type is represented using a different payoff matrix. The leader does not know the follower's type, but knows the probability distribution over them. The goal is to find the optimal mixed strategy for the leader to commit to, given that the defender could be facing any of the follower types.

16.2.3 Solution Concept: Strong Stackelberg Equilibrium

The solution to a security game is a mixed strategy for the defender that maximizes the expected utility of the defender, given that the attacker learns the mixed strategy of the defender and chooses a best response for himself. This solution concept is known as a Stackelberg equilibrium [27].

The most commonly adopted version of this concept in related literature is called strong Stackelberg equilibrium (SSE) [4, 9, 35, 45]. An SSE for security games is informally defined as follows (the formal definition of SSE is not introduced for brevity, and can instead be found in [25]):

Definition 1. A pair of strategies form a *SSE* if they satisfy

1. The defender plays a best response, that is, the defender cannot get a higher payoff by choosing any other strategy.
2. The attacker plays a best response, that is, given a defender strategy, the attacker cannot get a higher payoff by attacking any other target.
3. The attacker breaks ties in favor of the leader.

The assumption that the follower will always break ties in favor of the leader in cases of indifference is reasonable because in most cases the leader can induce the favorable strong equilibrium by selecting a strategy arbitrarily close to the equilibrium that causes the follower to strictly prefer the desired strategy [45]. Furthermore an SSE exists in all Stackelberg games, which makes it an attractive solution concept compared to versions of Stackelberg equilibrium with other tie-breaking rules. Finally, although initial applications relied on the SSE solution concept, we have since proposed new solution concepts that are more robust against various uncertainties in the model [1, 37, 50] and have used these robust solution concepts in some of the later applications.

16.3 Deployed Real-World Security Applications

In this section, we describe several deployed and emerging applications of the Stackelberg game framework in different real-world domains. Besides describing successful transitions of research, our aim is to set the stage for later sections in which we discuss the research challenges that arise.

16.3.1 ARMOR for Los Angeles International Airport

Los Angeles International Airport (LAX) is the largest destination airport in the USA and serves 60–70 million passengers per year. The LAX police use diverse measures to protect the airport, which include vehicular checkpoints, police units patrolling the roads to the terminals, patrolling inside the terminals (with canines), and security screening and bag checks for passengers. The application of our game-theoretic approach is focused on two of these measures: (1) placing vehicle checkpoints on inbound roads that service the LAX terminals, including both location and timing, and (2) scheduling patrols for bomb-sniffing canine units at the different LAX terminals. The eight different terminals at LAX have very different

Fig. 16.1 LAX checkpoints are deployed using ARMOR

characteristics, like physical size, passenger loads, international versus domestic flights, etc. These factors contribute to the differing risk assessments of these eight terminals. Furthermore, the numbers of available vehicle checkpoints and canine units are limited by resource constraints. Thus, it is challenging to optimally allocate these resources to improve their effectiveness while avoiding patterns in the scheduled deployments.

The ARMOR system (Assistant for Randomized Monitoring over Routes) focuses on two of the security measures at LAX (checkpoints and canine patrols) and optimizes security resource allocation using Bayesian Stackelberg games. Take the vehicle checkpoints model as an example. Assuming that there are n roads, the police's strategy is placing $m < n$ checkpoints on these roads where m is the maximum number of checkpoints. ARMOR randomizes allocation of checkpoints to roads. The adversary may conduct surveillance of this mixed strategy and may potentially choose to attack through one of these roads. ARMOR models different types of attackers with different payoff functions, representing different capabilities and preferences for the attacker. ARMOR uses DOBSS (Decomposed Optimal Bayesian Stackelberg Solver) [35] to compute the defender's optimal strategy. ARMOR has been successfully deployed since August 2007 at Fig. 16.1.

16.3.2 IRIS for US FAMS

The US FAMS allocates air marshals to flights originating in and departing from the USA to dissuade potential aggressors and prevent an attack should one occur. Flights are of different importance based on a variety of factors such as the numbers of passengers, the population of source and destination, and international flights from different countries. Security resource allocation in this domain is significantly

more challenging than for ARMOR: a limited number of air marshals need to be scheduled to cover thousands of commercial flights each day. Furthermore, these air marshals must be scheduled on tours of flights that obey various constraints (e.g., the time required to board, fly, and disembark). Simply finding schedules for the marshals that meet all of these constraints is a computational challenge. Our task is made more difficult by the need to find a randomized policy that meets these scheduling constraints, while also accounting for the different values of each flight.

Against this background, the IRIS system (Intelligent Randomization In Scheduling) has been developed and deployed by FAMS since October 2009 to randomize schedules of air marshals on international flights. In IRIS, the targets are the set of n flights and the attacker could potentially choose to attack one of these flights. The FAMS can assign $m < n$ air marshals that may be assigned to protect these flights. Since the number of possible schedules exponentially increases with the number of flights and resources, DOBSS is no longer applicable to the FAMS domain. Instead, IRIS uses the much faster ASPEN algorithm [16] to generate the schedule for thousands of commercial flights per day.

16.3.3 PROTECT for USCG

The USCG's mission includes maritime security of the US coasts, ports, and inland waterways; a security domain that faces increased risks due to threats such as terrorism and drug trafficking. Given a particular port and the variety of critical infrastructure that an attacker may attack within the port, USCG conducts patrols to protect this infrastructure; however, while the attacker has the opportunity to observe patrol patterns, limited security resources imply that USCG patrols cannot be at every location 24/7. To assist the USCG in allocating its patrolling resources, the PROTECT (Port Resilience Operational/Tactical Enforcement to Combat Terrorism) model has been designed to enhance maritime security. It has been in use at the port of Boston since April 2011, and is also in use at the port of New York since February 2012 (Fig. 16.2). Similar to previous applications ARMOR and IRIS, PROTECT uses an attacker–defender Stackelberg game framework, with USCG as the defender against terrorists that conduct surveillance before potentially launching an attack.

The key idea in PROTECT is also that unpredictability creates situations of uncertainty for an enemy and can be enough to deem a target less appealing. While randomizing patrol patterns is key, PROTECT also addresses the fact that the targets are of unequal value, understanding that the attacker will adapt to whatever patrol patterns USCG conducts. The output of PROTECT is a schedule of patrols which includes when the patrols are to begin, what critical infrastructure to visit for each patrol, and what activities to perform at each critical infrastructure.

While PROTECT builds on previous work, it offers key innovations. First, this system is a departure from the assumption of perfect attacker rationality noted in

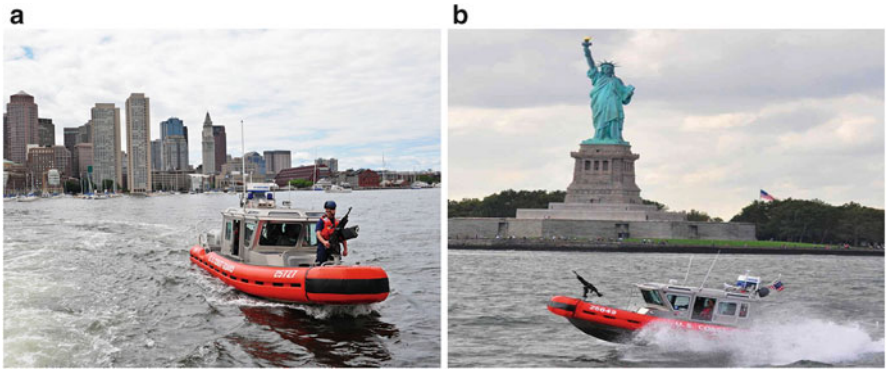


Fig. 16.2 USCG boats patrolling the ports of Boston and NY. (**a**) PROTECT is being used in Boston. (**b**) Extending PROTECT to NY

previous work, relying instead on a quantal response model [31] of the attacker's behavior. Second, to improve PROTECT's efficiency, a compact representation of the defender's strategies is used by exploiting equivalence and dominance. Finally, the evaluation of PROTECT for the first time provides real-world data: (1) comparison of human-generated vs PROTECT schedules, and (2) results from an Adversarial Perspective Team's (APT) (human mock attackers) analysis. The PROTECT model has now been extended to other US ports like Los Angeles/Long Beach and is moving towards nationwide deployment.

16.3.4 Ferry Protection for the USCG

Another problem that USCG faces is the protection of ferries, including the Staten Island Ferry in New York, from potential terrorist attacks from water. We developed a game-theoretic system for scheduling escort boat patrols to protect ferries, and this has been deployed at the Staten Island Ferry since 2013 [10] (Fig. 16.3). The key research challenge is the fact that the ferries are continuous moving in a continuous domain, and the attacker could attack at any moment in time. This type of moving targets domain leads to game-theoretic models with continuous strategy spaces, which presents computational challenges. Our theoretical work showed that while it is safe to discretize the defender's strategy space, discretizing the attacker's strategy space would result in loss of utility. We developed a novel algorithm that uses a compact representation for the defender's mixed-strategy space while being able to exactly model the attacker's continuous strategy space. The implemented algorithm, running on a laptop, is able to generate daily schedules for escort boats with guaranteed expected utility values.

Fig. 16.3 Escort boats protecting the Staten Island Ferry use strategies generated by our system

Fig. 16.4 TRUSTS for transit systems. (**a**) Los Angeles Metro. (**b**) Barrier-free entrance to transit system

16.3.5 TRUSTS for Security in Transit Systems

Urban transit systems face multiple security challenges, including deterring fare evasion, suppressing crime and counter-terrorism. In particular, in some urban transit systems, including the Los Angeles Metro Rail system, passengers are legally required to purchase tickets before entering but are not physically forced to do so (Fig. 16.4). Instead, security personnel are dynamically deployed throughout the transit system, randomly inspecting passenger tickets. This proof-of-payment fare collection method is typically chosen as a more cost-effective alternative to direct fare collection, i.e., when the revenue lost to fare evasion is believed to be less than what it would cost to directly preclude it. In the case of Los Angeles Metro,

with approximately 300,000 riders daily, this revenue loss can be significant; the annual cost has been estimated at $5.6 million [13]. The Los Angeles Sheriffs Department (LASD) deploys uniformed patrols on board trains and at stations for fare-checking (and for other purposes such as crime prevention). The LASD's current approach relies on humans for scheduling the patrols, which places a tremendous cognitive burden on the human schedulers who must take into account all of the scheduling complexities (e.g., train timings, switching time between trains, and schedule lengths).

The TRUSTS system (Tactical Randomization for Urban Security in Transit Systems) models the patrolling problem as a leader–follower Stackelberg game [51]. The leader (LASD) pre-commits to a mixed-strategy patrol (a probability distribution over all pure strategies), and riders observe this mixed strategy before deciding whether to buy the ticket or not. Both ticket sales and fines issued for fare evasion translate into revenue for the government. Therefore the utility for the leader is the total revenue (total ticket sales plus penalties). The main computational challenge is the exponentially many possible patrol strategies, each subject to both the spatial and temporal constraints of travel within the transit network under consideration. To overcome this challenge, TRUSTS uses a compact representation of the strategy space which captures the spatiotemporal structure of the domain.

The LASD conducted field tests of this TRUSTS system in the LA Metro in 2012, and one of the feedback comments from the officers was that patrols are often interrupted due to execution uncertainty such as emergencies and arrests. Utilizing techniques from planning under uncertainty [in particular Markov Decision Processes (MDPs)], we proposed a general approach to dynamic patrolling games in uncertain environments, which provides patrol strategies with contingency plans [20]. This led to schedules now being loaded onto smartphones and given to officers. If interruptions occur, the schedules are then automatically updated on the smartphone app. The LASD has conducted successful field evaluations using the smartphone app, and the TSA is currently evaluating it towards nationwide deployment.

Crime presents a serious problem in transit systems like LA Metro. Furthermore, unlike terrorists that strategically plans an attack, criminals are often opportunistic, in that their decisions are based on the available opportunities encountered. For the crime problem, we developed a new game-theoretic model that utilizes recent advances in criminology on modeling opportunistic criminals, and novel efficient algorithms that achieve speed-ups by exploiting the spatiotemporal structure of the domain [53].

16.3.6 Fishery Protection for USCG

Fisheries are a vital natural resource from both an ecological and economic standpoint. However, fish stocks around the world are threatened with collapse due to illegal, unreported, and unregulated (IUU) fishing. In the USA, the Coast Guard

(USCG) is tasked with the responsibility of protecting and maintaining the nation's fisheries. To this end, the USCG deploys resources (both air and surface assets) to conduct patrols over fishery areas in order to deter and mitigate IUU fishing. Due to the large size of these patrol areas and the limited patrolling resources available, it is impossible to protect an entire fishery from IUU fishing at all times. Thus, an intelligent allocation of patrolling resources is critical for security agencies like the USCG.

The MIDAS algorithm was developed to address the types challenges faced in natural resource conservation domains such as fishery protection. In stark contrast to counter-terrorism settings, there is frequent interaction between the defender and attacker in these resource conservation domains. This distinction is important for three reasons. First, due to the comparatively low stakes of the interactions, rather than a handful of persons or groups, the defender must protect against numerous adversaries (potentially hundreds or even more), each of which may behave differently. Second, frequent interactions make it possible to collect data on the actions of the adversaries actions over time. Third, the adversaries are less strategic given the short planning windows between actions. Combining these factors, MIDAS models a population of boundedly rational adversaries and utilizes available data to learn the behavior models of the adversaries using the subjective utility quantal response (SUQR) model in order to improve the way the defender allocates its patrolling resources.

MIDAS has been successfully deployed and evaluated by the USCG in the Gulf of Mexico. Historical data on fish stock densities, USCG air and surface patrols, as well as IUU sightings and interdictions was used to construct the game model. Between July and September 2014, six aircraft patrols were generated weekly to protect a 80 by 60 nautical mile area on the US–Mexico border off the coast of Texas. This region represents a critical fishery for red snapper, a species that is highly lucrative to fish, and as such observes a high volume of IUU fishing. This evaluation period in the Gulf of Mexico represents the most sophisticated real-world deployment of security games to date. MIDAS is currently under review by the USCG and is being considered for further deployment in the Gulf of Mexico as well as in other fisheries nationwide.

16.4 Emerging Real-World Security Applications

16.4.1 Networked Domains

Beyond the deployed applications above, there are a number of emerging application areas. One such area of great importance is securing urban city networks, transportation networks, computer networks, and other network centric security domains. For example, after the terrorist attacks in Mumbai of 2008 [8], the Mumbai police have started setting up vehicular checkpoints on roads. We can model the problem

faced by the Mumbai police as a security game between the Mumbai police and an attacker. In this urban security game, the pure strategies of the defender correspond to allocations of resources to edges in the network—for example, an allocation of police checkpoints to roads in the city. The pure strategies of the attacker correspond to paths from any *source* node to any *target* node—for example, a path from a landing spot on the coast to the airport. The strategy space of the defender grows exponentially with the number of available resources, whereas the strategy space of the attacker grows exponentially with the size of the network. In addressing this computational challenge, novel algorithms based on incremental strategy generation have been able to generate randomized defender strategies that scale up to the entire road network of Mumbai [19].

The Stackelberg game framework can also be applied to adversarial domains that exhibit "contagious" actions for each player. For example, word-of-mouth advertising/viral marketing has been widely studied by marketers trying to understand why one product or video goes "viral" while others go unnoticed. Counter-insurgency is the contest for the support of the local leaders in an armed conflict and can include a variety of operations such as providing security and giving medical supplies. These efforts carry a social effect beyond the action taken that can cause advantageous ripples through the neighboring population. Moreover, multiple intelligent parties attempt to leverage the same social network to spread their message, necessitating an adversary-aware approach to strategy generation. Game-theoretic approaches can be used to generate resource-allocations strategies for such large-scale, real-world networks [41, 42]. This interaction can be modeled as a graph with one player attempting to spread influence while another player attempts to stop the probabilistic propagation of that influence by spreading their own influence. This "blocking" problem models situations faced by governments/peacekeepers combatting the spread of terrorist radicalism and armed conflict with daily/weekly/monthly visits with local leaders to provide support and discuss grievances [15].

Game-theoretic methods are also appropriate for modeling resource allocation in cyber-security such as packet selection and inspection for detecting potential threats in large computer networks. The problem of attacks on computer systems and corporate computer networks gets more pressing each year. A number of intrusion detection and monitoring systems are being developed, e.g., deep packet inspection method that periodically selects a subset of packets in a computer network for analysis. The attacking/protecting problem can be formulated as a game between two players: the attacker (or the intruder) and the defender (the detection system). The actions of the attacker can be seen as sending malicious packets from a controlled computer to vulnerable computers. The objective of the defender is to prevent the intruder from succeeding by selecting the packets for inspection and subsequently thwarting the attack. However, packet inspections cause unwanted latency and hence the defender has to decide where and how to inspect network traffic. The computational challenge is efficiently computing the optimal defending strategies for such network scenarios [43].

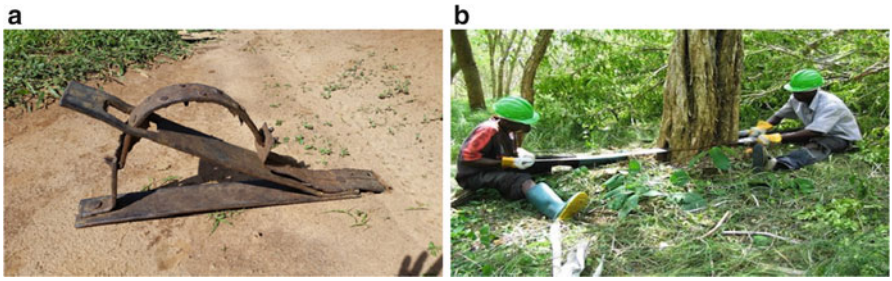


Fig. 16.5 Examples of illegal activities in green security domains. (**a**) An illegal trapping tool. (**b**) Illegally cutting trees.

16.4.2 Green Security Domains

A number of our newer applications are focused on resource conservation through suppression of environmental crime. One area is protecting forests [22], where we must protect a continuous forest area from extractors by patrols through the forest that seek to deter such extraction activity (Fig. 16.5). With limited resources for performing such patrols, a patrol strategy will seek to distribute the patrols throughout the forest, in space and time, in order to minimize the resulting amount of extraction that occurs or maximize the degree of forest protection. This problem can be formulated as a Stackelberg game and the focus is on computing optimal allocations of patrol density [22].

Endangered species poaching is reaching critical levels as the populations of these species plummet to unsustainable numbers. The global tiger population, for example, has dropped over 95 % from the start of the 1900s and has resulted in three out of nine species extinctions. Depending on the area and animals poached, motivations for poaching range from profit to sustenance, with the former being more common when profitable species such as tigers, elephants, and rhinos are the targets. To counter poaching efforts and to rebuild the species' populations, countries have set up protected wildlife reserves and conservation agencies tasked with defending these large reserves. Because of the size of the reserves and the common lack of law enforcement resources, conservation agencies are at a significant disadvantage when it comes to deterring and capturing poachers. Agencies use patrolling as a primary method of securing the park. Due to their limited resources, however, patrol managers must carefully create patrols that account for many different variables (e.g., limited patrol units to send out, multiple locations that poachers can attack at varying distances to the outpost). Our proposed system Protection Assistant for Wildlife Security (PAWS) aims to assist conservation agencies in their critical role of patrol creation by predicting where poachers will attack and optimizing patrol routes to cover those areas.

16.5 Scale Up to Real-World Problem Sizes

The wide use of Stackelberg games has inspired theoretical and algorithmic progress leading to the development of fielded applications, as described in Sect. 16.3. For example, DOBSS [35], an algorithm for solving Bayesian Stackelberg games, is central to the fielded application ARMOR in use at the Los Angeles International Airport [17]. Conitzer and Sandholm [9] gave complexity results and algorithms for computing optimal commitment strategies in Bayesian Stackelberg games, including both pure- and mixed-strategy commitments.

These early works assumed that the set of pure strategies for the players are given explicitly. Many real-world problems, like the FAMS and urban road networks, present billions of pure strategies to both the defender and the attacker. Such large problem instances cannot even be represented in modern computers, let alone solved using previous techniques. We have proposed models and algorithms that compute optimal defender strategies for massive real-world security domains [16, 18].

16.5.1 Scale Up with Defender Pure Strategies

In this section, we describe one particular algorithm ASPEN, that computes SSE in domains with a *very large* number of pure strategies (up to billions of actions) for the defender [16]. ASPEN builds on the insight that in many real-world game-theoretic problems, there exist solutions with *small support sizes*, which are mixed strategies in which only a small set of pure strategies are played with positive probability [28]. ASPEN exploits this by using a *strategy generation* approach for the defender, in which defender pure strategies are iteratively generated and added to the optimization formulation.

As an example, let us consider the problem faced by the FAMS. There are currently tens of thousands of commercial flights flying each day, and public estimates state that there are thousands of air marshals that are scheduled daily by the FAMS [24]. Air marshals must be scheduled on tours of flights that obey logistical constraints (e.g., the time required to board, fly, and disembark). An example of a schedule is an air marshal assigned to a round trip from Los Angeles to New York and back.

ASPEN [16] casts this problem as a security game, where the attacker can choose any of the flights to attack, and each air marshal can cover one schedule. Each schedule here is a feasible set of targets that can be covered together; for the FAMS, each schedule would represent a flight tour which satisfies all the logistical constraints that an air marshal could fly. A *joint schedule* then would assign every air marshal to a flight tour, and there could be exponentially many joint schedules in the domain. A pure strategy for the defender in this security game is a joint schedule. As mentioned previously, ASPEN employs strategy generation since all the defender pure strategies cannot be enumerated for such a massive problem. ASPEN decomposes the problem into a *master* problem and a *slave* problem, which

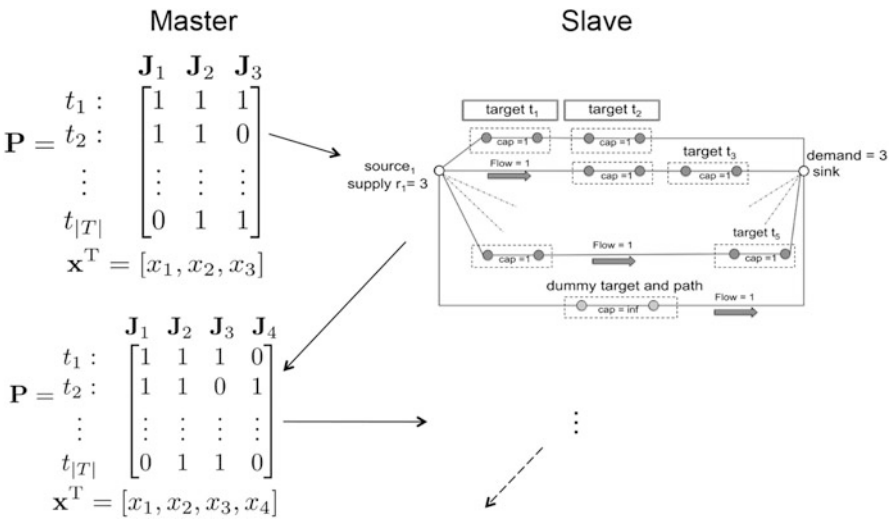


Fig. 16.6 Strategy generation employed in ASPEN: the schedules for a defender are generated iteratively. The *slave* problem is a novel minimum-cost integer flow formulation that computes the new pure strategy to be added to $\mathbf{P}$; $\mathbf{J}_4$ is computed and added in this example

are then solved iteratively. Given a number of pure strategies, the master solves for the defender and the attacker optimization constraints, while the slave is used to generate a new pure strategy for the defender in every iteration.

The iterative process is graphically depicted in Fig. 16.6. The master operates on the pure strategies (joint schedules) generated thus far, which are represented using the matrix $\mathbf{P}$. Each column of $\mathbf{P}$, $\mathbf{J}_j$, is one pure strategy (or joint schedule). An entry P_{ij} in the matrix $\mathbf{P}$ is 1 if a target t_i is covered by joint-schedule $\mathbf{J}_j$, and 0 otherwise. The objective of the master problem is to compute $\mathbf{x}$, the optimal mixed strategy of the defender over the pure strategies in $\mathbf{P}$. The objective of the slave problem is to generate the best joint schedule to add to $\mathbf{P}$. The best joint schedule is identified using the concept of *reduced costs*, which measures if a pure strategy can potentially increase the defender's expected utility (the details of the approach are provided in [16]). While a naïve approach would be to iterate over all possible pure strategies to identify the pure strategy with the maximum potential, ASPEN uses a novel minimum-cost integer flow problem to efficiently identify the best pure strategy to add. ASPEN always converges on the optimal mixed strategy for the defender.

Employing strategy generation for large optimization problems is not an "out-of-the-box" approach, the problem has to be formulated in a way that allows for domain properties to be exploited. The novel contribution of ASPEN is to provide a linear formulation for the master and a minimum-cost integer flow formulation for the slave, which enables the application of strategy generation techniques. Additionally, ASPEN also provides a branch-and-bound heuristic to reason over attacker actions. This branch-and-bound heuristic provides a further order of magnitude speed-up, allowing ASPEN to handle the massive sizes of real-world problems.

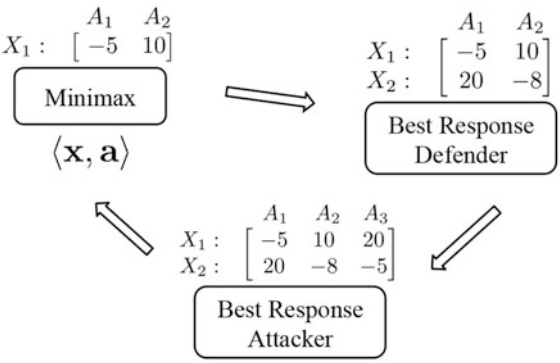


Fig. 16.7 Strategy generation employed in RUGGED: the pure strategies for both the defender and the attacker are generated iteratively

16.5.2 Scale Up with Defender and Attacker Pure Strategies

In domains such as the urban network security setting described in Sect. 16.4, the number of pure strategies of both the defender and the attacker are exponentially large. In this section, we describe the RUGGED algorithm [18], which generates pure strategies for both the defender and the attacker.

RUGGED models the domain as a zero-sum game, and computes the minimax equilibrium, since the minimax strategy is equivalent to the SSE in zero-sum games. Figure 16.7 shows the working of RUGGED: at each iteration, the minimax module generates the optimal mixed strategies $\langle \mathbf{x}, \mathbf{a} \rangle$ for the two players for the current payoff matrix, the Best Response Defender module generates a new strategy for the defender that is a best response against the attacker's current strategy $\mathbf{a}$, and the Best Response Attacker module generates a new strategy for the attacker that is a best response against the defender's current strategy $\mathbf{x}$. The rows X_i in the figure are the pure strategies for the defender, they would correspond to an allocation of checkpoints in the urban road network domain. Similarly, the columns A_j are the pure strategies for the attacker, they represent the attack paths in the urban road network domain. The values in the matrix represent the payoffs to the defender. The algorithm stops when neither of the generated best responses improve on the current minimax strategies.

The contribution of RUGGED is to provide the mixed-integer formulations for the best response modules which enable the application of such a strategy generation approach. RUGGED can compute the optimal solution for deploying up to 4 resources in real-city network with as many as 250 nodes within a reasonable time frame of 10 h (the complexity of this problem can be estimated by observing that both the best response problems are NP-hard themselves [18]). More recent work [19] builds on RUGGED and proposes SNARES, which allows scale-up to the entire city of Mumbai, with 10–15 checkpoints.

16.5.3 Scale Up with Mobile Resources and Moving Targets

In this section, we describe the CASS (Solver for Continuous Attacker Strategy) algorithm [10] for solving security problems where the defender has mobile patrollers to protect a set of mobile targets against the attacker who can attack these moving targets at any time during their movement. In these security problems, the sets of pure strategies for both the defender and attacker are continuous w.r.t the continuous spatial and time components of the problem domain. The CASS algorithm attempts to compute the optimal mixed strategy for the defender without discretizing the attacker's continuous strategy set; it exactly models this set using sub-interval analysis which exploits the piecewise-linear structure of the attacker's expected utility function. The insight of CASS is to compactly represent the defender's mixed strategies as a *marginal* probability distribution, overcoming the short-coming of an exponential number of pure strategies for the defender.

As a domain example, in the problem of protecting ferries described in Sect. 16.3.4, there are a number of ferries carrying hundreds of passengers in many waterside cities. These ferries are attractive targets for an attacker who can approach the ferries with a small boat packed with explosives at any time; this attacker's boat may only be detected when it comes close to the ferries. Small, fast, and well-armed patrol boats can provide protection to such ferries by detecting the attacker within a certain distance and stop him from attacking with the armed weapons. However, the numbers of patrol boats are often limited, thus the defender cannot protect the ferries at all times and locations.

CASS casts this problem as a *zero-sum* security game in which targets move along a *one-dimensional* domain, i.e., a straight line segment connecting two terminal points. This *one-dimensional* assumption is valid as in real-world domains such as ferry protection, ferries normally move back-and-forth in a straight line between two terminals (i.e., ports) around the world. Although the targets' locations vary w.r.t time changes, these targets have a fixed daily schedule, meaning that determining the locations of the targets at a certain time is straightforward. The defender has mobile patrollers (i.e., boats) that can move along between two terminals to protect the targets. While the defender is trying to protect the targets, the attacker will decide to attack a certain target at a certain time. The probability that the attacker successfully attacks depends on the positions of the patroller at that time. Specifically, each patroller possesses a protective circle of radius within which she can detect and try to intercept any attack, whereas she is incapable of detecting the attacker prior to that radius.

In CASS, the defender's strategy space is discretized and her mixed strategy is compactly represented using flow distributions. Figure 16.8 shows an example of a ferry transition graph in which each node of the graph indicates a particular pair of (location and time step) for the target. Here, there are three location points, namely A, B, and C on a straight line where B lies between A and C. Initially, the target is at one of these location points at the 5-min time step. Then the target moves to the next location point which is determined based on the connectivity between these

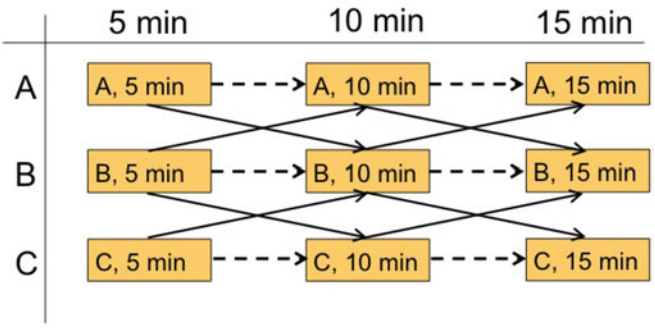


Fig. 16.8 An example of a ferry transition graph

points at the 10-min time step, and so on. For example, if the target is at the location point A at the 5-min time step, denoted by (A, 5 min) in the transition graph, it can move to the location point B or stay at location point A at the 10-min time step. The defender follows this transition graph to protect the target. A pure strategy for the defender is defined as a trajectory of this graph, e.g., the trajectory including (A, 5 min), (B, 10 min), and (C, 15 min) indicates a pure strategy for the defender. One key challenge of this representation for the defender's pure strategies is that the transition graph consists of an exponential number of trajectories, i.e., $O(N^T)$ where N is the number of location points and T is the number of time steps. To address this challenge, CASS proposes a compact representation of the defender's mixed strategy. Instead of directly computing a probability distribution over pure strategies for the defender, CASS attempts to compute the marginal probability that the defender will follow a certain edge of the transition graph, e.g., the probability of being at the node (A, 5 min) and moving to the node (B, 10 min). CASS shows that *any strategy in full representation can be mapped into a compact representation* as well as *compact representation does not lead to any loss in solution quality*. This compact representation allows CASS to reformulate the resource-allocation problem as computing the optimal *marginal* coverage of the defender over a number of $O(NT)$ the edges of the transition graph.

16.5.4 Scale Up with Continuous Domains and Boundedly Rational Attacker

As discussed in Sect. 16.3, natural resource conservation domains such as fishery protection introduce a unique set of challenges which must be addressed, namely *scalability* and *robustness*. For scalability, the defender is responsible for protecting a large patrol area and therefore must consider a large strategy space. Even if the patrol area is discretized into a grid or graph structure, the defender must still reason over an exponential number of patrol strategies. For robustness, the defender

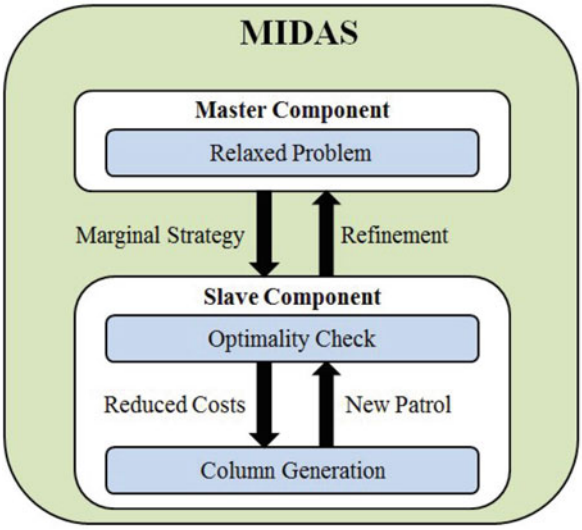


Fig. 16.9 Overview of the multiple iterative process within the MIDAS algorithm

must protect against *multiple* boundedly rational adversaries. Bounded rationality models, such as the quantal response (QR) model [31] and the SUQR model [32], introduce stochastic actions, relaxing the strong assumption in classical game theory that all players are perfectly rational and utility maximizing. These models are able to better predict the actions of human adversaries and thus lead the defender to choose strategies that perform better in practice. However, both QR and SUQR are non-linear models resulting in a computationally difficult optimization problem for the defender.

Previous work on boundedly rational adversaries has considered the challenges of scalability and robustness separately, in [47, 48] and [14, 49], respectively. The MIDAS algorithm was introduced to merge these two research threads for the first time by addressing scalability and robustness simultaneously. Figure 16.9 provides a visual overview of how MIDAS operates as an iterative process. Given the sheer complexity of the game being solved, the problem is decomposed using a master–slave formulation. The master utilizes multiple simplifications to create a relaxed version of the original problem which is more efficient to solve. First, a piecewise-linear approximation of the security game is taken to make the optimization problem both linear and convex. This is a modified version of the approach in [47], replacing the QR model of the adversary with SUQR and considering a robust maximin formulation over a set of boundedly rational adversaries. Second, the complex spatiotemporal constraints associated with patrols are initially ignored and then incrementally added back using cut generation.

Due to the relaxations, solving the master produces a marginal strategy $\mathbf{x}$ which is a probability distribution over targets. However, the defender ultimately needs a probability distribution over patrols. Additionally, since not all of the spatiotemporal constraints are considered in the master, the relaxed solution $\mathbf{x}$ may not be a feasible solution to the original problem. Therefore, the slave checks if the marginal strategy

x can expressed as a linear combination, i.e., probability distribution, of patrols by computing a one-norm minimization. If the one-norm distance is zero, the marginal distribution can be translated to a feasible pure strategy distribution which is in fact the optimal solution to the original problem. Otherwise, the marginal distribution is infeasible for the original problem. However, given the exponential number of patrol strategies, even performing this optimality check is intractable. Thus, column generation is used *within* the slave where only a small set of patrols is considered initially in the optimality check and the set is expanded over time. Much like previous examples of column generation in security games, e.g., [16], new patrols are added by solving a minimum-cost network flow problem using reduced cost information from the optimality check. If the optimality check fails, then the slave generates a cut which is returned to refine and constrain the master, incrementally bringing it closer to the original problem. The entire process is repeated until an optimal solution is found.

16.6 Address Uncertainty in Real-World Problems

Addressing uncertainty is a key challenge of solving real-world security problems. Traditional SSGs often assume that the defender has perfect information about the game payoff matrix as well as the attacker's behaviors. Moreover, she is supposed to be capable of exactly executing her patrolling strategy. However, due to limited data, the defender cannot precisely estimate such aspects, i.e., the payoff matrix or attacker's behaviors. Also, there is no guarantee that the defender can exactly follow the patrolling schedule as a result of unseen events that could change her patrolling strategy. These types of uncertainty could deteriorate the effectiveness of the defender's strategy and thus it is important for the defender to address them when generating strategy. This section of the book chapter describes several game-theoretic solutions to deal with uncertainty in SSGs.

16.6.1 Security Patrolling with Dynamic Execution Uncertainty

In security problems such as fare inspections in the Los Angeles Metro Rail system as described in Sect. 16.3.5, the targets, e.g., trains normally follow predetermined schedules, thus timing is an important aspect which determines the effectiveness of the defender's patrolling schedules (the defender needs to be at the right location at a specific time in order to protect these moving targets). However, as a result of execution uncertainty (e.g., emergencies or errors), the defender could not carry out her planned patrolling schedule in later time steps. For example, in real-world trials for TRUSTS carried out by Los Angeles Sheriff's Department (LASD), there

is interruption (due to writing citations, felony arrests, and handling emergencies) in a significant fraction of the executions, causing the officers to miss the train they are supposed to catch as following the pre-generated patrolling schedule.

In this section, we present the Bayesian Stackelberg game model for security patrolling with dynamic execution uncertainty introduced by Jiang et al. [20] in which the uncertainty is represented using MDPs. The key advantage of this game-theoretic model is that patrol schedules which are computed based on Stackelberg equilibrium have contingency plans to deal with interruptions and are robust against execution uncertainty. Specifically, the security problem with execution uncertainty is represented as a two-player Bayesian Stackelberg game between the defender and the attacker. The defender has multiple patrol units while there are also multiple types of attackers which are unknown to the defender. A (naive) patrol schedule consists of a set of sequenced commands in the following form: at time t, the patrol unit should be at location l, and execute patrol action a. This patrol action a will take the unit to the next location and time if successfully executed. However, due to execution uncertainty, the patrol unit may end up at a different location and time. Figure 16.10 shows an example of execution uncertainty in a transition graph where if the patrol unit is currently at location A at the 5-min time step, she is supposed to take the on-train action to move to location B in the next time step. However, unlike CASS for ferry protection in which the defender's action is deterministic, there is a 10 % chance that she will still stay at location A due to execution uncertainty. These interactions of the defender with the environment when executing patrol can be represented as an MDP.

A key challenge of computing the SSE for this type of security problem is that the dimension of the space of mixed strategies for the defender is exponential in the number of states in terms of the defender's times and locations. Nevertheless, in many domains, the utilities have additional *separable* structure that the defender can exploit to efficiently compute an SSE of patrolling games with execution uncertainty. Specifically, when there exist *unit* utilities such as that both players' utilities can be represented as a linear combination of these *unit* utilities, the defender's Markov strategy can be obtained based on the marginal probabilities of

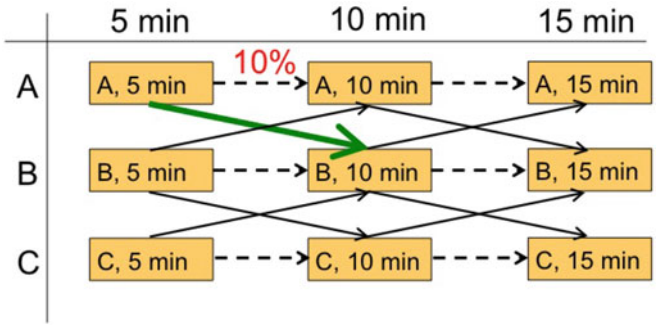


Fig. 16.10 An example of execution uncertainty in a transition graph

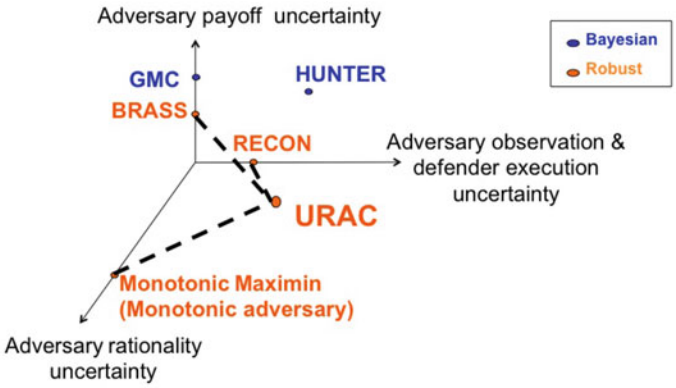


Fig. 16.11 Uncertainty space and algorithms

each patrolling unit reaching a state $s = (t, l)$, and taking action a. Here, the *unit* utilities only depend on a certain patrolling unit's state and action and a certain type of attacker. Therefore, instead of directly computing the mixed strategy, the defender attempts to compute the marginal probabilities which have dimensions polynomial in the sizes of the MDPs (the details of this approach are provided in [20]).

16.6.2 Security Patrolling with Unified Uncertainty Space

In this section, we present the two leading approaches for addressing uncertainty in security games in which the timing is not taken into account (which is different from the MDP-based approach described in the previous section). We first summarize the major types of uncertainties in security games as a three-dimensional uncertainty space with the following three dimensions (Fig. 16.11): (1) uncertainty in the adversary's payoff; (2) uncertainty related to the defender's strategy (including uncertainty in the defender's execution and the attacker's observation); and (3) uncertainty in the adversary's rationality. These dimensions refer to three key attributes which affect both players' utilities. The origin of the uncertainty space corresponds to the case with no uncertainty. Figure 16.11 also shows existing algorithms for addressing uncertainty in SSGs which follow the two main approaches: (1) modeling uncertainties based on Bayesian Stackelberg game models and (2) applying robust-optimization techniques. For example, BRASS [36] is a robust algorithm that only addresses attacker-payoff uncertainty while URAC (Unified Robust Algorithmic framework for addressing unCertainties) [33] is a unified robust algorithm that handles all types of uncertainty. In addition, HUNTER (Handling UNcerTainty Efficiently using Relaxation) [52] is a Bayesian-based algorithm that addresses all types of uncertainty except for the attacker-rationality uncertainty. While the Bayesian-based approach assumes a known distribution of uncertainties beforehand, the robust approach does not assume such prior knowledge.

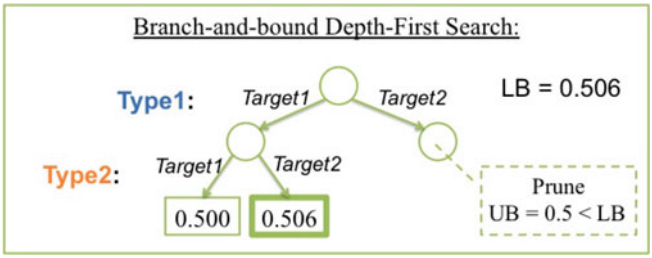


Fig. 16.12 Branch-and-bound depth first search

In the following, we will describe the two algorithms which are representatives of these two approaches: HUNTER (based on the Bayesian-based approach) and URAC (based on the robust approach).

16.6.2.1 Bayesian Approach

Overall, HUNTER is a novel algorithm for solving Bayesian Stackelberg games that can be used together with sample average approximation technique to solve Stackelberg games with uncertainty in the defender's execution and the attacker's observation [52]. Specifically, HUNTER attempts to compute the optimal mixed strategy for the defender against multiple attacker types with a prior distribution over the types. By exploiting the fact that the attacker is a perfectly rational player who will attack the optimal target with highest utility, HUNTER applies a best-first search for efficiently pruning the search tree that results from assigning attacker types to pure strategies as shown in Fig. 16.12. In other words, HUNTER first constructs the search tree by iteratively searching through all attacker types and all corresponding pure strategies for that attacker type. At each leaf node, the linear program at that node provides an optimal strategy for the defender such that the attacker's best response for every attacker type is the chosen target at that leaf node. Moreover, at internal nodes of the search tree (which corresponds to a partial assignment in which responses of a subset of attacker types are fixed), upper bounds and lower bounds of the optimal SSG solution are computed, which are then used to prune the search tree. As the size of the search tree is exponential in the number of targets and number of attacker types, finding tight upper bounds and lower bounds at internal nodes are essential in order to efficiently prune the search tree.

The key idea of HUNTER is to provide a tractable linear relaxation of Bayesian Stackelberg games that provides an upper bound efficiently at each of HUNTER's internal nodes based on finding a convex hull of all feasible solutions of the corresponding linear program at internal nodes. Figure 16.13 illustrates an example of constructing a convex hull of feasible solution regions of a two-target Bayesian security game with two attacker types. In Fig. 16.13a, each square corresponds to a partial assignment of an attacker type to a pure strategy, i.e., attacked target. The set

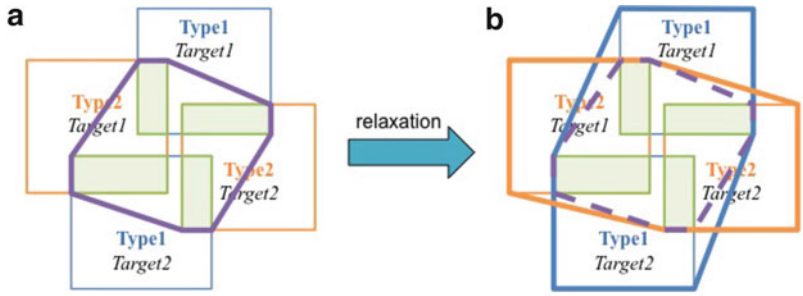


Fig. 16.13 An example of constructing a convex hull of feasible solution regions of a two-target Bayesian security games with two attacker types. HUNTER

of feasible solutions for the defender is the union of four disjoint green regions. As the optimal solution of a linear program is an extreme point of its feasible region, the linear program w.r.t the green regions is equivalent to a linear program with the same objective but w.r.t the convex hull of these four regions. However, the number of the disjoint regions is exponential in the number of targets and number of attacker types, finding a convex hull for these regions is computational. Therefore, HUNTER derives the relaxation of a Bayesian Stackelberg game by considering simpler convex hulls (of a small number of disjoint sets) (the blue and yellow regions shown in Fig. 16.13b) of which intersection is a super set of the convex hull of green regions. By solving this relaxation problem, HUNTER obtains an upper bound for the optimal solution of the Bayesian security game.

16.6.2.2 Robust Approach

In this section, we present the robust URAC algorithm for addressing a combination of all uncertainty types [33]. Consider an SSG where there is uncertainty in the attacker's payoff, the defender's strategy (including the defender's execution and the attacker's observation), and the attacker's behavior, URAC represents all these uncertainty types (except for the attacker's behaviors) using uncertainty interval. Instead of knowing exactly values of these game attributes, the defender only has prior information w.r.t the upper bounds and lower bounds of these attributes. For example, the attacker's reward if successfully attacking a target t is known to lie within the interval $[1, 3]$. Furthermore, URAC assumes the attacker monotonically responds to the defender's strategy. In other words, the higher the expected utility of a target, the more likely that the attacker will attack that target; however, the precise attacking probability is unknown for the defender. This monotonicity assumption is motivated by the Quantal Response model—a well-known human behavioral model for capturing the attacker's decision making [31].

Based on these uncertainty assumptions, URAC attempts to compute the optimal strategy for the defender by maximizing her utility against the worst-case scenario

of uncertainty. The key challenge of this optimization problem is that it involves several types of uncertainty, resulting in multiple minimization steps for determining the worst-case scenario. Nevertheless, URAC introduces a unified representation of all these uncertainty types as a uncertainty set of attacker's responses. Intuitively, despite of any type of uncertainty mentioned above, what finally affects the defender's utility is the attacker's response, which is unknown to the defender due to uncertainty. As a result, URAC can represent the robust-optimization problem as a single maximin problem.

However, the infinite uncertainty set of the attacker's responses depends on the planned mixed strategy for the defender, making this maximin problem difficult to solve if directly applying the traditional method (i.e., taking the dual maximization of the inner minimization of maximin and merging it with the outer maximization—maximin now can be represented a single maximization problem). Therefore, URAC proposes a divide-and-conquer method in which the defender's strategy set is divided into subsets such that the uncertainty set of the attacker's responses is the same for every defender strategy within each subset. This division leads to multiple sub-maximin problems which can be solved by using the traditional method. The optimal solution of the original maximin problem is now can be computed as a maximum over all the sub-maximin problems.

16.7 Current Research

In this section we highlight several areas that we are actively doing research on, and point out some of the open research challenges.

16.7.1 Scalability

Driven by the growing complexity of applications, a sequence of algorithms for solving security games have been developed including DOBSS [35], ERASER [25], ASPEN [16], and RUGGED [18]. However, existing algorithms still cannot scale up to very large-scale domains. While RUGGED/SNARES computes optimal solutions much faster than any of the previous approaches, much work remains to be done for it to be applicable to complex heterogenous settings on large networks.

Besides strategy generation, another approach for dealing with an exponential number of pure strategies is to compactly represent mixed strategies as marginal probabilities of coverage on each of the targets. Because of the utility structure of security games, such marginal probabilities are sufficient to express the expected utility of the defender. Kiekintveld et al. [25] used this approach in ERASER to formulate the problem of computing SSE as a compact mixed-integer linear program. However, this approach is unable to deal with complex constraints on the defender resources [26]. Nevertheless, we have recently been able to use this

approach for certain patrolling domains, including fare-enforcement patrols in urban transit systems [51] and boat patrols for protecting ferries [10]. In these domains a pure strategy is a patrol of a certain time duration over a set of locations, and the number of such pure strategies grow exponentially in the time duration. We were able to compactly represent mixed strategies as fractional flows on the *transition graph*, in which vertices are time–location pairs and arcs represent possible actions. This allowed us to formulate the optimization problems compactly which led to improved scalability. An open problem is to find other types of security domains in which the strategy space can be compactly represented. Another is to develop a hybrid approach that combines marginals and strategy generation.

16.7.2 Robustness

Classical game theory solution concepts often make assumptions on the knowledge, rationality, and capability (e.g., perfect recall) of players. Unfortunately, these assumptions could be wrong in real-world scenarios. Algorithms for the defender's optimal strategy have been proposed to take into account various uncertainties faced in the domain, including payoff noise [52], execution/observation error [50], and uncertain capability [1]. However, previous works assumed that the attacker knows (or with a small noise) the defender's mixed strategy. Recently An et al. [2] proposed a formal framework to model the attacker's belief update process as he observes instantiations of the defender's mixed strategy. The resulting optimization problem for the defender is non-linear and scalable computation remains an open issue. Furthermore, maximin is one of the leading robust method which is widely applied for addressing uncertainty in security games, which is known to be overly conservative. Minimax regret—an alternative less conservative robust criteria has just been applied recently to address payoff uncertainty [34]. The resulting optimization problem for using minimax regret is non-linear non-convex in both the defender strategy and the attacker's payoff and is thus computationally difficult. Moreover, addressing a combination of uncertainty using minimax regret has not been solved.

16.7.3 Adversary Modeling

One required research direction is addressing bounded rationality of human adversaries. This is a fundamental problem that can affect the performance of our game-theoretic solutions, since algorithms based on the assumption of the perfectly rational adversary are not robust to deal with deviations of the adversary from the optimal response. Recently, there has been some research on applying ideas from behavioral game theory (e.g., prospect theory [23] and quantal response [30]) within security game algorithms. One line of approaches is based on the quantal response model to predict the behaviors of the human adversary, and then to

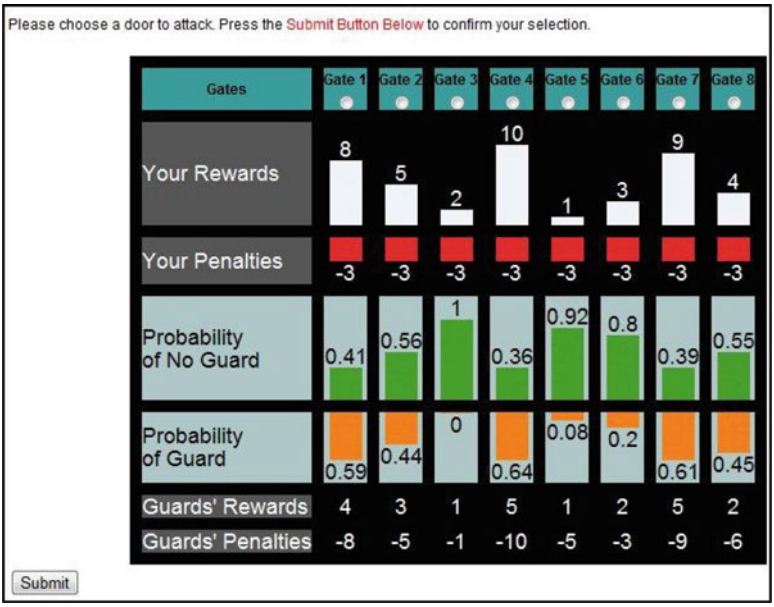


Fig. 16.14 Interface of the Guards and Treasures game to simulate the LAX security scenario

compute optimal defender strategies against such behavior of the adversary. These include BRQR [46] which follows the logit quantal response (QR) [30] model, and subsequent work on SUQR models [32]. The parameters of these models are estimated by experimental tuning. Figure 16.14 shows the interface of an interactive game used in our human subject experiments, based on the security scenario at the LAX airport. The source code is available here. [1] Given the details for each target, the participants playing this game were asked to choose a target to attack. Data from a large set of participants on the Amazon Mechanical Turk (AMT) were collected and used to learn the parameters of the behavioral models to predict future attacks.

Experiments with the Guards and Treasures game were conducted only as a single-shot game where the adversary would observe the defender's strategy and then choose a target to attack and then the game would be over. While this may be true for domains like counter-terrorism, in other real-world domains like fisheries protection, or wildlife crime, there are repeated interactions between the defender and the adversary, where the game progresses in "rounds." We call this a Repeated SSG (RSSG) where in each round the defender would play a particular strategy and the adversary would observe that strategy and act accordingly. In order to simulate this scenario and conduct experiments to identify adversary behavior in such repeated settings, an online RSSG game was developed (shown in Fig. 16.15) and deployed.

[1] http://teamcore.usc.edu/projects/BGT/experiment.html.

Fig. 16.15 Interface of the Wildlife Poaching game to simulate an RSSG

In our game, human subjects play the role of poachers looking to place a snare to hunt a hippopotamus in a protected wildlife park. The portion of the park shown in the map is actually a Google Maps view of a portion of the QENP in Uganda. The region shown is divided into a 5×5 grid, i.e., 25 distinct cells. Overlaid on the Google Maps view of the park is a heat-map, which represents the rangers' mixed strategy x—a cell i with higher coverage probability x_i is shown more in red, while a cell with lower coverage probability is shown more in green. As the subjects play the game and click on a particular region on the map, they were given detailed information about the poacher's reward, penalty, and coverage probability at that region. However, the participants are unaware of the exact location of the rangers while playing the game, i.e., they do not know the pure strategy that will be played by the rangers, which is drawn randomly from mixed strategy x shown on the game interface. In our game, there were nine rangers protecting this park, with each ranger protecting one grid cell. Therefore, at any point in time, only 9 out of the 25 distinct regions in the park are protected. A player succeeds if he places a snare in a region which is not protected by a ranger, else he is unsuccessful. Similar to the Guards and Treasures game, here also we recruited human subjects on AMT and asked them to play this game repeatedly for a set of rounds with the defender strategy changing per round based on the behavioral model being used to learn the adversary's behavior.

While behavioral models like (QR) [30] and SUQR [32] assume that there is a homogeneous population of adversaries, in the real-world we face heterogeneous populations of adversaries. Therefore Bayesian SUQR was proposed to learn the

behavioral model for each attack [49]. PAWS is an application which was originally created using Bayesian SUQR. However, in real-world security domains, we may have very limited data, or may only have some limited information on the biases displayed by adversaries. An alternative approach is based on robust optimization: instead of assuming a particular model of human decision making, try to achieve good defender expected utility against a range of possible models. One instance of this approach is MATCH [37], which guarantees a bound for the loss of the defender to be within a constant factor of the adversary loss if the adversary responds non-optimally. Another robust solution concept is monotonic maximin [21], which tries to optimize defender utility against the worst-case monotonic adversary behavior, where monotonicity is the property that actions with higher expected utility is played with higher probability. Recently, there has been attempts to combine such robust-optimization approaches with available behavior data [14] for RSSGs. However, an open question of research is how these proposed models and algorithms will fare against human subjects in RSSGs. Furthermore, since real-world human attackers are sometimes distributed coalitions of socially, culturally, and cognitively biased agents, we may need significant interdisciplinary research to build in social, cultural, and coalitional biases into our adversary models.

16.7.4 Multi-Objective Optimization

In existing applications such as ARMOR, IRIS, and PROTECT, the defender is trying to maximize a single objective. However, there are domains where the defender has to consider multiple objectives simultaneously. Multi-objective security games (MOSGs) have been proposed to address the challenges of domains with multiple incomparable objectives [7]. In an MOSG, the threats posed by the attacker types are treated as different objective functions which are not aggregated, thus eliminating the need for a probability distribution over attacker types. Unlike Bayesian security games which have a single optimal solution, MOSGs have a set of Pareto-optimal (non-dominated) solutions which is referred to as the Pareto frontier. By presenting the Pareto frontier to the end-user, they may be able to better understand the structure of their problem as well as the trade-offs between different security strategies.

16.7.5 Evaluations: Lab Evaluation via Simulation and Field Evaluation

Evaluation in itself is a major challenge given the real-world deployment of these systems. It is difficult to define a baseline for the purpose of evaluation in security applications, as safety often trumps costs. Our evaluation focuses on presenting the benefit of our approach over prior approaches to security. We have conducted a

Lab Evaluation	Field Evaluation: Patrol quality Unpredictable? Cover?	Field Evaluation: Tests against adversaries
Simulated adversary	Compare real schedules	"Mock attackers"
Human subject adversaries	Scheduling competition	Capture rates of real adversaries
	Expert evaluation	

Fig. 16.16 Field evaluation

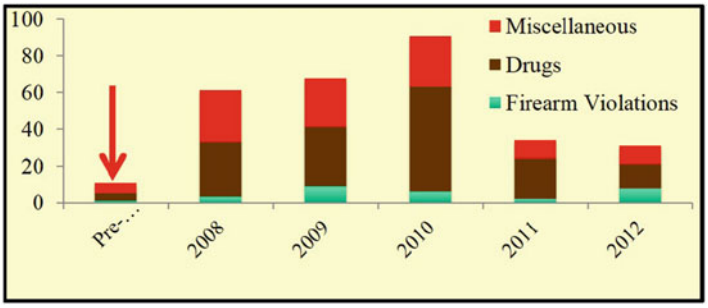


Fig. 16.17 ARMOR evaluation results

number of such evaluations: simulations, human subjects in the lab, assessment by domain experts internal and external to agencies deploying these applications, data from deployments (such as number of citations to fare-evaders), and adversary perspective teams (mock attacker teams) before and after deployment have all been used. We have already discussed simulations and human subject experiments in other parts of this chapter. Moreover, there are other evaluation approaches that we have tried, which are summarized in Fig. 16.16. In the following, we will discuss two of these approaches.

1. Data from deployment: data from the field, before and after deployment, supports our claim about improved security with our game-theoretic approach. Figure 16.17 shows the number of detected violations after ARMOR was deployed at LAX airport. As can be seen, the number of detected violations increased after our deployment and decreased in later years, suggesting better detection and deterrence effect of our approach. The patrol schedule for Boston port before and after our deployment of PROTECT (Fig. 16.18) clearly shows that there was a definite pattern in the patrols before PROTECT. In particular, there was low patrol for all targets on day 2, which could have been exploited by an attacker. In contrast, PROTECT provides almost the same level of patrol every day, with higher value targets patrol more often.

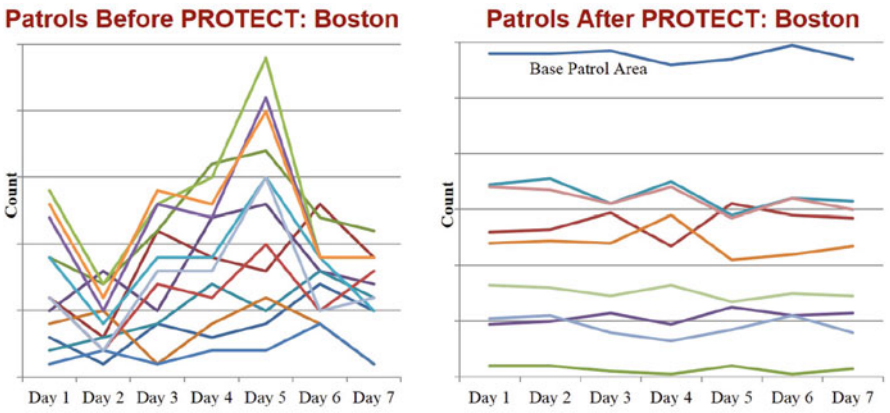


Fig. 16.18 PROTECT evaluation results: pre deployment (*left*) and post deployment patrols (*right*)

2. Mock attacker team: The USCG created an APT, a mock attacker team, to better understand the adversaries view of targets in the Boston port. This team, in addition to understanding the adversary's viewpoint, also gauged the effectiveness of patrol activities before and after deployment of PROTECT. The APT incorporates the adversary's known intent, capabilities, skills, commitment, resources, and cultural influences. In addition, it helps in identifying the level of deterrence projected at and perceived by the adversary. This analysis led to the conclusion that the effectiveness of deterrence increased from the before to after PROTECT deployment.

More detailed evaluations are discussed in the publications on the applications [11, 17, 39, 51], and more of these are discussed in [40].

16.8 Conclusion

Security is recognized as a world-wide challenge and game theory is an increasingly important paradigm for reasoning about complex security resource allocation. While the deployed game-theoretic applications have provided a promising start, very significant amount of research remains to be done. These are large-scale interdisciplinary research challenges that call upon multiagent researchers to work with researchers in other disciplines, be "on the ground" with domain experts, and examine real-world constraints and challenges that cannot be abstracted away.

References

1. An, B., Tambe, M., Ordonez, F., Shieh, E., Kiekintveld, C.: Refinement of strong Stackelberg equilibria in security games. In: Proceedings of the 25th Conference on Artificial Intelligence, pp. 587–593 (2011)
2. An, B., Kempe, D., Kiekintveld, C., Shieh, E., Singh, S., Tambe, M., Vorobeychik, Y.: Security games with limited surveillance. In: Conference on Artificial Intelligence (AAAI) (2012)
3. Avenhaus, R., von Stengel, B., Zamir, S.: Inspection games. In: Aumann, R.J., Hart, S. (eds.) Handbook of Game Theory, vol. 3, Chap. 51, pp. 1947–1987. North-Holland, Amsterdam (2002)
4. Breton, M., Alg, A., Haurie, A.: Sequential Stackelberg equilibria in two-person games. J. Optim. Theory Appl. **59**(1), 71–97 (1988)
5. Brown, G., Carlyle, M., Kline, J., Wood, K.: A two-sided optimization for theater ballistic missile defense. Oper. Res. **53**, 263–275 (2005)
6. Brown, G., Carlyle, M., Salmeron, J., Wood, K.: Defending critical infrastructure. Interfaces. **36**, 530–544 (2006)
7. Brown, M., An, B., Kiekintveld, C., Ordonez, F., Tambe, M.: Multi-objective optimization for security games. In: Proceedings of The 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2012)
8. Chandran, R., Beitchman, G.: Battle for Mumbai ends, death toll rises to 195. Times of India (29 November 2008). http://articles.timesofindia.indiatimes.com/2008-11-29/india/27930171_1_taj-hotel-three-terrorists-nariman-house
9. Conitzer, V., Sandholm, T.: Computing the optimal strategy to commit to. In: Proceedings of the ACM Conference on Electronic Commerce (ACM-EC), pp. 82–90 (2006)
10. Fang, F., Jiang, A., Tambe, M.: Optimal patrol strategy for protecting moving targets with multiple mobile resources. In: AAMAS (2013)
11. Fave, F.M.D., Brown, M., Zhang, C., Shieh, E., Jiang, A.X., Rosoff, H., Tambe, M., Sullivan, J.: Security games in the field: an initial study on a transit system(extended abstract). In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS) [Short paper] (2014)
12. Gatti, N.: Game theoretical insights in strategic patrolling: model and algorithm in normal-form. In: ECAI-08, pp. 403–407 (2008)
13. Hamilton, B.A.: Faregating Analysis. Report Commissioned by the LA Metro (2007). http://boardarchives.metro.net/Items/2007/11_November/20071115EMACItem27.pdf
14. Haskell, W.B., Kar, D., Fang, F., Tambe, M., Cheung, S., Denicola, L.E.: Robust protection of fisheries with compass. In: Proceedings of the 28th AAAI Conference on Artificial Intelligence, pp. 2978–2983 (2014)
15. Howard, N.J.: Finding optimal strategies for influencing social networks in two player games. Master's thesis, MIT, Sloan School of Management (2011)

16. Jain, M., Kardes, E., Kiekintveld, C., Ordonez, F., Tambe, M.: Security games with arbitrary schedules: a branch and price approach. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 792–797 (2010)
17. Jain, M., Tsai, J., Pita, J., Kiekintveld, C., Rathi, S., Tambe, M., Ordonez, F.: Software assistants for randomized patrol planning for the LAX airport police and the federal air marshal service. Interfaces **40**, 267–290 (2010)
18. Jain, M., Korzhyk, D., Vanek, O., Pechoucek, M., Conitzer, V., Tambe, M.: A double oracle algorithm for zero-sum security games on graphs. In: Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2011)
19. Jain, M., Tambe, M., Conitzer, V.: Security scheduling for real-world networks. In: AAMAS (2013)
20. Jiang, A., Yin, Z., Kraus, S., Zhang, C., Tambe, M.: Game-theoretic randomization for security patrolling with dynamic execution uncertainty. In: AAMAS (2013)
21. Jiang, A.X., Nguyen, T.H., Tambe, M., Procaccia, A.D.: Monotonic maximin: a robust Stackelberg solution against boundedly rational followers. In: Conference on Decision and Game Theory for Security (GameSec) (2013)
22. Johnson, M., Fang, F., Yang, R., Tambe, M., Albers, H.: Patrolling to maximize pristine forest area. In: Proceedings of the AAAI Spring Symposium on Game Theory for Security, Sustainability and Health (2012)
23. Kahneman, D., Tvesky, A.: Prospect theory: an analysis of decision under risk. Econometrica **47**(2), 263–291 (1979)
24. Keteyian, A.: TSA: Federal Air Marshals. http://www.cbsnews.com/stories/2010/02/01/earlyshow/main6162291.shtml (2010). Retrieved 1 Feb 2011
25. Kiekintveld, C., Jain, M., Tsai, J., Pita, J., Tambe, M., Ordonez, F.: Computing optimal randomized resource allocations for massive security games. In: Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 689–696 (2009)
26. Korzhyk, D., Conitzer, V., Parr, R.: Complexity of computing optimal Stackelberg strategies in security resource allocation games. In: Proceedings of the 24th AAAI Conference on Artificial Intelligence, pp. 805–810 (2010)
27. Leitmann, G.: On generalized Stackelberg strategies. J. Optim. Theory Appl. **26**(4), 637–643 (1978)
28. Lipton, R., Markakis, E., Mehta, A.: Playing large games using simple strategies. In: EC: Proceedings of the ACM Conference on Electronic Commerce, pp. 36–41. ACM, New York, NY (2003)
29. Lye, K., Wing, J.M.: Game strategies in network security. Int. J. Inf. Secur. **4**(1–2), 71–86 (2005)
30. McFadden, D.: Quantal choice analysis: a survey. Ann. Econ. Soc. Meas. **5**(4), 363–390 (1976)
31. McKelvey, R.D., Palfrey, T.R.: Quantal response equilibria for normal form games. Game Econ. Behav. **10**(1), 6–38 (1995)
32. Nguyen, T.H., Yang, R., Azaria, A., Kraus, S., Tambe, M.: Analyzing the effectiveness of adversary modeling in security games. In: Conference on Artificial Intelligence (AAAI) (2013)
33. Nguyen, T., Jiang, A., Tambe, M.: Stop the compartmentalization: unified robust algorithms for handling uncertainties in security games. In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2014)
34. Nguyen, T.H., Yadav, A., An, B., Tambe, M., Boutilier, C.: Regret-based optimization and preference elicitation for Stackelberg security games with uncertainty. In: Proceedings of the National Conference on Artificial Intelligence (AAAI) (2014)

35. Paruchuri, P., Pearce, J.P., Marecki, J., Tambe, M., Ordonez, F., Kraus, S.: Playing games with security: an efficient exact algorithm for Bayesian Stackelberg games. In: Proceedings of the 7th International Conference on Autonomous Agents and Multiagent Systems (AAMAS), pp. 895–902 (2008)

36. Pita, J., Jain, M., Ordóñez, F., Tambe, M., Kraus, S., Magori-Cohen, R.: Effective solutions for real-world Stackelberg games: when agents must deal with human uncertainties. In: The Eighth International Conference on Autonomous Agents and Multiagent Systems (2009)

37. Pita, J., John, R., Maheswaran, R., Tambe, M., Kraus, S.: A robust approach to addressing human adversaries in security games. In: European Conference on Artificial Intelligence (ECAI) (2012)

38. Arce, D.G., Sandler, T.: Terrorism and game theory. Simul. Gaming **34**(3), 319–337 (2003)

39. Shieh, E., An, B., Yang, R., Tambe, M., Baldwin, C., DiRenzo, J., Maule, B., Meyer, G.: PROTECT: a deployed game theoretic system to protect the ports of the United States. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2012)

40. Taylor, M., Kiekintveld, C., Tambe, M.: Evaluating deployed decision-support systems for security: challenges, analysis, and approaches. In: Tambe, M. (ed.) Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned. Cambridge University Press, Cambridge (2011)

41. Tsai, J., Nguyen, T.H., Tambe, M.: Security games for controlling contagion. In: Conference on Artificial Intelligence (AAAI) (2012)

42. Tsai, J., Qian, Y., Vorobeychik, Y., Kiekintveld, C., Tambe, M.: Bayesian security games for controlling contagion. In: Proceedings of the ASE/IEEE International Conference on Social Computing(SocialCom) (2013)

43. Vanek, O., Yin, Z., Jain, M., Bosansky, B., Tambe, M., Pechoucek, M.: Game-theoretic resource allocation for malicious packet detection in computer networks. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2012)

44. von Stackelberg, H.: Marktform und Gleichgewicht. Springer, Vienna (1934)

45. von Stengel, B., Zamir, S.: Leadership with commitment to mixed strategies. Technical Report, LSE-CDAM-2004-01, CDAM Research Report (2004)

46. Yang, R., Kiekintveld, C., Ordonez, F., Tambe, M., John, R.: Improving resource allocation strategy against human adversaries in security games. In: IJCAI (2011)

47. Yang, R., Ordonez, F., Tambe, M.: Computing optimal strategy against quantal response in security games. In: Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, vol. 2, pp. 847–854 (2012)

48. Yang, R., Jiang, A.X., Tambe, M., Ordóñez, F.: Scaling-up security games with boundedly rational adversaries: a cutting-plane approach. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, pp. 404–410. AAAI Press, Menlo Park (2013)

49. Yang, R., Ford, B., Tambe, M., Lemieux, A.: Adaptive resource allocation for wildlife protection against illegal poachers. In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2014)

50. Yin, Z., Jain, M., Tambe, M., Ordonez, F.: Risk-averse strategies for security games with execution and observational uncertainty. In: Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI), pp. 758–763 (2011)

51. Yin, Z., Jiang, A., Johnson, M., Tambe, M., Kiekintveld, C., Leyton-Brown, K., Sandholm, T., Sullivan, J.: TRUSTS: scheduling randomized patrols for fare inspection in transit systems. In: Proceedings of the 24th Conference on Innovative Applications of Artificial Intelligence (IAAI) (2012)

52. Yin, Z., Tambe, M.: A unified method for handling discrete and continuous uncertainty in Bayesian Stackelberg games. In: International Conference on Autonomous Agents and Multiagent Systems (AAMAS) (2012)
53. Zhang, C., Jiang, A.X., Short, M.B., Brantingham, P.J., Tambe, M.: Modeling crime diffusion and crime suppression on transportation networks: an initial report. In: SNSC 2013: The AAAI Fall Symposium 2013 on Social Networks and Social Contagion (2013)

Chapter 17
Scattering of Plane Electromagnetic Waves by Radially Inhomogeneous Spheres: Asymptotics and Special Functions

Michael A. Pohrivchak, John A. Adam, and Umaporn Nuntaplook

Abstract A brief historical introduction to the visual and wave-theoretic consequences of high frequency electromagnetic scattering by large spheres is given, with special emphasis on backscattering. Exact electromagnetic solutions for radially inhomogeneous dielectric lenses are unavailable for many functional dependences of the refractive index on the radial distance, so the high-frequency behavior based on an asymptotic analysis of the exact solution has been obtained in very few cases. In this chapter existing results for the asymptotic behavior of backscattered radiation are extended to a broader class of refractive index profiles. Additionally, by exploiting some known results from quantum mechanics, asymptotic solutions for two scalar problems (decoupled from the electromagnetic cases) are derived for the case of small variations in the refractive index across the scattering sphere. By using a Liouville transformation the electromagnetic wavenumber-dependent scattering potential is converted to a wavenumber-independent form, and the resulting inverse problem is solved for several refractive index profiles.

Keywords EM scattering • TE/TM modes • Rainbow • Watson transform • Back-scattering • Asymptotic expansions • Bessel's equation • Whittaker's equation • Hypergeometric equation • Radial Debye potentials • Refractive index profiles • Scattering potentials • Riccati-Bessel functions • Born approximation • Liouville transformation • Maxwell Fish-Eye

M.A. Pohrivchak
The Naval Research Laboratory, Washington, DC, USA
e-mail: mporkchop@gmail.com

J.A. Adam (✉)
Department of Mathematics & Statistics, Old Dominion University, Norfolk, VA, USA
e-mail: jadam@odu.edu

U. Nuntaplook
Department of Mathematics, Faculty of Science, Mahidol University, Bangkok, Thailand
e-mail: ununtaploook@gmail.com

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_17

17.1 Historical Introduction

(i) General

When light (or electromagnetic radiation in general) is deflected in some manner from its direction of travel, it is said to be scattered. There are several mechanisms that contribute to the scattering of light by particles in the atmosphere: reflection, refraction, and diffraction being the most common, though they are not mutually exclusive effects. The size of the particles determines which mechanism is the predominant one. The optical aspects are complicated because of the range of particle sizes compared with the wavelengths of visible light: approximately 400–700 nm. A convenient measure of relative size is the radius-to-wavelength ratio. When this ratio is at least about ten, the particles are considered to be large, and it is convenient to regard light in terms of rays. This is the domain of geometrical optics, although all three processes mentioned above can occur. Light rays can be partially reflected from the surface of the particles, refracted on passing through its interior, or diffracted ("bent") around the edges. All three mechanisms are exhibited in the phenomenon of the rainbow so we shall start this review by discussing rainbow formation. It is well known that light is refracted and reflected by raindrops to produce this beautiful colored arc in the sky. Less familiar is the third important mechanism—diffraction—a consequence of the wavelike properties of light. This is responsible for some of the more subtle rainbow features—pale fringes below the top of the bow (called *supernumerary bows*), and also iridescence in clouds near the sun, and concentric colored rings around the moon. To understand the phenomenon of scattering from a more analytic point of view we need to recall some basic physics. An electromagnetic wave has, not surprisingly, both an electric and magnetic field that are functions of time and space as it propagates. The direction of propagation and the directions of these fields form a mutually orthogonal triad, and when an electromagnetic field encounters an electron bound to a molecule, the electron is accelerated by the electric field of the wave. It's a type of "chicken and egg" situation, because an accelerated electron will also radiate electromagnetic energy in the form of waves in all directions (to some extent), and this is the scattered radiation, components of which will be discussed in this chapter.

It has been said that "Descartes knew where to hang the rainbow in the sky, but only Newton could paint it." But by the middle of the eighteenth century, the contributions of Descartes and Newton notwithstanding, observations of supernumerary bows were a persistent reminder of the inadequacy of current theories of the rainbow. By focusing attention on the light *wavefronts* incident on a spherical drop, rather than the rays normal to them, it is easier to appreciate the self-interference of such a wave as it becomes "folded" onto itself as a result of refraction and reflection within the drop, the true extent of the rainbow is revealed. The primary rainbow is in fact the *first interference maximum* in an oscillatory pattern, the second and third maxima being the first and

second supernumerary bows, respectively (and so on). The angular spacing of these bands depends on the size of the droplets producing them. The width of individual bands and the spacing between them decrease as the drops get larger. If drops of many different sizes are present, these supernumerary arcs tend to overlap somewhat and smear out what would have been obvious interference bands for droplets of uniform size. This is why these pale blue or pink or green bands are then most noticeable near the top of the rainbow: it is the near-sphericity of the smaller drops that enable them to contribute to this part of the bow; larger drops are distorted from sphericity by the aerodynamic forces acting upon them. Nearer the horizon a wide range of drop size contributes to the bow, but at the same time it tends to blur the interference bands. In principle, similar interference effects also occur above the secondary rainbow, though they are very rare. "Thus the supernumerary rainbows proved to be the midwife that delivered the wave theory of light to its place of dominance in the nineteenth century" [1].

It is important to recognize that not only were the Cartesian and Newtonian theories unable to account for the presence of supernumerary bows, but also they both predicted an abrupt transition between regions of illumination and shadow (as at the edges of Alexander's dark band, when rays only giving rise to the primary and secondary bows are considered). In the wave theory of light such sharp boundaries are softened by diffraction, which occurs when the normal interference pattern responsible for rectilinear propagation of light is distorted in some way. Diffraction effects are particularly prevalent in the vicinity of caustics. In 1835 Potter showed that the rainbow ray may be interpreted as a caustic, i.e. the envelope of the system of rays constituting the rainbow. The word caustic means "burning," and caustics are associated with regions of high intensity illumination (with geometrical optics predicting an infinite intensity there). Thus the rainbow problem is essentially that of determining the intensity of (scattered) light in the neighborhood of a caustic. This was exactly what Airy attempted to do several years later in 1838. The principle behind Airy's approach was established by Huygens in the seventeenth century: Huygens' principle regards every point of a wavefront as a secondary source of waves, which in turn defines a new wavefront and hence determines the subsequent propagation of the wave. Airy reasoned that if one knew the amplitude distribution of the waves along any complete wavefront in a raindrop, the distribution at any other point could be determined by Huygens' principle. Using the standard assumptions of diffraction theory, he formulated the local intensity of scattered light in terms of a "rainbow integral," subsequently renamed the Airy integral in his honor; it is related to the now familiar Airy function. It is analogous to the Fresnel integrals which also arise in diffraction theory. There is a natural and fundamental parameter, the size parameter, β, which is useful in determining the domain of validity of the Airy approximation; it is defined as the ratio of the droplet circumference to the wavelength λ of light. In terms of the wavenumber k this is $\beta = 2\pi r/\lambda$, r being the droplet radius. Typically, for sizes ranging from fog droplets to large

raindrops, β ranges from about 100 to several thousand. Airy's approximation is a good one only for $\beta \gtrsim 5000$ and angles sufficiently close to that of the rainbow ray (the ray of minimum deviation from the direction of incidence). On the other hand, the "why is the sky blue?" scattering problem—Rayleigh scattering—requires only one term because the scatterers are molecules which are much smaller than a wavelength of light, so the simplest truncation—retaining only the first term—is perfectly adequate.

Airy theory has been called the incomplete "complete" answer. It did go beyond the models of the day in that it quantified the dependence upon the raindrop size of (1) the rainbow's angular width, (2) its angular radius, and (3) the spacing of the supernumerary bows. Also, unlike the models of Descartes and Newton, Airy's predicted a non-zero distribution of light intensity in Alexander's dark band (the darker region between the primary and secondary bows), and a finite intensity at the angle of minimum deviation (as noted above, the earlier theories predicted an infinite intensity there). However, spurred on by Maxwell's recognition that light is part of the electromagnetic spectrum, and the subsequent publication of his mathematical treatise on electromagnetic waves, several mathematical physicists sought a more complete theory of scattering, because it had been demonstrated by then that the Airy theory failed to predict precisely the angular position of many laboratory-generated rainbows. Among them were the German physicist Gustav Mie who published a paper in 1908 on the scattering of light by homogeneous spheres in a homogeneous medium, and Peter Debye who independently developed a similar theory for the scattering of electromagnetic waves by spheres. Mie's theory was intended to explain the colors exhibited by colloidally dispersed metal particles, whereas Debye's work, based on his 1908 thesis, dealt with the problem of light pressure on a spherical particle. In fact, Ludvig Lorenz, a Danish theorist, preceded Mie by about 15 years in the treatment of the scattering of electromagnetic waves by spheres. The resulting body of knowledge is usually referred to as Mie theory, and typical computations based on it are formidable compared with those based on Airy theory, unless the drop size is sufficiently small. A similar (but scalar) formulation arises in the scattering of sound waves by an impenetrable sphere, studied by Lord Rayleigh and others in the nineteenth century. Mie theory is based on the solution of Maxwell's equations of electromagnetic theory for a monochromatic plane wave from infinity incident upon a homogeneous isotropic sphere of radius r. The surrounding medium is transparent (as the sphere may be), homogeneous, and isotropic. The incident wave induces forced oscillations of both free and bound charges in synchrony with the applied field, and this induces a secondary electric and magnetic field, each of which has components inside and outside the sphere. Of crucial importance in the theory are what are termed the scattering amplitudes for the two independent polarizations, θ being the angular variable; these amplitudes can be expressed as an infinite sum called a partial wave expansion. Each term (or "partial wave") in the expansion is defined in terms of combinations of Legendre functions of the first kind, Riccati–Bessel and Riccati–Hankel functions [2, 3].

Although in principle the rainbow problem can be "solved" with enough computer time and resources, numerical solutions by themselves offer little or no insight into the physics of the phenomenon. However, there was a significant mathematical development in the early twentieth Century that eventually had a profound impact on the study of scalar and vector scattering: The Watson transform, originally introduced in 1918 by Watson in connection with the diffraction of radio waves around the earth [4] [1] (and subsequently modified by several mathematical physicists in studies of the rainbow problem), is a method for transforming the slowly converging partial-wave series into a rapidly convergent expression involving an integral in the complex angular-momentum plane. This allows the above transformation to effectively "redistribute" the contributions to the partial wave series into a few points in the complex plane—specifically poles (called Regge poles in elementary particle physics) and saddle-points. Such a decomposition means that instead of identifying angular momentum with certain discrete real numbers, it is now permitted to move continuously through complex values. However, despite this modification, the poles and saddle points have profound physical interpretations in the rainbow problem ([5] and the references therein).

(ii) The Backscattering Problem

The backscattering problem refers to a special case of the Mie solution for which the radiation is scattered in a direction 180° from the direction of the incident field. Kerker [6] gave a detailed account of backscattering from dielectric spheres (including coated spheres). The topic is naturally of fundamental importance in radar techniques. Targets include cloud droplets, rain, snow, hail, flocks of birds or insections, weather formations (e.g., thunderstorms, tornados), aircraft, satellites, even the moon and planets. The following quote from a paper by Inada and Plonus [7] (see also [8]) is quite illuminating (citations for the papers by Rubinow and Nussenzveig have been added):

> The exact solution to the scattering of a plane electromagnetic wave by a dielectric sphere was obtained by Mie in 1908. The Mie solution...given in the form of an infinite series, has a limitation in that it converges very slowly when the radius of the sphere exceeds a few wavelengths. This difficulty was overcome by Watson...in 1918 for the problem of wave propagation around the earth. The method is to transform the slowly convergent Mie series into a rapidly convergent series. This treatment is known as the Watson transformation. Unlike the problem of the perfectly conducting sphere..., the scattering problem is not as well understood. The methods of correcting geometrical optics for the perfectly conducting sphere are not applicable in the case of dielectric (or penetrable) spheres. The problem of backscattering is further complicated by the waves existing inside the sphere which could contribute significantly. Among several studies on dielectric spheres, Rubinow [9]... and Nussenzveig [10]...investigated the problem of scattering of a scalar wave from a penetrable sphere at high frequencies. The scalar case has many common features with ... quantum mechanics problems... We know that by applying the Watson transformation to the exact Mie series for a perfectly conducting sphere, the scattered fields are given as a sum of optics and residue contributions.

[1] It is interesting to note that Watson mentions possible communication with inhabitants of Mars!

> The question that now arises [for a dielectric sphere] is the following: How do the residue contributions affect the backscattering fields?. . .[They] are physically connected with two different types of surface waves; one type is a "creeping wave" analogous to that of a perfectly conducting sphere, which encircles the dielectric sphere, and the other type is a wave which enters the sphere and then emerges as a surface wave. The latter is unique to a dielectric sphere. . .

One of the most mathematically sophisticated studies of backscattering (for a specific power-law-dependent class of dielectrics) can be found in the 1974 Ph.D. dissertation of Brockman [11]. He investigated high frequency far field backscattering of a plane time harmonic monochromatic electromagnetic wave by a class of radially inhomogeneous spheres (this is also known as the high-frequency backscattered field). He applied a Watson transformation on the high frequency exact solution, thereby converting the Mie series to a contour integration in the complex frequency plane. By deforming the contour, various contributions were extracted from the segments on the contour, including residue series that converge rapidly at high frequencies, as well as several line integrations. However, Brockman found it helpful to examine the geometrical optics regime as well (see also [12]). This enabled him to identify possible ray contributions to backscattering, such as the front axial directly reflected ray, the rear axial ray, glory rays (rays entering the sphere that exit in the backscatter direction without experiencing any surface wave behavior), and the backscattered rainbow ray (or stationary glory ray). In addition to these, Brockman applied the ray technique to the creeping waves (these travel along the outer surface of the sphere), whispering gallery modes (traveling around the inside surface of the sphere) and what he referred to as partial surface waves: a hybrid combination of geometric optics rays and creeping waves.

17.2 Theory

The main goal in this section is to determine the leading order estimate of the far backscattered electromagnetic field at short wavelengths for a rather more general class of refractive index profiles than was considered in the seminal (but very succinct) paper by Uslenghi and Weston [13]. The mathematical details they provided were very sparse, and required a great deal of effort to derive [14], so in and of itself the authors believe this is a significant contribution to the subject, especially since the available class of refractive index profiles to which this is applicable has been expanded considerably. The far backscattered field is given by an infinite series which converges slowly at short wavelengths. The Watson transformation will be employed to speed up the convergence of this series by converting the series to a contour integral. Once this is done, the radial eigenfunctions will be derived for fields of magnetic- and electric-type. These eigenfunctions are necessary in order to calculate the asymptotic expansions for the transverse electric (TE) and transverse magnetic (TM) modes. Once these expansions are obtained, the Mie solutions [15] will be calculated which will allow for the determination of the high-frequency backscattered field. Consider, then an incident plane electromagnetic

wave propagating in the positive z-direction with the free space wavenumber k, whose electric vector

$$\mathbf{E}^{inc.} = \tilde{\mathbf{e}} e^{ikz} \tag{17.1}$$

has unit amplitude and is polarized in the direction of the constant unit vector $\tilde{\mathbf{e}}$. We note that $k = 2\pi/\lambda$, where λ is the wavelength of the incident plane wave given by Eq. (17.1). After interacting with the scattering particle (a sphere of radius $\hat{a}$) it produces the far backscattered field (which corresponds to a linear combination of outgoing spherical waves) [13]

$$\mathbf{E}^{b.s.} = \tilde{\mathbf{e}} \frac{e^{ikr}}{ikr} \sum_{n=1}^{\infty} (-1)^n \left(n + \frac{1}{2}\right)(a_n - b_n), \tag{17.2}$$

where

$$a_n = -\frac{\psi_n'(k\hat{a}) - M_n \psi_n(k\hat{a})}{\zeta_n'(k\hat{a}) - M_n \zeta_n(k\hat{a})}, \tag{17.3}$$

$$b_n = -\frac{\psi_n'(k\hat{a}) - \tilde{M}_n \psi_n(k\hat{a})}{\zeta_n'(k\hat{a}) - \tilde{M}_n \zeta_n(k\hat{a})}, \tag{17.4}$$

$$\psi_n(k\hat{a}) = \sqrt{\frac{\pi k\hat{a}}{2}} J_{n+\frac{1}{2}}(k\hat{a}), \qquad \zeta_n(k\hat{a}) = \sqrt{\frac{\pi k\hat{a}}{2}} H^{(1)}_{n+\frac{1}{2}}(k\hat{a}), \tag{17.5}$$

and

$$M_n = \frac{1}{k\hat{a}} \left[\frac{S_n^{(1)'}(x)}{S_n^{(1)}(x)}\right]_{x=1}, \tag{17.6a}$$

$$\tilde{M}_n = \frac{1}{k\hat{a}} \left[\frac{T_n^{(1)'}(x)}{T_n^{(1)}(x)}\right]_{x=1}, \tag{17.6b}$$

where the prime indicates differentiation with respect to the argument. The functions $\psi_n(k\hat{a})$ and $\zeta_n(k\hat{a})$ are the Riccati–Bessel functions. It should be noted that the subscript n appearing in the above equations is not the refractive index profile; it represents the separation constant. The refractive index profile will be denoted by the function $R(x)$, where $x = r/\hat{a}$ is the radial distance from the center $r = 0$ of the sphere, normalized to the radius $\hat{a}$ of the sphere. The functions M_n and $\tilde{M}_n$ are known as the transverse electric (TE) and transverse magnetic (TM) modes, respectively. With respect to the TE/TM modes, there is no electric/magnetic field in the direction of wave propagation, respectively. As a result, the functions $S_n(x)$ appearing in Eq. (17.6a) are known as the radial eigenfunctions for fields of magnetic-type. Similarly, the functions $T_n(x)$ appearing in Eq. (17.6b) are known as the radial eigenfunctions for fields of electric-type. In order for the leading order estimate of the far backscattered field to be calculated for short wavelengths, the Mie coefficients which are given by Eqs. (17.3) and (17.4) must be determined. Before this can be accomplished, the asymptotic expansions for the TE and TM modes

must be calculated using Eqs. (17.6a) and (17.6b). This requires the determination of the radial eigenfunctions for fields of magnetic- and electric-type, respectively. The radial eigenfunctions $S_n^{(1)}(x)$ and $T_n^{(1)}(x)$ are those particular solutions of the radial differential equations

$$S_n''(x) + \left\{ [k\hat{a}R(x)]^2 - \frac{n(n+1)}{x^2} \right\} S_n(x) = 0 \tag{17.7}$$

and

$$T_n''(x) - 2\frac{R'(x)}{R(x)} T_n'(x) + \left\{ [k\hat{a}R(x)]^2 - \frac{n(n+1)}{x^2} \right\} T_n(x) = 0. \tag{17.8}$$

Note that the radial eigenfunctions will be required to be finite over the interval $0 \leq x \leq 1$ for the refractive index profiles considered here.

17.2.1 Profile 1

Consider first the refractive index profile given by

$$R(x) = \frac{c_0 x^{(\alpha/2)-1}}{1 + x^{\alpha}}. \tag{17.9}$$

17.3 Converting the Sum to a Contour Integral

The high-frequency backscattered field that is given by Eq. (17.2) converges extremely slowly in the limit $k\hat{a} \to \infty$, in other words, for short wavelengths. As a result, we will utilize the Watson transformation which replaces a slowly converging series with a contour integral. This integral converges at a much faster rate than the series. Let $d_0 = 2/\alpha$. If we consider $\nu = n + 1/2$ as a complex number, then, using the definition in Eq. (17.10), which will be required later,

$$\gamma_{\pm} = \frac{1}{2}\sqrt{1 \pm \frac{2}{\alpha} + \frac{4}{\alpha^2}\nu^2}, \tag{17.10}$$

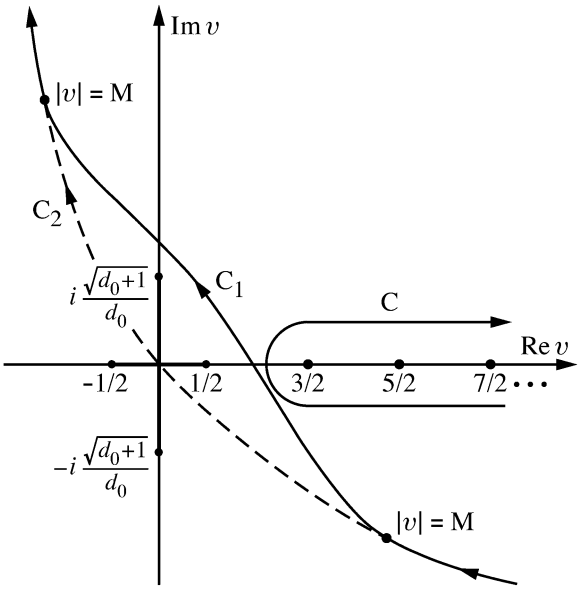
we see that γ_- has branch points at

$$\nu = \pm\frac{1}{d_0}\sqrt{d_0 - 1} \tag{17.11}$$

and γ_+ at

$$\nu = \pm i\frac{\sqrt{d_0 + 1}}{d_0}. \tag{17.12}$$

Fig. 17.1



As a result of the location of the poles of the integrand, we choose the branch cuts in the complex ν-plane along the real axis between $-1/2$ and $+1/2$. Along the imaginary axis, we choose the branch cuts between

$$-i\frac{\sqrt{d_0+1}}{d_0} \text{and} + i\frac{\sqrt{d_0+1}}{d_0}.$$

Now we will replace the summation in Eq. (17.2) with a line integral taken along the clockwise contour C of Fig. 17.1, which encloses those poles of the integrand that are located at $\nu = p + 1/2$, where p is a positive integer.

By following a transformation of the type of Watson's, the line integral along C is replaced by the sum of:

- A line integral whose contour consists of a path C_1 extending from the fourth through the first to the second quadrant, plus the arc of a circle of large radius with center at $\nu = 0$ extending from the second through the first to the fourth quadrant, and
- A residue series due to the poles of the integrand which lie in the first quadrant. The contour C_1 crosses the real ν-axis between $1/2$ and $3/2$ and the imaginary ν-axis above $+i\left(\sqrt{d_0+1}\right)/d_0$, avoiding the branch cuts. The result obtained thus far is still exact. Hence, upon using the Watson transformation, we determine that the high-frequency backscattered field is given by

$$\mathbf{E}^{b.s.} \sim -\tilde{\mathbf{e}}\frac{e^{ikr}}{2kr}\int_{C_1}\frac{\nu}{\cos\pi\nu}\left(a_{\nu-\frac{1}{2}} - b_{\nu-\frac{1}{2}}\right)d\nu, \qquad (d_0 k\hat{a} \gg 1). \quad (17.13)$$

Figure 17.1 illustrates the branch cuts and the contours that will be used to evaluate Eq. (17.13). The quantity $(a_{\nu-\frac{1}{2}} - b_{\nu-\frac{1}{2}})$ in Eq. (17.13) must now be evaluated for $|\nu| = O\left((k\hat{a})^{1/2+\epsilon}\right)$ where ϵ is an arbitrarily small positive number. Before this can be accomplished, the radial eigenfunctions for fields of magnetic- and electric-type have to be computed. Next the asymptotic expansions for the TE and TM modes must be determined. Once this is done, the Mie solutions can be evaluated and, as a result, the high-frequency backscattered field can be calculated.

17.4 Radial Eigenfunctions for Fields of Magnetic-Type

First consider fields of magnetic-type. From Westcott [16], the radial eigenfunctions for the refractive index profile $R(x)$ in Eq. (17.9) are given by

$$u(r) = [b_0^{\frac{1}{\alpha}} r]^{\frac{1+(c-1)\alpha}{2}} (1 + b_0 r^\alpha)^{\frac{A}{2}} {}_2F_1(a, b; c; -b_0 r^\alpha), \tag{17.14}$$

where ${}_2F_1(a, b; c; z)$ is the hypergeometric function. Writing this in terms of x and renaming the function $S_n^{(1)}(x)$, it is found (after much algebra, Pohrivchak [14]) that

$$S_n^{(1)}(x) = x^{\frac{1+(c-1)\alpha}{2}} (1 + x^\alpha)^{\frac{A}{2}} {}_2F_1(a, b; c; -x^\alpha), \tag{17.15}$$

where

$$a = \beta + \frac{2}{\alpha}\nu, \beta = \frac{A}{2}.$$

In due course, the solution (17.15) can be rewritten as

$$S_n^{(1)}(x) = x^{\nu+\frac{1}{2}} (1 + x^\alpha)^\beta {}_2F_1(\beta + \frac{2}{\alpha}\nu, \beta; 1 + \frac{2}{\alpha}\nu; -x^\alpha). \tag{17.16}$$

17.5 Radial Eigenfunctions for Fields of Electric-Type

For fields of electric-type, the radial eigenfunctions for the refractive index profile $R(x)$ in Eq. (17.9) are given by Westcott [16]

$$u(r) = [b_0^{\frac{1}{\alpha}} r]^{\frac{1+(c-1)\alpha}{2}} (1 + b_0 r^\alpha)^{\frac{a-b+1}{2}} {}_2F_1(a, c - b; c; -b_0 r^\alpha). \tag{17.17}$$

Again, after much algebraic manipulation, this can be written as

$$T_n^{(1)}(x) = x^{\alpha\gamma_- + \frac{\alpha}{2} - \frac{1}{2}} (1 + x^\alpha)^{\beta-1} {}_2F_1(\beta + \gamma_- + \gamma_+, \beta + \gamma_- - \gamma_+; 1 + 2\gamma_-; -x^\alpha). \tag{17.18}$$

17.6 Asymptotic Expansions for the TE Modes

We will now use the expression for the radial eigenfunctions for fields of magnetic-type given by Eq. (17.16) to calculate the asymptotic expansions for the TE modes. To that end, we must determine M_n which is given by Eq. (17.6a). Again, most of the technical details will be omitted. They may be found in [14]. From Eq. (17.6a), we find that the TE modes are given by

$$M_n = \frac{1}{k\hat{a}}\left[\frac{\alpha\beta + 2(\nu+\frac{1}{2})}{2} - \frac{\alpha\beta(\beta+\frac{2}{\alpha}\nu)}{1+\frac{2}{\alpha}\nu}\frac{{}_2F_1(\beta+\frac{2}{\alpha}\nu+1,\beta+1;2+\frac{2}{\alpha}\nu;-1)}{{}_2F_1(\beta+\frac{2}{\alpha}\nu,\beta;1+\frac{2}{\alpha}\nu;-1)}\right]$$
$$= \frac{1+\alpha\beta+2\nu}{2k\hat{a}} - \frac{\alpha\beta(\beta+\frac{2}{\alpha}\nu)}{k\hat{a}(1+\frac{2}{\alpha}\nu)}\frac{{}_2F_1(\beta+\frac{2}{\alpha}\nu+1,\beta+1;2+\frac{2}{\alpha}\nu;-1)}{{}_2F_1(\beta+\frac{2}{\alpha}\nu,\beta;1+\frac{2}{\alpha}\nu;-1)}. \tag{17.19}$$

Eventually we arrive at the result that

$$\frac{{}_2F_1(\beta+\frac{2}{\alpha}\nu+1,\beta+1;2+\frac{2}{\alpha}\nu;-1)}{{}_2F_1(\beta+\frac{2}{\alpha}\nu,\beta;1+\frac{2}{\alpha}\nu;-1)}$$
$$= \frac{1+\frac{2}{\alpha}\nu}{2\beta}\left[1 - \frac{(\frac{1}{\alpha}\nu-\frac{\beta}{2})\Gamma(\frac{1}{\alpha}\nu-\frac{\beta}{2})\Gamma\left(\frac{1+\frac{2}{\alpha}\nu+\beta}{2}\right)}{(\frac{1}{\alpha}\nu+\frac{\beta}{2})\Gamma(\frac{1}{\alpha}\nu+\frac{\beta}{2})\Gamma\left(\frac{1+\frac{2}{\alpha}\nu-\beta}{2}\right)}\right]. \tag{17.20}$$

Hence

$$M_{\nu-\frac{1}{2}} = \frac{1}{2k\hat{a}}\left[1 + \frac{\alpha(\frac{2}{\alpha}\nu-\beta)\Gamma(\frac{1}{\alpha}\nu-\frac{\beta}{2})\Gamma(\frac{1}{\alpha}\nu+\frac{1+\beta}{2})}{\Gamma(\frac{1}{\alpha}\nu+\frac{\beta}{2})\Gamma(\frac{1}{\alpha}\nu+\frac{1-\beta}{2})}\right]. \tag{17.21}$$

From the following formula

$$\frac{\Gamma(z+a)}{\Gamma(z+b)} \sim z^{a-b}\sum_{k=0}^{\infty}\frac{G_k(a,b)}{z^k} \sim z^{a-b}\left[G_0(a,b) + \frac{G_1(a,b)}{z} + \frac{G_2(a,b)}{z^2} + O(z^{-3})\right], \tag{17.22}$$

where

$$G_0(a,b) = 1; G_1(a,b) = \frac{1}{2}(a-b)(a+b-1);$$
$$G_2(a,b) = \frac{1}{12}\binom{a-b}{2}\left[3(a+b-1)^2 - (a-b+1)\right],$$

we are able eventually to determine that

$$M_{\nu-\frac{1}{2}} \sim \left(1-\frac{d_0}{2}-i\right)+\left(\frac{3}{4}-\frac{1}{2d_0}\right)(k\hat{a})^{-1} +\frac{i}{2}\left(\frac{\nu}{k\hat{a}}\right)^2+\left(\frac{2-d_0-4i}{16d_0^2}\right)(k\hat{a})^{-2}+\frac{i}{8}\left(\frac{\nu}{k\hat{a}}\right)^4 +O\left[\frac{\nu}{(k\hat{a})^3}\right]+O\left[\frac{\nu^3}{(k\hat{a})^4}\right]+O\left[(k\hat{a})^{-3}\right]. \tag{17.23}$$

17.7 Asymptotic Expansions for the TM Modes

We will now use the expression for the radial eigenfunctions for fields of electric-type given by Eq. (17.18) to calculate the asymptotic expansions for the TM modes. To that end, we must determine $\tilde{M}_n$ which is given by Eq. (17.6b). This is an even more complicated task than for the TE modes. First, we must obtain an expression for the derivative of the radial eigenfunctions for fields of electric-type. If we define

$$P=\frac{{}_2F_1(\beta+\gamma_-+\gamma_++1,\beta+\gamma_--\gamma_++1;2+2\gamma_-;-1)}{{}_2F_1(\beta+\gamma_-+\gamma_+,\beta+\gamma_--\gamma_+;1+2\gamma_-;-1)}, \tag{17.24}$$

then it is found that

$$\tilde{M}_n=\alpha\left[\frac{\beta-1+2\gamma_-}{2k\hat{a}}+\frac{1-\frac{1}{\alpha}}{2k\hat{a}}-\frac{[(\beta+\gamma_-)^2-\gamma_+^2]}{k\hat{a}(1+2\gamma_-)}P\right]. \tag{17.25}$$

After some manipulation of the hypergeometric functions, we are able to rewrite the expression for P in a slightly more useful form:

$$P=\frac{{}_2F_1(\beta+\gamma_-+\gamma_++1,1-\beta+\gamma_-+\gamma_+;2+2\gamma_-;\frac{1}{2})}{2\cdot{}_2F_1(\beta+\gamma_-+\gamma_+,1-\beta+\gamma_-+\gamma_+;1+2\gamma_-;\frac{1}{2})}. \tag{17.26}$$

The integral representation in Magnus et al. [17]

$${}_2F_1(b,\lambda;C;\frac{h}{z})=\frac{z^\lambda}{\Gamma(\lambda)}\int_0^\infty t^{\lambda-1}e^{-zt}{}_1F_1(b;C;ht)dt, \tag{17.27}$$

which is valid for

$$\mathrm{Re}\,z>\mathrm{Re}\,h>0, \qquad \mathrm{Re}\,\lambda>0,$$

may be applied to Eq. (17.26). Letting $b = \beta + \gamma_- + \gamma_+ + 1, \lambda = 1 - \beta + \gamma_- + \gamma_+, C = 2 + 2\gamma_-, h = 1$, and $z = 2$ in the numerator of Eq. (17.26) yields

$$
\begin{aligned}
&{}_2F_1(\beta + \gamma_- + \gamma_+ + 1, 1 - \beta + \gamma_- + \gamma_+; 2 + 2\gamma_-; \frac{1}{2}) \\
&= \frac{2^{1-\beta+\gamma_-+\gamma_+}}{\Gamma(1 - \beta + \gamma_- + \gamma_+)} \int_0^\infty t^{\gamma_-+\gamma_+-\beta} e^{-2t} {}_1F_1(\beta + \gamma_- + \gamma_+ + 1; 2 + 2\gamma_-; t) dt.
\end{aligned}
\tag{17.28}
$$

Similarly, letting $b = \beta + \gamma_- + \gamma_+, \lambda = 1 - \beta + \gamma_- + \gamma_+, C = 1 + 2\gamma_-, h = 1$, and $z = 2$ in the denominator of Eq. (17.26) gives us the result

$$
\begin{aligned}
&{}_2F_1(\beta + \gamma_- + \gamma_+, 1 - \beta + \gamma_- + \gamma_+; 1 + 2\gamma_-; \frac{1}{2}) \\
&= \frac{2^{1-\beta+\gamma_-+\gamma_+}}{\Gamma(1 - \beta + \gamma_- + \gamma_+)} \int_0^\infty t^{\gamma_-+\gamma_+-\beta} e^{-2t} {}_1F_1(\beta + \gamma_- + \gamma_+; 1 + 2\gamma_-; t) dt.
\end{aligned}
\tag{17.29}
$$

Hence

$$
P = \frac{\int_0^\infty t^{\gamma_-+\gamma_+-\beta} e^{-2t} {}_1F_1(\beta + \gamma_- + \gamma_+ + 1; 2 + 2\gamma_-; t) dt}{2 \int_0^\infty t^{\gamma_-+\gamma_+-\beta} e^{-2t} {}_1F_1(\beta + \gamma_- + \gamma_+; 1 + 2\gamma_-; t) dt}. \tag{17.30}
$$

In the notation of the Digital Library of Mathematical Functions (http://dlmf.nist.gov/)

$$
{}_1F_1(a; b; z) \equiv M(a, b, z).
$$

This is also known as Kummer's function. The closely related Olver's function is denoted by $\mathbf{M}(a, b, z)$, where

$$
M(a, b, z) = \Gamma(b)\mathbf{M}(a, b, z) = e^z M(b - a, b, -z); \tag{17.31a}
$$

$$
\mathbf{M}(a, b, -z) = \frac{z^{\frac{1}{2}(1-b)}}{\Gamma(a)} \int_0^\infty e^{-\tau} \tau^{a-\frac{1}{2}b-\frac{1}{2}} J_{b-1}(2\sqrt{z\tau}) d\tau. \tag{17.31b}
$$

Combining Eqs. (17.31a) and (17.31b), it follows that

$$
M(a, b, z) = \frac{e^z z^{\frac{1}{2}(1-b)} \Gamma(b)}{\Gamma(b - a)} \int_0^\infty e^{-\tau} \tau^{\frac{1}{2}(b-1)-a} J_{b-1}(2\sqrt{z\tau}) d\tau. \tag{17.32}
$$

In view of this we can write

$$
\begin{aligned}
&{}_1F_1(\beta+\gamma_-+\gamma_++1;2+2\gamma_-;t)\\
&=\frac{e^t t^{-\frac{1}{2}-\gamma_-}\Gamma(2+2\gamma_-)}{\Gamma(1-\beta+\gamma_--\gamma_+)}\int\limits_0^\infty e^{-\tau}\tau^{-\beta-\gamma_+-\frac{1}{2}}J_{1+2\gamma_-}(2\sqrt{t\tau})d\tau. \qquad (17.33)
\end{aligned}
$$

In a similar fashion it can be determined that

$$
{}_1F_1(\beta+\gamma_-+\gamma_+;1+2\gamma_-;t)=\frac{e^t t^{-\gamma_-}\Gamma(1+2\gamma_-)}{\Gamma(1-\beta+\gamma_--\gamma_+)}\int\limits_0^\infty e^{-\tau}\tau^{-\beta-\gamma_+}J_{2\gamma_-}(2\sqrt{t\tau})d\tau. \qquad (17.34)
$$

The integral representations in these last two equations are valid provided

$$
k\hat{a}-|\nu|\gg 1. \qquad (17.35)
$$

Proceeding on this basis equation (17.30) may be rewritten as

$$
P=\left(\frac{1}{2}+\gamma_-\right)\left\{\frac{\int\limits_0^\infty t^{\gamma_+-\beta-\frac{1}{2}}e^{-t}dt\int\limits_0^\infty e^{-\tau}\tau^{-\beta-\gamma_+-\frac{1}{2}}J_{1+2\gamma_-}(2\sqrt{t\tau})d\tau}{\int\limits_0^\infty t^{\gamma_+-\beta}e^{-t}dt\int\limits_0^\infty e^{-\tau}\tau^{-\beta-\gamma_+}J_{2\gamma_-}(2\sqrt{t\tau})d\tau}\right\}. \qquad (17.36)
$$

In Eq. (17.36), we make the change of variables

$$
t=u^2=k\hat{a}\xi^2, \qquad \tau=w^2=k\hat{a}\eta^2, \qquad (17.37)
$$

and then apply Sommerfeld's integral representation for a Bessel function of the first kind of order μ,

$$
J_\mu(2uw)=\frac{1}{2\pi}\int\limits_\Sigma d\tau e^{i\mu\tau-2iuw\sin\tau} \qquad (17.38)
$$

where the contour Σ begins at $\tau=-\pi+i\infty$ and ends at $\tau=\pi+i\infty$. After several intermediate steps we find that

$$
P=\left(\frac{\frac{1}{2}+\gamma_-}{k\hat{a}}\right)\frac{\int\limits_\Sigma d\tau e^{i(1+2\gamma_-)\tau}\int\limits_0^\infty d\eta e^{-k\hat{a}\eta^2-(2\gamma_++2\beta)\ln\eta}\int\limits_0^\infty d\xi e^{-k\hat{a}\xi^2-2ik\hat{a}\xi\eta\sin\tau+(2\gamma_+-2\beta)\ln\xi}}{\int\limits_\Sigma d\tau e^{2i\gamma_-\tau}\int\limits_0^\infty d\eta e^{-k\hat{a}\eta^2-(2\gamma_++2\beta-1)\ln\eta}\int\limits_0^\infty d\xi e^{-k\hat{a}\xi^2-2ik\hat{a}\xi\eta\sin\tau+(2\gamma_+-2\beta+1)\ln\xi}}. \qquad (17.39)
$$

This result is exact and valid for $|\nu| = O\left[(k\hat{a})^{\frac{1}{2}+\epsilon}\right]$. We can asymptotically evaluate the integrals by using the method of steepest descents. It is found (for $c_0 = 2$)

$$P \sim -\left(\frac{1/2+\gamma_-}{d_0 k\hat{a}}\right)[1+\tan f(\nu)]\left\{1+O\left(\frac{\nu}{k\hat{a}}\right)+O\left[\frac{\nu^3}{(k\hat{a})^2}\right]+O\left(\frac{1}{k\hat{a}}\right)\right\}, \tag{17.40}$$

where

$$f(\nu) = \frac{\pi}{4} - \frac{\pi}{2} d_0 k\hat{a} + \pi\gamma_- - \frac{1}{2}\arctan\frac{1}{2}. \tag{17.41}$$

After all this analysis the asymptotic expansions for the TM modes are given by

$$\tilde{M}_{\nu-\frac{1}{2}} \sim \left[1 - \frac{d_0}{2} + \tan f(\nu)\right]\left\{1+O\left(\frac{\nu}{k\hat{a}}\right)+O\left[\frac{\nu^3}{(k\hat{a})^2}\right]+O\left(\frac{1}{k\hat{a}}\right)\right\}. \tag{17.42}$$

17.8 The High-Frequency Backscattered Field

We have calculated the asymptotic expansions for the TE and TM modes in the previous two sections. We are now in a position to determine the difference of the Mie solutions, given by $a_n - b_n$, which appears in the expression of the high-frequency backscattered field given by Eq. (17.13). Once this is accomplished, we will be able to achieve the main objective of this chapter and determine the leading order estimate of the high-frequency backscattered field for a specific value of the positive real constant d_0. The Debye asymptotic expansions are used for the Bessel functions appearing in

$$(a_{\nu-\frac{1}{2}} - b_{\nu-\frac{1}{2}}).$$

In particular, for $|\nu| = O\left[(k\hat{a})^{1/2+\epsilon}\right]$, we will use the asymptotic relations

$$H_\nu^{(1)}(k\hat{a}) \sim \sqrt{\frac{2}{\pi\rho k\hat{a}}}\left\{1+O\left(\frac{1}{k\hat{a}}\right)+O\left[\frac{\nu^2}{(k\hat{a})^3}\right]\right\} \tag{17.43}$$

and

$$H_\nu^{(1)'}(k\hat{a}) \sim \sqrt{\frac{2}{\pi\rho k\hat{a}}}\left\{i+O\left(\frac{1}{k\hat{a}}\right)+O\left[\left(\frac{\nu}{k\hat{a}}\right)^2\right]\right\}, \tag{17.44}$$

where

$$\rho = \exp\left[i\left(\pi\nu + \frac{\pi}{2} - 2k\hat{a} - \frac{\nu^2}{k\hat{a}}\right)\right]\left\{1+O\left[\left(\frac{\nu}{k\hat{a}}\right)^2\right]+O\left[\frac{\nu^4}{(k\hat{a})^3}\right]\right\}. \tag{17.45}$$

The contour integral in the high-frequency backscattered field is now given by

$$\int_{C_1} \frac{\nu}{\cos \pi \nu} \left(a_{\nu-\frac{1}{2}} - b_{\nu-\frac{1}{2}} \right) d\nu = 2 \int_{C_1} \frac{\nu e^{-i\pi\nu}}{1 + e^{-2i\pi\nu}} \left(a_{\nu-\frac{1}{2}} - b_{\nu-\frac{1}{2}} \right) d\nu, \ (d_0 k\hat{a} \gg 1). \tag{17.46}$$

Noting that $\nu = n + 1/2$ it follows (eventually) that

$$a_{\nu-1/2} - b_{\nu-1/2} = a_n - b_n = \frac{(M_n - \tilde{M}_n)(\psi_n \zeta_n' - \psi_n' \zeta_n)}{(\zeta_n' - M_n \zeta_n)(\zeta_n' - \tilde{M}_n \zeta_n)} = \frac{i(M_n - \tilde{M}_n)}{(\zeta_n' - M_n \zeta_n)(\zeta_n' - \tilde{M}_n \zeta_n)}. \tag{17.47}$$

Hence asymptotically

$$a_n - b_n \sim \frac{1 - i \tan f(\nu)}{(\zeta_n' - M_n \zeta_n)(\zeta_n' - \tilde{M}_n \zeta_n)} \left[1 + O\left(\frac{\nu}{k\hat{a}} \right) + O\left(\frac{1}{k\hat{a}} \right) \right]. \tag{17.48}$$

After much algebra equation (17.46) may be written as

$$\begin{aligned} &\int_{C_1} \frac{\nu}{\cos \pi \nu} \left(a_{\nu-\frac{1}{2}} - b_{\nu-\frac{1}{2}} \right) d\nu \\ &\sim \frac{e^{-2ik\hat{a}}}{(d_0/4 - 1/2 + i)} \int_{C_1} \frac{\nu e^{-i\nu^2/k\hat{a}}}{1 + e^{-2i\pi\nu}} \left\{ \frac{1 - i \tan f(\nu)}{1 + i\,(1 - d_0/2) + i \tan f\,(\nu)} \right. \\ &\left. \left[1 + O\left(\frac{\nu}{k\hat{a}} \right) + O\left(\frac{1}{k\hat{a}} \right) \right] \right\} d\nu. \end{aligned} \tag{17.49}$$

For simplicity we will only consider the case where

- the optical rays do not make more than one turn about the center of the lens and
- at least one ray emerges in the backscattering direction.

These considerations yield the following bounds on d_0:

$$1 \leq d_0 \leq 2.$$

It transpires from detailed calculations that the contributions to the backscattered field arising from the poles enclosed by the contour C_1 and by the semicircle at infinity *cannot* be neglected when compared with the contour integral contribution if $d_0 < 2$. This means that the dominant term in the high-frequency backscattered field does not arise from specular reflection as in the case of a "lens" for which $d_0 = 2$. Therefore the dominant term in the high-frequency backscattered field is not obtainable by evaluating the contour integral by the saddle point method in this case. We will now turn our attention to the evaluation of the integral in Eq. (17.49) when $d_0 = 2$. This will allow for the determination of the leading order estimate of the high-frequency backscattered field by using Eq. (17.13). Hence

$$\mathbf{E}^{b.s} \sim -\tilde{\mathbf{e}}\frac{e^{ikr}}{2ikr}e^{-2ik\hat{a}} \times \int_{C_1} \frac{\nu}{1+e^{-2i\pi\nu}} e^{-i\frac{\nu^2}{k\hat{a}}} \left(\frac{1-i\tan f(\nu)}{1+i\tan f(\nu)}\right)\left[1+O\left(\frac{\nu}{k\hat{a}}\right)+O\left(\frac{1}{k\hat{a}}\right)\right]d\nu. \tag{17.50}$$

Now after some rearrangements

$$\frac{1-i\tan f(\nu)}{1+i\tan f(\nu)} = -ie^{2i\pi k\hat{a}-2i\pi\gamma_- +i\arctan(1/2)}.$$

Using this result in Eq. (17.50) provides the high-frequency backscattered field when $d_0 = 2$ as

$$\mathbf{E}^{b.s} \sim \tilde{\mathbf{e}}\frac{e^{ikr}}{2kr}e^{i[2k\hat{a}(\pi-1)+\arctan(1/2)]} \times \int_{C_1} \nu \frac{e^{-i\left(\nu^2/k\hat{a}+2\pi\gamma_-\right)}}{1+e^{-2i\pi\nu}}\left[1+O\left(\frac{\nu}{k\hat{a}}\right)+O\left(\frac{1}{k\hat{a}}\right)\right]d\nu. \tag{17.51}$$

To proceed, let M be a positive number, large compared with unity but independent of $k\hat{a}$. In other words, M can be described as follows:

$$M \gg 1, \qquad \lim_{k\hat{a}\to\infty}\frac{M}{k\hat{a}} = 0.$$

When $d_0 = 2$, it can be shown that the line integral along the arc of the circle vanishes as the radius tends to infinity. Also, the contributions to the backscattered field due to the poles in the first quadrant may be neglected because we only want the dominant term of the high-frequency backscattered field, and this arises from an asymptotic estimate of the line integral along the contour C_1. Next we can split the contour C_1 into three parts, by singling out the portion near $\nu = 0$ along which $|\nu| < M$ (see Fig. 17.1). Along this central portion, we have that

$$e^{-i\nu^2/k\hat{a}} \sim 1, \qquad (|\nu| < M)$$

so that the corresponding integral is $O(1)$, whereas the integral along the entire contour C_1 is $O(k\hat{a})$. Since what is required is only the leading term in the asymptotic estimate, we may neglect the central portion of C_1. Along the remaining part of the contour, it is determined that

$$\gamma_- \sim \nu, \qquad (|\nu| > M).$$

We find that

$$\int_{C_1} d\nu \frac{\nu e^{-i\frac{\nu^2}{k\hat{a}}}}{1+e^{-2i\pi\nu}} \sim \int_{C_2} d\nu \frac{\nu e^{-i\frac{\nu^2}{k\hat{a}}}}{1+e^{-2i\pi\nu}} + O(1),$$

where the contour C_2 in the uncut ν plane consists of that portion of C_1 along which $|\nu| > M$, plus the dashed line of Fig. 17.1. This results in the following modification of Eq. (17.51):

$$\mathbf{E}^{b.s} \sim -\tilde{\mathbf{e}}\frac{e^{ikr}}{2kr}e^{i[2k\hat{a}(\pi-1)+\arctan(1/2)]} \times \left\{\int_{C_2} \nu \frac{e^{-i\nu^2/k\hat{a}}}{1+e^{-2i\pi\nu}}\left[1+O\left(\frac{\nu}{k\hat{a}}\right)+O\left(\frac{1}{k\hat{a}}\right)\right]d\nu + O(1)\right\}. \quad (17.52)$$

Now

$$\int_{C_2} d\nu \frac{\nu e^{-i\frac{\nu^2}{k\hat{a}}}}{1+e^{-2i\pi\nu}}\left[1+O\left(\frac{\nu}{k\hat{a}}\right)+O\left(\frac{1}{k\hat{a}}\right)\right] \sim \int_{C_2} \frac{\nu e^{-i\nu^2/k\hat{a}}}{1+e^{-2i\pi\nu}}d\nu + \int_{C_2} \frac{\frac{\nu^2}{k\hat{a}}e^{-i\nu^2/k\hat{a}}}{1+e^{-2i\pi\nu}}d\nu + \frac{1}{k\hat{a}}\int_{C_2} \frac{\nu e^{-i\nu^2/k\hat{a}}}{1+e^{-2i\pi\nu}}d\nu$$

$$\equiv I_1 + I_2 + \frac{1}{k\hat{a}}I_1. \quad (17.53)$$

It may be shown that

$$I_1 \sim -\frac{k\hat{a}}{2}e^{-i\pi/2}. \quad (17.54)$$

Similarly

$$I_2 \sim -\frac{k\hat{a}}{2}e^{-i\frac{\pi}{2}}O((k\hat{a})^{-\frac{1}{2}}). \quad (17.55)$$

Hence the terms containing integrals in the backscattered field reduce to

$$-\frac{k\hat{a}}{2}e^{-i\pi/2}\left[1+O((k\hat{a})^{-\frac{1}{2}})+O((k\hat{a})^{-1})\right] \sim -\frac{k\hat{a}}{2}e^{-i\pi/2}\left[1+O((k\hat{a})^{-\frac{1}{2}})\right]. \quad (17.56)$$

Finally then, the leading order estimate of the high-frequency backscattered field is found to be

$$\mathbf{E}^{b.s} \sim \tilde{\mathbf{e}}\frac{\hat{a}}{4r}e^{i\left[kr+2k\hat{a}(\pi-1)-\pi/2+\arctan\frac{1}{2}\right]}\left\{1+O[(k\hat{a})^{-\frac{1}{2}}]\right\}. \quad (17.57)$$

It should be possible to consider other refractive index profiles that are based on the hypergeometric equation and to derive, as above the radial eigenfunctions for fields of magnetic- and electric-type. Two possibilities will be mentioned briefly here; more details may be found in [14]. The first is given (in terms of $x = r/\hat{a}$) by

$$R_2(x) = \frac{c_0}{x(1+x^\alpha)}, \tag{17.58}$$

and the second is

$$R_3(x) = \frac{c_0}{x\sqrt{1+x^\alpha}}. \tag{17.59}$$

As above, the results of Westcott [16] may be used in each case to determine the radial eigenfunctions for fields of magnetic-type and electric-type in terms of the original physical parameters for each system. Rather than so doing here, we provide below the comprehensive (and previously unpublished) details for the formulation of the radial "Schrödinger-type" equations, stated originally in [16].

17.9 Verification of Solutions from [16]

In that paper the author provides solutions for several wavenumbers $m(r)$ in the medium to the differential equation

$$\frac{d^2u}{dr^2} + \left\{ m_{eff}^2(r) - \frac{n(n+1)}{r^2} \right\} u = 0, \tag{17.60}$$

where

$$m_{eff}^2(r) = m^2(r) \qquad \text{for fields of magnetic-type, and} \tag{17.61a}$$

$$m_{eff}^2(r) = m^2(r) - m(r)\frac{d^2}{dr^2}\left\{\frac{1}{m(r)}\right\} \qquad \text{for fields of electric-type.} \tag{17.61b}$$

Note that $m(r) = k\tilde{n}(r)$, where k is the free space wavenumber and $\tilde{n}(r)$ is the refractive index profile. In this section we will verify that the solutions given by Westcott satisfy Eq. (17.60). Since many of the second derivative calculations of the provided solutions are rather long, here we only summarize the results for the second derivative that were derived. We denote solutions of Eq. (17.60) for fields of electric-type by $u_E(r)$ and for fields of magnetic-type by $u_M(r)$.

17.10 Profile Based on Bessel's Equation

Consider first the wavenumber

$$m(r) = ar^b, \tag{17.62}$$

where a and b are constants. With this wavenumber equation (17.60) has solution

$$u(r) \propto r^{\frac{1}{2}} Z_\nu(z), \tag{17.63}$$

where Z_ν denotes any solution of Bessel's equation of order ν and

$$z = \frac{a}{1+b} r^{b+1}.$$

Hence

$$u''(r) \propto r^{-2} \left[(b+1)^2 \nu^2 - (b+1)^2 z^2 - \frac{1}{4} \right] u(r). \tag{17.64}$$

The order ν is different for fields of electric- and magnetic-type.

17.10.1 Fields of Electric-Type

For this case

$$\nu^2 = \frac{b}{1+b} + \left\{ \frac{2n+1}{2(1+b)} \right\}^2. \tag{17.65}$$

Using Eq. (17.61b) it follows that

$$m_{eff}^2(r) = a^2 r^{2b} - \frac{b(b+1)}{r^2}. \tag{17.66}$$

Using Eqs. (17.64) and (17.65), it is found that

$$u_E''(r) = r^{-2}[b(b+1) + n(n+1) - a^2 r^{2(b+1)}] u_E(r),$$

It is readily shown from this that Eq. (17.60) is satisfied.

17.10.2 Fields of Magnetic-Type

For this case

$$\nu = \frac{n + \frac{1}{2}}{b + 1}, \tag{17.67}$$

with

$$m_{eff}^2(r) = m^2(r) = a^2 r^{2b}. \tag{17.68}$$

Hence

$$u_M''(r) = r^{-2}[n(n+1) - a^2 r^{2(b+1)}]u_M(r). \tag{17.69}$$

From this it is easy to verify that Eq. (17.60) is satisfied once again.

17.11 Profiles Based on Whittaker's Equation

Westcott [16] considered two wavenumber profiles in this context.

17.11.1 Profile 1

Now let

$$m(r) = \frac{a}{r \ln br}, \tag{17.70}$$

where a and b are constants. For fields of electric-type, independent solutions are given in [16] as

$$u_E(r) \propto r^{\frac{1}{2}} W_{\pm c,d}\{\pm(2n+1)\ln br\}, \tag{17.71}$$

where

$$c = -(2n+1)^{-1}, d = \sqrt{\frac{1}{4} - a^2},$$

and $W_{\pm c,d}(z)$ is Whittaker's function. Hence (after some algebra)

$$u_E''(r) = r^{-2}\left[\frac{1}{\ln br} - \frac{a^2}{(\ln br)^2} + n(n+1)\right]u_E(r).$$

Given that

$$m_{eff}^2(r) = \frac{a^2}{r^2(\ln br)^2} - \frac{1}{r^2 \ln br},$$

it follows that Eq. (17.60) is satisfied. For fields of magnetic-type independent solutions are given by

$$u_M(r) \propto (r \ln br)^{\frac{1}{2}} Z_\nu \left\{ \pm i \left(n + \frac{1}{2} \right) \ln br \right\}, \tag{17.72}$$

with $\nu = d$ (above). Since

$$u_M''(r) \propto [-a^2(r \ln br)^{-2} + n(n+1)r^{-2}] u_M(r),$$

and

$$m_{eff}^2(r) = a^2(r \ln br)^{-2}$$

it follows again that Eq. (17.60) is satisfied.

17.11.2 Profile 2

Now, consider the wavenumber

$$m(r) = \frac{a}{r\sqrt{\ln br}}, \tag{17.73}$$

where a and b are constants. For fields of electric-type, the independent solutions of Eq. (17.60) for this wavenumber are [16]

$$u_E(r) \propto r^{\frac{1}{2}} W_{\pm c,0}\{\pm(2n+1) \ln br\}, \tag{17.74}$$

where

$$c = \frac{a^2 - \frac{1}{2}}{2n+1}. \tag{17.75}$$

Again, since

$$u_E''(r) \propto \frac{1}{r^2} \left\{ \frac{(\frac{1}{2} - a^2)}{\ln br} - \frac{1}{4(\ln br)^2} + n(n+1) \right\} u_E(r),$$

and

$$m_{eff}^2(r) = \frac{1}{r^2}\left\{\frac{a^2}{\ln br} - \frac{1}{2\ln br} + \frac{1}{4(\ln br)^2}\right\},$$

we deduce that Eq. (17.60) is satisfied.

For fields of magnetic-type, the independent solutions corresponding to those in Eq. (17.74) are

$$u_M(r) \propto r^{\frac{1}{2}} W_{\pm c,\frac{1}{2}}\{\pm(2n+1)\ln br\}, \tag{17.76}$$

with

$$c = \frac{a^2}{2n+1}. \tag{17.77}$$

We find that

$$u_M''(r) \propto \frac{1}{r^2}\left\{-\frac{a^2}{\ln br} + n(n+1)\right\} u_M(r).$$

Given that

$$m_{eff}^2(r) = m^2(r) = \frac{a^2}{r^2 \ln br}$$

Eq. (17.60) is once again satisfied.

17.12 Profiles Based on the Hypergeometric Equation

For each of the three profiles that we will consider in this section, independent solutions for Eq. (17.60) for fields of electric- and magnetic-type, respectively, may be written as (where $z = -\beta r^\alpha$ with the constants α and β) [16]

$$u_E(r) \propto \left\{a_1 r^{\frac{1+(c-1)\alpha}{2}}(1-z)^{\frac{a-b+1}{2}}\,{}_2F_1(a, c-b; c; z) + a_2 r^{\frac{1-(c-1)\alpha}{2}}(1-z)^{\frac{a-b+1}{2}}\,{}_2F_1(1+a-c, 1-b; 2-c; z)\right\}; \tag{17.78a}$$

$$u_M(r) \propto \left\{a_1 r^{\frac{1+(c-1)\alpha}{2}}(1-z)^{\frac{1}{2}}\,{}_2F_1(a, b; c; z) + a_2 r^{\frac{1-(c-1)\alpha}{2}}(1-z)^{\frac{1}{2}}\,{}_2F_1(1+a-c, 1+b-c; 2-c; z)\right\}, \tag{17.78b}$$

where a_1 and a_2 are constants and ${}_2F_1(a,b;c;z)$ is Gauss's hypergeometric function. We note that the constants a, b, and c will be different for each profile. Let

$$L = \frac{c}{2} - \frac{c^2}{4}; M = \frac{A}{2} - \frac{A^2}{4}; N = \frac{1}{2}cA - ab; A = a + b - c + 1. \tag{17.79}$$

The constants a, b, and c in Eqs. ((17.78a)) and ((17.78b)) may be determined by Eq. (17.79) and another set of equations for L, M, and N that will be different for each profile. For future reference we note that

$$u_E''(r) \propto r^{-2} u_E(r) \left\{ \frac{\left[\left(N - \frac{1}{2}\right)\alpha^2 + \frac{1}{2}\right] z + \left[\left(\frac{1}{4} - M\right)\alpha^2 - \frac{1}{4}\right] z^2 - \left[\left(L - \frac{1}{4}\right)\alpha^2 + \frac{1}{4}\right]}{(1-z)^2} \right\}; \tag{17.80a}$$

$$u_M''(r) \propto r^{-2} u_M(r) \left\{ \frac{(N-M)\alpha^2 z^2 - N\alpha^2 z}{(1-z)^2} + \left(\frac{1}{4} - L\right)\alpha^2 - \frac{1}{4} \right\}. \tag{17.80b}$$

We will now consider three profiles and prove that the solutions of Eq. (17.60) are given by Eqs. ((17.78a)) and ((17.78b)) for fields of electric- and magnetic-type, respectively.

17.12.1 Profile 1

First, consider the wavenumber

$$m(r) = \frac{a_0}{r(1 + \beta r^\alpha)}, \tag{17.81}$$

where a_0 is a constant. For fields of electric-type with this wavenumber profile,

$$L = \left[a_0^2 - \frac{1}{4}(2n+1)^2\right]\alpha^{-2} + \frac{1}{4}; M = -\frac{3}{4} - \alpha^{-1} - \frac{1}{4}(2n+1)^2\alpha^{-2};$$

$$N = -\frac{1}{2}(2n+1)^2\alpha^{-2} - \frac{1}{2} - \alpha^{-1}. \tag{17.82}$$

Hence, because

$$m_{\mathit{eff}}^2(r) = \frac{a_0^2}{r^2(1-z)^2} + \frac{\alpha(\alpha+1)z}{r^2(1-z)}, \tag{17.83}$$

(recall that $z = -\beta r^\alpha$) it follows that

$$u_E''(r) + \left[m_{\mathit{eff}}^2(r) - \frac{n(n+1)}{r^2}\right] u_E(r) \propto 0. \tag{17.84}$$

For fields of magnetic-type L is the same as in Eq. (17.82), and

$$M = a_0^2\alpha^{-2}; N = 2a_0^2\alpha^{-2}. \tag{17.85}$$

Furthermore

$$m_{eff}^2(r) = \frac{a_0^2}{r^2(1-z)^2}, \tag{17.86}$$

so

$$u_M''(r) + \left[m_{eff}^2(r) - \frac{n(n+1)}{r^2}\right] u_M(r) \propto 0. \tag{17.87}$$

17.12.2 Profile 2

Next, we consider

$$m(r) = \frac{a_0 r^{\frac{\alpha}{2}-1}}{1+\beta r^\alpha}, \tag{17.88}$$

where a_0 is a constant. For fields of electric-type

$$L = \frac{1}{2}\alpha^{-1} - \frac{1}{4}(2n+1)^2\alpha^{-2}; M = -\frac{1}{2}\alpha^{-1} - \frac{1}{4}(2n+1)^2\alpha^{-2};$$
$$N = \left[a_0^2\beta^{-1} - \frac{1}{2}(2n+1)^2\right]\alpha^{-2}. \tag{17.89}$$

After some manipulation, we find that

$$u_E''(r) \propto r^{-2} u_E(r) \left\{ \frac{1}{4}\alpha^2 + n(n+1) + \frac{a_0^2\beta^{-1}z}{(1-z)^2} - \frac{\frac{1}{2}\alpha(1+z)}{1-z} \right\}, \tag{17.90}$$

and

$$m_{eff}^2(r) = r^{-2}\left\{ -\frac{a_0^2\beta^{-1}z}{(1-z)^2} - \frac{1}{4}\alpha^2 + \frac{\alpha}{2}\frac{(1+z)}{1-z} \right\}. \tag{17.91}$$

Once more, as required,

$$u_E''(r) + \left[m_{eff}^2(r) - \frac{n(n+1)}{r^2}\right] u_E(r) \propto 0. \tag{17.92}$$

For fields of magnetic-type

$$L = \frac{1}{4} - \frac{1}{4}(2n+1)^2\alpha^{-2}; M = N = -a_0^2\beta^{-1}\alpha^{-2}. \tag{17.93}$$

Upon using these definitions it is found that

$$u_M''(r) \propto r^{-2} u_M(r) \left[\frac{a_0^2 \beta^{-1} z}{(1-z)^2} + n(n+1) \right], \tag{17.94}$$

and

$$m_{eff}^2(r) = \frac{a_0^2 r^\alpha r^{-2}}{(1-z)^2} = -\frac{a_0^2 \beta^{-1} z}{(1-z)^2} r^{-2}. \tag{17.95}$$

Hence, it is determined that

$$u_M''(r) + \left[m_{eff}^2(r) - \frac{n(n+1)}{r^2} \right] u_M(r) \propto 0. \tag{17.96}$$

17.12.3 Profile 3

Finally, consider the wavenumber profile

$$m(r) = \frac{a_0}{r\sqrt{1+\beta r^\alpha}}, \tag{17.97}$$

where a_0, α, and β are constants. For fields of electric-type

$$L = \left[a_0^2 - \frac{1}{4}(2n+1)^2 \right] \alpha^{-2} + \frac{1}{2}; M = -\frac{1}{4}(2n+1)^2 \alpha^{-2} - \frac{1}{2}\alpha^{-1};$$

$$N = \left[a_0^2 - \frac{1}{2}(2n+1)^2 \right] \alpha^{-2} - \frac{1}{2}\alpha^{-1}. \tag{17.98}$$

Using the above definitions (again, after much algebra) we find that

$$u_E''(r) \propto r^{-2} u_E(r) \left\{ -\frac{a_0^2}{1-z} + n(n+1) - \frac{\frac{1}{2}\alpha z}{1-z} + \frac{\frac{1}{4}\alpha^2 z^2 - \frac{1}{2}\alpha^2 z}{(1-z)^2} \right\} \tag{17.99}$$

and

$$m_{eff}^2(r) = r^{-2} \left\{ \frac{a_0^2}{1-z} + \frac{\frac{1}{2}\alpha z}{1-z} + \frac{\frac{1}{2}\alpha^2 z - \frac{1}{4}\alpha^2 z^2}{(1-z)^2} \right\}, \tag{17.100}$$

whence

$$u_E''(r) + \left[m_{eff}^2(r) - \frac{n(n+1)}{r^2} \right] u_E(r) \propto 0. \tag{17.101}$$

For fields of magnetic-type

$$L = \frac{1}{4} + \left\{ a_0^2 - \frac{1}{4}(2n+1)^2 \right\} \alpha^{-2}; M = 0; N = a_0^2 \alpha^{-2}. \tag{17.102}$$

Again, it is found that

$$u_M''(r) \propto r^{-2} u_M(r) \left\{ -\frac{a_0^2}{1-z} + n(n+1) \right\} \tag{17.103}$$

and

$$m_{eff}^2(r) = m^2(r) = \frac{a_0^2 r^{-2}}{1-z}, \tag{17.104}$$

so that

$$u_M''(r) + \left[m_{eff}^2(r) - \frac{n(n+1)}{r^2} \right] u_M(r) \propto 0. \tag{17.105}$$

17.13 A Further Quantum Mechanical Connection

Aspects of plane wave electromagnetic scattering by a radially inhomogeneous sphere are discussed. The vector problem is reduced to two scalar radial "Schrödinger-like" equations, and a connection with time-independent potential scattering theory is exploited to draw several conclusions about specific refractive index profiles [18].

The refractive index $n(r)$ (which may be complex) is a function of the radial coordinate only, and the sphere has radius a. For $r > a$, $n(r) \equiv 1$. A time-harmonic dependence of the field quantities, $\exp(-i\omega t)$ is assumed throughout. The governing equation for the electric field $E(r, \theta, \phi)$ is

$$\nabla \times \nabla \times \mathbf{E} - k^2 n^2(r) \mathbf{E} = \mathbf{0}. \tag{17.106}$$

The wavenumber k is $2\pi/\lambda$, λ being the wavelength. As shown in [19], the solution may be found by expanding the electric field in terms of vector spherical harmonics in terms of the so-called transverse electric (TE) and transverse magnetic (TM) modes, respectively:

$$\mathbf{M}_{l,m}(r, \theta, \phi) = \frac{e^{im\phi}}{kr} S_l(r) \mathbf{X}_{l,m}(\theta), \tag{17.107a}$$

$$\mathbf{N}_{l,m}(r, \theta, \phi) = \frac{e^{im\phi}}{k^2 n^2(r)} \left[\frac{1}{r} \frac{dT_l(r)}{dr} \mathbf{Y}_{l,m}(\theta) + \frac{T_l(r)}{r^2} \mathbf{Z}_{l,m}(\theta) \right]. \tag{17.107b}$$

The vector angular functions in Eqs. (17.107a), (17.107b) are defined in a spherical coordinate system as

$$\mathbf{X}_{l,m}(\theta) = \langle 0, i\pi_{l,m}(\theta), -\tau_{l,m}(\theta)\rangle, \tag{17.108a}$$

$$\mathbf{Y}_{l,m}(\theta) = \langle 0, \tau_{l,m}(\theta), -i\tau_{l,m}(\theta)\rangle, \tag{17.108b}$$

$$\mathbf{Z}_{l,m}(\theta) = \left\langle l(l+1) P_l^m(\cos\theta), 0, 0\right\rangle, \tag{17.108c}$$

where $P_l^m(\cos\theta)$ is an associated Legendre polynomial of degree l and order m. The corresponding scalar angular functions are defined as

$$\pi_{l,m}(\theta) = \frac{m}{\sin\theta} P_l^m(\cos\theta), \tag{17.109a}$$

$$\tau_{l,m}(\theta) = \frac{dP_l^m(\cos\theta)}{d\theta}. \tag{17.109b}$$

The functions $S_l(r)$ and $T_l(r)$ are called the radial Debye potentials, and they respectively satisfy the equations

$$\frac{d^2S_l(r)}{dr^2} + \left[k^2n^2(r) - \frac{l(l+1)}{r^2}\right] S_l(r) = 0, \tag{17.110a}$$

$$\frac{d^2T_l(r)}{dr^2} - \left(\frac{2}{n(r)}\frac{dn(r)}{dr}\right)\frac{dT_l(r)}{dr} + \left[k^2n^2(r) - \frac{l(l+1)}{r^2}\right] T_l(r) = 0. \tag{17.110b}$$

In addition to the appropriate matching conditions at $r = a$ these potentials must also satisfy the boundary conditions $S_l(0) = 0$ and $T_l(0) = 0$. Equation (17.110b) may be rewritten in terms of the dependent variable $U_l(r)$, where $T_l(r) = n(r) U_l(r)$ to become

$$\frac{d^2U_l(r)}{dr^2} + \left[k^2n^2(r) - n(r)\frac{d^2}{dr^2}\left(\frac{1}{n(r)}\right) - \frac{l(l+1)}{r^2}\right] U_l(r) = 0. \tag{17.111}$$

Provided that $n(0) \neq 0$, $U_l(0) = 0$. Both Eqs. (17.110a) and (17.111) may be placed in the form of the canonical time-independent Schrödinger equation, namely

$$\frac{d^2S_l(r)}{dr^2} + \left[k^2 - V_S(r) - \frac{l(l+1)}{r^2}\right] S_l(r) = 0, \tag{17.112a}$$

$$\frac{d^2U_l(r)}{dr^2} + \left[k^2 - V_U(r) - \frac{l(l+1)}{r^2}\right] U_l(r) = 0, \tag{17.112b}$$

where the k-dependent "scattering potentials" $V_S(r)$ and $V_U(r)$ are defined, respectively, in $[0, a]$ as

$$V_S(r) = k^2\left[1 - n^2(r)\right], \tag{17.113a}$$

$$V_U(r) = k^2\left[1 - n^2(r) + \frac{n(r)}{k^2}\frac{d^2}{dr^2}\left(\frac{1}{n(r)}\right)\right]. \tag{17.113b}$$

for the TE and TM modes, respectively (the potentials are both identically zero for $r > a$). These potentials are identical for the case of a uniform refractive index. $V_U(r)$ will be regarded as a small perturbation of the potential $V_S(r)$, so we also define

$$\varepsilon(r) \equiv V_U(r) - V_S(r) = n(r)\frac{d^2}{dr^2}\left(\frac{1}{n(r)}\right). \tag{17.114}$$

It is a standard result for potentials vanishing sufficiently fast at infinity [2–4] that as $r \to \infty$

$$S_l(r) \sim \sin\left(r - \frac{\pi l}{2} + \delta_l^S(k)\right), \tag{17.115a}$$

$$U_l(r) \sim \sin\left(r - \frac{\pi l}{2} + \delta_l^U(k)\right). \tag{17.115b}$$

Here $\delta_l^S(k)$ and $\delta_l^U(k)$ are the phase shifts induced by each potential, respectively. Multiplying Eqs. (17.7a) and (17.7b) by $U_l(r)$ and $S_l(r)$, respectively, subtracting and integrating we obtain

$$U_l(r)\frac{dS_l(r)}{dr} - S_l(r)\frac{dU_l(r)}{dr} = -\int_0^r \varepsilon(\eta) S_l(\eta) U_l(\eta)\, d\eta. \tag{17.116}$$

Utilizing the asymptotic expressions in (17.115a), (17.115b), we have, in the limit as $r \to \infty$,

$$k\sin\left[\delta_l^U(k) - \delta_l^S(k)\right] = -\int_0^\infty \varepsilon(r) S_l(r) U_l(r)\, dr = -\int_0^{ka} \varepsilon(r) S_l(r) U_l(r)\, dr, \tag{17.117}$$

since $n(r)$ is constant for $r > ka$ (or $r > a$). Thus far this equation is exact. If we now consider $\varepsilon(r)$ to be sufficiently small that $U_l(r) \approx S_l(r)$, then $\left|\delta_l^U(k) - \delta_l^S(k)\right| << 1$ and we have the relation

$$\delta_l^U(k) \approx \delta_l^S(k) \pm \frac{1}{k}\int_0^{ka} \varepsilon(r)\left[S_l(r)\right]^2 dr. \tag{17.118}$$

Whether $\delta_l^U(k) > \delta_l^S(k)$ or not clearly depends on the concavity of $n(r)$. A further approximation can be made if the scattering potential $V_S(r)$ is constant (specifically, $V_S = k^2\left(1 - N^2\right)$ for $n = N$, $r \leq a$), for then the solution for Eq. (17.112a) can be expressed in terms of a Riccati–Bessel function of the first kind, i.e.

$$S_l(r) = \left(\frac{\pi N k r}{2}\right)^{1/2} J_{l+1/2}(Nkr). \tag{17.119}$$

Then we have that

$$\begin{aligned}\delta_l^U(k) &\approx \delta_l^S(k) \pm \frac{\pi N}{2}\int_0^a \left\{n(r)\frac{d^2}{dr^2}\left(\frac{1}{n(r)}\right)\right\}\left[J_{l+1/2}(Nkr)\right]^2 r dr \\ &\equiv \delta_l^S(k) \pm \frac{\pi N}{2}\mathcal{I}(a).\end{aligned} \tag{17.120}$$

In the case of a small perturbation about $V_S = 0$, i.e. for which $n = N = 1$, the term $\delta_l^S(k)$ in Eq. (17.120) is zero, and the resulting approximation for $\delta_l^U(k)$ is related to the first Born approximation in quantum scattering theory [20]. In particular, if $\varepsilon(r) = Dr^{-s}$, D being some constant, a closed form solution for I can be found as $a \to \infty$ [21], namely

$$\mathcal{I}(\infty) = \int_0^\infty \left[J_{l+1/2}(Nkr)\right]^2 r^{1-s} dr = \frac{1}{2}\left(\frac{Nk}{2}\right)^{s-2} \frac{\Gamma(s-1)\,\Gamma\left(l - \frac{1}{2}s + \frac{3}{2}\right)}{\left[\Gamma\left(\frac{1}{2}s\right)\right]^2 \Gamma\left(l + \frac{1}{2}s + \frac{1}{2}\right)}, \tag{17.121}$$

provided $s > 1$ and $2l > s - 3$. The question may be asked: what $n(r)$ profiles give rise to $\varepsilon(r) = Dr^{-s}$ (where $D > 0$)? Writing $p(r) = [n(r)]^{-1}$ we are led to consider solutions of the equation

$$r^s \frac{d^2 p(r)}{dr^2} - Dp(r) = 0. \tag{17.122}$$

The general solution to this equation may be expressed in terms of modified Bessel functions, but we do not pursue this direction here.

17.14 A Liouville Transformation

As defined in Eqs. (17.113a) and (17.113b), the "potentials" $V_S(r)$ and $V_U(r)$ are also k-dependent, which is not the case in potential scattering theory [22]. This has an important consequence: unlike the quantum mechanical case, here pure

"bound state" solutions, that is, real square-integrable solutions corresponding to $k^2 < 0$ ($\mathrm{Im}\, k > 0$) do not exist. It can be readily proven [23] for the TE mode [Eq. (17.112a)] that

$$\int_0^\infty \left[\left| \frac{dS_l(r)}{dr} \right|^2 + \frac{l(l+1)}{r^2} |S_l(r)|^2 \right] dr = k^2 \int_0^\infty n^2(r) |S_l(r)|^2 dr. \quad (17.123)$$

This cannot be satisfied for $k^2 < 0$ for a real and positive refractive index $n(r)$. In [24] the corresponding result is established from Eq. (17.112b) for $U_l(r)$. Furthermore, a Liouville transformation may be used to define a new k-independent potential. Using the following simultaneous changes of independent and dependent variables in Eq. (17.110a)

$$r \to \rho : \rho(r) = \int_0^r n(s) ds, \quad (17.124a)$$

$$u_l \to \psi_l : \psi_l(\rho) = (n(r))^{1/2} u_l(r). \quad (17.124b)$$

Clearly $n(r)$ must be integrable and non-negative (in naturally occurring circumstances $n \geq 1$ and $n(r) = 1$ for $r > a$); also $\rho(0) = 0$. It is easy to establish the following results:

$$(i)\ \rho(r) = \rho_0 + r - a, r \geq a, \text{where} \rho_a = \int_0^a n(s) ds;$$

$$(ii)\ \rho(r) \sim r, r \to \infty;$$

$$(iii)\ r(\rho) = \int_0^\rho \frac{ds}{v(s)}, \text{where} v(\rho) = n(r(\rho)).$$

Furthermore, by applying (17.124a) and (17.124b) to Eq. (17.112a) we find that

$$\left[\frac{d^2}{d\rho^2} - \frac{l(l+1)}{R^2(\rho)} + k^2 \right] \psi_l(r) = V(\rho) \psi_l(\rho), \quad (17.125)$$

where

$$R(\rho) = v(\rho) r(\rho) \sim n(0) \rho, \rho \to 0, \text{and} V(\rho) = [v(\rho)]^{-1/2} \frac{d^2}{d\rho^2} [v(\rho)]^{1/2}. \quad (17.126)$$

Clearly $v(\rho)$ should be at least twice differentiable. Now the new "potential" $V(\rho)$ is independent of the wavenumber k. Note also that $V(\rho) = 0$ for $\rho > \rho_a$. It is of interest to determine the "shape" of the potential $V(\rho)$ by inverting $\rho(r)$ for

various choices of physical $n(r)$ profiles for $r \in [0, a]$ (with $n(0) = n_0$, $n(a) = n_a$ and $n(r) = 1$ for $r > a$). In what follows only the non-zero potential shapes with be stated (corresponding to $\rho \in [0, \rho_a]$. Thus [24] for

$$n(r) = n_a \left[1 - c^2 \left(\frac{r-a}{a}\right)^2\right]^{-1}; V(\rho) = \frac{c^2}{n_a^2} > 0, \tag{17.127a}$$

where c is a real constant, i.e. the potential is a spherical barrier. For the profile [25]

$$n(r) = (A + Br)^{-1}, A = n_0^{-1}, B = \frac{n_0 - n_a}{a n_0 n_a}; V(\rho) = \frac{B^2}{4} > 0, \tag{17.127b}$$

also a barrier. For the important Maxwell Fish-Eye profile [26],

$$n(r) = n_0 \left(1 + Br^2\right)^{-1}, B = \frac{n_0 - n_a}{a^2 n_a}; V(\rho) = -\frac{B}{n_0^2}. \tag{17.127c}$$

In this case, the new potential is a spherical well or barrier as $n_0 > n_a$ or $n_0 < n_a$, respectively. In the latter case the singularity occurring in $n(r)$ is moot since it arises for $r > a$. In all the other cases investigated thus far [27], including $n(r) = n_0 \exp(-\alpha r)$; $n_0 \cos \alpha r$ and $n_0 \cosh \alpha r$, the potentials $V(\rho)$ are rather complicated functions, and there are no significant advantages to using the Liouville transformation in these cases. It is therefore of interest to examine what profiles $n(r)$ give rise to constant potentials $V(\rho)$. In Eq. (17.126) let $y(\rho) = [v(\rho)]^{1/2}$ and $V(\rho) = V_0$, where V_0 is a constant of either sign. Then it follows that

$$\frac{d^2 y}{d\rho^2} - V_0 y = 0, \tag{17.128}$$

the general solution being expressible in terms of real or complex exponential functions as $V_0 > 0$ (potential barrier) or $V_0 < 0$ (potential well), respectively. In r-space, $V_0 < 0$ corresponds to a constant refractive index $n = N = \left(1 + |V_0| k^{-2}\right)^{1/2} > 1$, so we proceed with this physically realistic case. Writing the general solution of Eq. (17.128) as

$$y(\rho) = C \cos\left(|V_0|^{1/2} \rho + \eta\right), \tag{17.129}$$

where C and η are constants, it follows that

$$r(\rho) = \int_0^\rho \frac{ds}{v(s)} = \left(C^2 |V_0|^{1/2}\right)^{-1} \left[\tan\left(|V_0|^{1/2} \rho + \eta\right) - \tan \eta\right]. \tag{17.130}$$

This can be inverted to yield

$$\rho(r) = \int_0^r n(s)ds = |V_0|^{-1/2} \left\{ \arctan\left[C^2 |V_0|^{1/2} r + \tan\eta \right] - \eta \right\}. \tag{17.131}$$

Therefore

$$n(r) = \rho'(r) = \frac{C}{1 + [Br + \tan\eta]^2}, \tag{17.132a}$$

where $C = n_0 \sec^2\eta$ and η can be determined from the requirement that $n(a) = n_a$. This is a generalization of the Maxwell Fish-Eye profile in Eq. (17.127c). The corresponding result for $V_0 > 0$ is

$$n(r) = \frac{C}{1 - [Br + \tanh\eta]^2}. \tag{17.132b}$$

Note that in this case a singularity exists for $r > 0$ at $r = B^{-1}(1 - \tanh\eta)$.

17.15 Summary

Aside from a historical introduction to the visual consequences of electromagnetic scattering by large spheres in general, and backscattering in particular, a major contribution of this paper to the literature is to provide derivations of several results stated therein but with no or limited details provided. A second contribution is the extension of these same results to other refractive index profiles (in the case of electromagnetic backscattering). Since exact electromagnetic solutions for radially inhomogeneous dielectric lenses are available only for few functional dependences of the refractive index on the radial distance, the high-frequency behavior based on an asymptotic analysis of the exact solution has been obtained in very few cases. This article based on [13] has extended the range of possible profiles by generalizing the parameters of the original mathematical model. When an exact solution is not available, a high-frequency estimate may be obtained by performing an asymptotic analysis of the differential equations satisfied by the radial eigenfunctions. But even when an exact solution is available, it may be easier to proceed directly with an asymptotic solution of the differential equation [28]. Additionally, by exploiting some known results from quantum mechanics, we are able to derive asymptotic solutions for two scalar problems (decoupled from the electromagnetic cases) for the case of small variations in the refractive index across the scattering sphere. Finally, by using a Liouville transformation we are able to convert the electromagnetic wavenumber-dependent scattering potential to a wavenumber-independent one, and

solve the resulting inverse problem for several refractive index profiles. The reader interested in further details for both the historical and mathematical aspects of this chapter should consult references [29] and [30], respectively.

References

1. Lee, R.L. Jr., Fraser, A.B.: The Rainbow Bridge: Rainbows in Art, Myth, and Science. Pennsylvania State University Press, University Park (2001)
2. Adam, J.A.: In: Bourama, T. (eds.) Advances in Interdisciplinary Mathematical Research: Applications to Engineering, Physical and Life Sciences. Springer Proceedings in Mathematics & Statistics, vol. 37, pp. 57–96. Springer, New York (2013)
3. Logan N.A.: Proc. IEEE **53**, 773–785 (1965)
4. Watson, G.N.: Proc. R. Soc. (Lond.) **A95**, 83–99 (1918)
5. Adam, J.A.: Not. Am. Math. Soc. **49**, 1360–1371 (2002)
6. Kerker, M.: The Scattering of Light and Other Electromagnetic Radiation. Academic, New York (1969)
7. Inada, H., Plonus, M.A.: IEEE Trans. Antennas Propag. **AP-18**, 89–99 (1970)
8. Inada, H., Plonus, M.A.: IEEE Trans. Antennas Propag. **AP-18**, 649–660 (1970)
9. Rubinow, S.I.: Ann. Phys. **14**, 305–332 (1961)
10. Nussenzveig, H.M.: J. Math. Phys. **10**, 82–124 (1969); **10**, 125–176 (1969)
11. Brockman, C.L. Jr.: High frequency electromagnetic wave backscattering from radially inhomogeneous dielectric spheres, University of California, Los Angeles. Ph.D., Electrical Engineering (1974)
12. Brockman, C.L. Jr., Alexopoulos, N.G.: Appl. Opt. **16**, 166–174 (1977)
13. Uslenghi, P.L.E., Weston, V.H.: Appl. Sci. Res. **23**, 147–163 (1970)
14. Pohrivchak, M.A.: Ray- and wave-theoretic approach to electromagnetic scattering from radially inhomogeneous spheres and cylinders. Ph.D. dissertation, Old Dominion University (2014)
15. Born, M., Wolf, E.: Principles of Optics, 7th edn. Cambridge University Press, Cambridge (1999)
16. Westcott, B.S.: Electron. Lett. **4**, 572–573 (1968)
17. Magnus, W., Oberhettinger, F., Soni, R.: Formulas and Theorems for the Special Functions of Mathematical Physics. Springer, New York (1966)
18. Nuntaplook, U., Adam, J.A.: Applied Mathematics E-Notes (in press)
19. Johnson, B.R.: J. Opt. Soc. Am. **A10**, 343–352 (1993)
20. Schiff, L.I.: Quantum Mechanics, 3rd edn. McGraw-Hill, New York (1968)
21. de Alfaro, V., Regge, T.: Potential Scattering. North-Holland Publishing Company, Amsterdam (1965)
22. Mott, N.F., Massey, H.S.W.: The Theory of Atomic Collisions, 3rd edn. Clarendon Press, Oxford (1965)
23. Eftimiu, C.: J. Math. Phys. **23**, 2140–2146 (1982)
24. Eftimiu, C.: Inverse electromagnetic scattering for radially inhomogeneous dielectric spheres, in Inverse methods in electromagnetic imaging. In: Proceedings of the NATO Advanced Research Workshop, Bad Windsheim, Part 1 (A85-48926, 24-70). D. Reidel Publishing Company, Dordrecht (1983)

25. Adam, J.A., Laven, P.: Appl. Opt. **46** 922–929 (2007)
26. Leonhardt, U., Philbin, T.: Geometry and Light: The Science of Invisibility, Dover Publications, New York (2010)
27. Adam. J. A.: unpublished notes
28. Uslenghi, P.L.E.: In: Antennas and Propagation Society International Symposium Conference Proceedings: APSURSI, vol. 23, pp. 389–390 (1985)
29. Kragh, H.: Appl. Opt. **30**, 4688–4695 (1991)
30. Adam, J.A.: In: Kokhanovsky, A. (ed.) Light Scattering Reviews, vol. 9, Chap. 3. Springer, Berlin (2014)

Chapter 18
An Introduction to Symplectic Coordinates

M. AÍvarez-Ramírez and Rodríguez José Antonio García

Abstract In this paper we review the relations between Hamiltonian systems and the symplectic geometry in a simple context. We use them to reduce the degrees of freedom of the system. In particular they are used to obtain the solutions of the two-body problem.

Keywords Symplectic geometry • Hamiltonian function • Mathieu transformation

18.1 Introduction

A practical difficulty that arises in many problems of mechanics is the selection of a good system of coordinates. Each one has its own advantages and drawbacks, and thus there is not a best one; moreover, changing systems of coordinates usually involves several and difficult computations. However there is a consensus on the preference of systems that preserve symplectic structures and reflect the symmetries of the specific problem. In this paper, we shall examine some basic properties of the symplectic geometry and use it in a classical example in celestial mechanic in order to persuade the reader of this claim.

This example is the two-body problem, where the orbits of two particles in $\mathbb{R}^2$, subject to their mutual gravitational interaction, are studied. The underlying symplectic properties will provide us with a clean way to analyze this dynamical system, and they are going to be used along this paper to obtain the solution.

The analysis of the two-body problem starts with the choice of the vector space $\mathbb{R}^8$ of positions and momenta of the two particles as the phase space, and the construction of a symplectic structure where the energy is the Hamiltonian of the system. The two-body problem preserves the center of mass, the linear and angular momenta, and the energy; the next steps will use them to reduce the complexity of the problem, and finally to give the general solution.

M. AÍvarez-Ramírez (✉) • R.J. A. García
Departamento de Matemáticas, UAM-Iztapalapa, San Rafael Atlixco 186, Col. Vicentina, 09340 Iztapalapa, México, D.F., Mexico
e-mail: mar@xanum.uam.mx; agar@xanum.uam.mx

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_18

The current work is presented as follows. Section 18.2 has a brief introduction to Hamiltonian systems and the two-body problem is presented. Section 18.3 provides some elements of symplectic geometry such as the definition of symplectic spaces and symplectic changes of coordinates.

There is a discussion of several relations between Hamiltonian systems and symplectic geometry in Sect. 18.4. Poisson brackets, integrals, and Liouville integrability are introduced here, along with the statements of the theorems of Liouville and Noether.

Section 18.5 deals with the Mathieu transformation. This will allow us to reduce the degrees of freedom in a Hamiltonian system. Finally, Sect. 18.6 provides the reader with some comments and references that are helpful to extend the knowledge about the symplectic geometry and some related topics.

18.2 Hamiltonian Systems

In order to avoid the complications of the more general definition, we start with a simple definition of a Hamiltonian system, see [1, 3] for more general but abstract definitions.

Let $H : \mathcal{O} \subset \mathbb{R}^{2n} \to \mathbb{R}$, or $H(\mathbf{q}, \mathbf{p})$ in its arguments, $\mathbf{q} = (q_1, \dots q_n)$, $\mathbf{p} = (p_1, \dots p_n) \in \mathbb{R}^n$, be a smooth function. Then the dynamical system

$$\dot{q}_i = \frac{\partial H}{\partial p_i}, \qquad \dot{p}_i = -\frac{\partial H}{\partial q_i} \qquad i = 1, 2, \dots, n \tag{18.1}$$

is a *Hamiltonian system*, and the function $H(\mathbf{q}, \mathbf{p})$ is the *Hamiltonian* of the system. The vector field in the right side of Eq. (18.1) is known as a *Hamiltonian vector field*. We will sometimes denote it by $\mathbf{X}_H$. We will also write by $\boldsymbol{\varphi}_H(t, \mathbf{q}_0, \mathbf{p}_0)$ the solution such that $\mathbf{q}(0) = \mathbf{q}_0$ and $\mathbf{p}(0) = \mathbf{p}_0$, and call it a *Hamiltonian flow* or the *flow of the Hamiltonian*.

The vector $\mathbf{q}$ comprises the configuration variables of the system, and $\mathbf{p}$ their canonically conjugate momenta. Each component of $\mathbf{q}$ has as conjugate momentum a component of $\mathbf{p}$. Loosely speaking, $\mathbf{q}$ describes the position of the system, and $\mathbf{p}$ the rate of change. The number n is the *number of degrees of freedom* of the Hamiltonian system. Then the phase space is $2n$-dimensional.

The Hamiltonian can be thought as the energy of the system. In fact, the law of the conservation of energy is expressed as

$$\frac{d}{dt} H\left(\mathbf{q}(t), \mathbf{p}(t)\right) = 0. \tag{18.2}$$

Hence the orbits are constrained in a single level set of the Hamiltonian.

If $\mathbf{z} = (\mathbf{q}, \mathbf{p})$, the system (18.1) can be written as

$$\dot{\mathbf{z}} = \mathbf{X}_H(\mathbf{z}) = \mathbf{J}\nabla H(z), \qquad \mathbf{J} = \begin{pmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{pmatrix}, \tag{18.3}$$

where $\mathbf{J}$ is a square matrix of order $2n$, formed by 4 square matrices of order n, $\mathbf{I}$ and $\mathbf{0}$ are the identity and zero matrix, respectively. The matrix $\mathbf{J}$ satisfies $\mathbf{J}^T = -\mathbf{J}$, $\mathbf{J}^2 = -\mathbf{I}$, and the $\det(\mathbf{J}) = 1$.

Example 2.1. Here we study the motion of two particles which are moving in a plane and attracting each other following the Newton laws. Let m_k, $\mathbf{q}_k$, and $\mathbf{p}_k$ be the mass, position, and momentum of the kth particle, without loss of generality we assume that $m_1 + m_2 = 1$ and the universal gravitational constant is set to one. Then the motion equation is

$$\dot{\mathbf{q}}_1 = \frac{\partial H}{\partial \mathbf{p}_1} = \frac{1}{m_1}\mathbf{p}_1, \qquad \dot{\mathbf{p}}_1 = -\frac{\partial H}{\partial \mathbf{q}_1} = m_1 m_2 \frac{\mathbf{q}_2 - \mathbf{q}_1}{|\mathbf{q}_2 - \mathbf{q}_1|^3},$$

$$\dot{\mathbf{q}}_2 = \frac{\partial H}{\partial \mathbf{p}_2} = \frac{1}{m_2}\mathbf{p}_2, \qquad \dot{\mathbf{p}}_2 = -\frac{\partial H}{\partial \mathbf{q}_2} = m_1 m_2 \frac{\mathbf{q}_1 - \mathbf{q}_2}{|\mathbf{q}_2 - \mathbf{q}_1|^3}.$$

The solutions of this system are of the form

$$(\mathbf{q}_1(t), \mathbf{q}_2(t), \mathbf{p}_1(t), \mathbf{p}_2(t)) \in \mathbb{R}^8.$$

This equation is a four degrees of freedom Hamiltonian system with Hamiltonian

$$H(\mathbf{q}_1, \mathbf{q}_2, \mathbf{p}_1, \mathbf{p}_2) = \frac{\mathbf{p}_1 \cdot \mathbf{p}_1}{2m_1} + \frac{\mathbf{p}_2 \cdot \mathbf{p}_2}{2m_2} - \frac{m_1 m_2}{|\mathbf{q}_2 - \mathbf{q}_1|}. \tag{18.5}$$

18.3 Symplectic Geometry

In this section we introduce some basic tools of symplectic geometry.

Definition 3.1. A *symplectic form* is a bilinear form ω on a vector space V of dimension even, such that

1. (skew-symmetric) For all $\mathbf{v}, \mathbf{w} \in V$ such that $\omega(\mathbf{v}, \mathbf{w}) = -\omega(\mathbf{w}, \mathbf{v})$.
2. (nondegenerate) If for every $\mathbf{w} \in V$ such that $\omega(\mathbf{v}, \mathbf{w}) = \mathbf{0}$, then $\mathbf{v} = \mathbf{0}$.

The vector space V is a *symplectic vector space*.

Example 3.2. The canonical example of a symplectic vector space is $\mathbb{R}^{2n}$ with the bilinear form $\omega(\mathbf{u}, \mathbf{v}) = \mathbf{J}\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^t \mathbf{J}\mathbf{v}$, where $\mathbf{u}^t$ denotes the transpose of $\mathbf{u}$.

It is usual to identify the vector space V with $\mathbb{R}^{2n}$ by picking a basis $\mathcal{B} = \{\mathbf{e}_1, \ldots \mathbf{e}_{2n}\}$, and the symplectic form with the matrix $\mathbf{A} = (a_{ij})$, where $a_{ij} = \omega(\mathbf{e}_i, \mathbf{e}_j)$. If this associated matrix is $\mathbf{J}$, then $\mathcal{B}$ is a *symplectic basis*. A basic fact is that any symplectic vector space has a symplectic basis.

Definition 3.3. Let (V, ω) be a symplectic space of dimension $2n$, a *Lagrangian subspace* $W \subset V$ is a linear subspace of V of dimension n such that for all $\mathbf{v}$, $\mathbf{w} \in W$: $\omega(\mathbf{v}, \mathbf{w}) = 0$.

Example 3.4. If $\mathcal{B} = \{\mathbf{e}_1, \ldots \mathbf{e}_{2n}\}$ is a symplectic basis, then $W_1 = \langle \mathbf{e}_1, \ldots \mathbf{e}_n \rangle$ and $W_2 = \langle \mathbf{e}_{n+1}, \ldots \mathbf{e}_{2n} \rangle$ are Lagrangian subspaces and $V = W_1 \oplus W_2$.

Example 3.5. Using the notation of Example 2.1, if $\mathbf{z} = (\mathbf{q}_1, \mathbf{q}_2, \mathbf{p}_1, \mathbf{p}_2)$ and $\mathbf{Z} = (\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{P}_1, \mathbf{P}_2) \in \mathbb{R}^8$, then the canonical symplectic form is

$$\mathbf{J}\mathbf{z} \cdot \mathbf{Z} = \mathbf{q}_1 \cdot \mathbf{P}_1 + \mathbf{q}_2 \cdot \mathbf{P}_2 - \mathbf{Q}_1 \cdot \mathbf{p}_1 - \mathbf{Q}_2 \cdot \mathbf{p}_2.$$

Definition 3.6. Let (V, ω) be a symplectic vector space. A linear map $T : V \to V$ is called a *linear symplectic transformation* if

$$\omega(T\mathbf{u}, T\mathbf{v}) = \omega(\mathbf{u}, \mathbf{v}).$$

The matrix $\mathbf{A}$ associated with T in a symplectic basis is called a *symplectic matrix*.

We observe that T is a linear symplectic transformation if and only if any associated symplectic matrix $\mathbf{A}$ satisfy $\mathbf{A}^t\mathbf{J}\mathbf{A} = \mathbf{J}$.

Symplectic matrices have several nice properties. For instance, their determinant is equal to one, their eigenvalues occur in quadruples $\{\lambda, \lambda^{-1}, \bar{\lambda}, \bar{\lambda}^{-1}\}$, and their transposes are also symplectic.

The set of symplectic matrices of $\mathbb{R}^{2n}$, denoted by $Sp(n)$, is a Lie group. It is easy to see that $Sp(1) = Sl(2)$, the Lie group of square matrices of order two with determinant equal to one. For bigger degrees of freedom we only have the proper inclusion.

Example 3.7. Let us continue with Example 3.4. The subspaces W_1 and W_2 are isomorphic to $\mathbb{R}^n$. Let $\mathcal{S} : W_1 \to W_1$ be a linear isomorphism and let $\mathbf{B}$ be the associated matrix. If $\mathbf{B}^{-t} = (\mathbf{B}^t)^{-1}$, thus

$$\mathbf{A} = \begin{pmatrix} \mathbf{B} & \mathbf{0} \\ \mathbf{0} & \mathbf{B}^{-t} \end{pmatrix}$$

is the associated matrix in the symplectic basis of a symplectic linear transformation defined in V.

Now we extend the concept of symplectic transformation to nonlinear functions.

Let $\mathcal{O}$ be an open set in $\mathbb{R}^{2n}$ and $\Gamma : \mathcal{O} \subset \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ be a smooth function. Then $\Gamma(\mathbf{z})$ is a *symplectic change of coordinates* is a diffeomorphism (i.e., $\Gamma(\mathbf{z})$ is one to one and smooth, and its inverse is also one to one and smooth) and its derivate $D[\Gamma(\mathbf{z})]$ is a linear symplectic transformation.

In other words, the derivate $D[\Gamma(\mathbf{z})]$ satisfies

$$D[\Gamma(\mathbf{z})]\,\mathbf{J}\,D[\Gamma(\mathbf{z})]^t = \mathbf{J}. \tag{18.6}$$

Let us observe that if $T : V \to V$ is a linear symplectic transformation then it is also a symplectic change of coordinates.

It is usual to write the property (18.6) using differential forms. Let $\mathbf{Q} = \mathbf{Q}(\mathbf{q}, \mathbf{p})$ and $\mathbf{P} = \mathbf{P}(\mathbf{q}, \mathbf{p})$ be a change of variables defined in a simple connected open set of $\mathbb{R}^n$. The change of variable is symplectic if and only if

$$\sum_{k=1}^{n} dq_k \wedge dp_k = \sum_{k=1}^{n} dQ_k \wedge dP_k.$$

This is equivalent to the exactness of the differential one-form

$$\omega = \sum_{i=1}^{n} p_i dq_i - P_i dQ_i.$$

18.4 Symplectic Geometry and Hamiltonians

In this section we discuss several relations between the symplectic structure of $\mathbb{R}^{2n}$ and the Hamiltonian equations.

A basic fact of the symplectic geometry is that it preserves the Hamiltonians. That is, if $\mathbf{z} = \Gamma(\mathbf{u})$ is a symplectic change of coordinates, then Eq. (18.3) becomes the Hamiltonian equation:

$$\dot{\mathbf{u}} = \mathbf{J}\nabla G(\mathbf{u}), \qquad \text{where } G = H \circ \Gamma.$$

In other words, the Hamiltonian obtained after a symplectic transformation is the one formed by replacing the old coordinates by the new ones. In addition, the Poincaré maps of Hamiltonian systems are symplectic changes of coordinates: if $\boldsymbol{\varphi}_H(t, \mathbf{z})$ is a Hamiltonian flow, then the time t_0-map given by $P(\mathbf{z}) = \boldsymbol{\varphi}_H(t_0, \mathbf{z})$ is a symplectic change of coordinates.

Some deeper relations between Hamiltonian systems and symplectic maps are expressed through the *Poisson bracket* defined by

$$\{F, G\} = \nabla F \cdot J\nabla G = \sum_{k=1}^{n} \left[\frac{\partial F}{\partial q_k} \frac{\partial G}{\partial p_k} - \frac{\partial F}{\partial p_k} \frac{\partial G}{\partial q_k} \right],$$

where $F, G : \mathbb{R}^{2n} \to \mathbb{R}$ are smooth functions.

The following properties are easily seen from the definition of the Poisson bracket.

– Skew-symmetric:

$$\{F, G\} = -\{G, F\}.$$

In particular, this implies that $\{F, F\} = 0$.

– Linearity:

$$\{\alpha F + \beta G, H\} = \alpha\{F, H\} + \beta\{G, H\}, \quad \text{for all} \quad \alpha, \beta \in \mathbb{R}.$$

– Jacobi identity:

$$\{F, \{G, H\}\} + \{G, \{H, F\}\} + \{H, \{F, G\}\} = 0.$$

The Poisson brackets provide us with an effective tool to study the integrals and other things related with Hamiltonians and symplectic geometry.

For the flow $\boldsymbol{\varphi}_H(t, \mathbf{z})$ of (18.3) we have

$$\begin{aligned} \frac{d}{dt} F[\boldsymbol{\varphi}(t, \mathbf{z})] &= \nabla F \cdot \frac{d}{dt} \boldsymbol{\varphi}(t, \mathbf{z}) \\ &= \nabla F \cdot \mathbf{J} \nabla H \left(\boldsymbol{\varphi}(t, \mathbf{z})\right) = \{F, H\} \left[\varphi(t, \mathbf{z})\right]. \end{aligned} \tag{18.7}$$

Hence the $\{F, H\}$ is the rate of change of the function F along the solutions of the Hamiltonian system associated with H. Since the Poisson bracket is skew-symmetric then $\{F, H\}$ is also the negative of the rate of change of the function H along the solutions of the Hamiltonian system associated with F. In addition, the Poisson brackets are preserved by symplectic transformations.

An *integral* is a smooth function $F : \mathbb{R}^{2n} \to \mathbb{R}$ which is constant along the solutions of (18.3); i.e., $F(\varphi(t)) = F(\mathbf{z})$ is constant for all $t \in \mathbb{R}$. The following theorem gives us some relations between the Poisson bracket and the integrals of a Hamiltonian system.

Theorem 4.1. *Let F, G, and H smooth functions be defined in the same subset of* $\mathbb{R}^{2n}$. *Then*

1. *F is an integral for (18.3) if and only if* $\{F, H\} = 0$.
2. *H is an integral for (18.3).*
3. *If F and G are integrals for (18.3), then* $\{F, G\} = 0$.

It is easy to determine if a given function F is an integral of a Hamiltonian system by computing its Poisson bracket with the Hamiltonian function; if it is identically zero, then F is an integral. However the selection of the possible candidates is very tricky. As an example the identity (18.2) reveals us that the Hamiltonian itself is an integral.

Now let us assume that the Hamiltonian function does not contain q_1, then $\frac{dp_1}{dt} = 0$, which means that p_1 is constant along the orbits of the system, in other words, p_1 is an integral of the Hamiltonian. In addition, the rest of the system can be thought as a system with a one less degree of freedom and a parameter p_1.

Other way to see the previous discussion is that the Lie group of displacements in the variable q_1 does not affect the Hamiltonian system, and it is related to the existence of the integral p_1. Let us note that this group is generated by the flow of the Hamiltonian $F(q_1, \dots, p_1, \dots) = p_1$.

The variables such as q_1 in the previous example that are not in the Hamiltonian are called *cyclic* and allow a reduction of the number of degrees of freedom.

Two integrals are in *involution* if their Poisson bracket is zero. The Liouville theorem essentially states that a Hamiltonian dynamical system of n degrees of freedom and with n functions F_i in involution, and linearly independent (their gradients are linearly independent) can be solved by quadratures.

Theorem 4.2 (Liouville). *Suppose that we are given n functions in involution on a symplectic 2n-dimensional phase space*

$$F_1, \dots, F_n, \qquad \{F_i, F_j\} = 0.$$

Consider a level set of the functions F_i given by

$$M_c = \{(\mathbf{q}, \mathbf{p}) \in M \subset \mathbb{R}^{2n} : \quad F_i = c_i, \quad i = 1, \dots n\}.$$

Assume that the n functions F_i are independent of M_c . In other words, the gradients DF_i are linearly independent at each point of M_c. Then

1. *M_c is a smooth manifold, invariant under the flow with $H = H(F_i)$.*
2. *If the manifold M is compact and connected, then it is diffeomorphic to the n-dimensional torus $\mathbb{T}^n = \{(\phi_1, \dots, \phi_n) \mod 2\pi\}$.*
3. *The phase flow with the Hamiltonian function H determines a conditionally periodic motion on M_c , i.e. in angular variables*

$$\frac{d\phi_i}{dt} = \alpha_i, \qquad \alpha_i = \alpha_i(F_j).$$

We end this section with a brief discussion of the Noether theorem in the context of Hamiltonian systems. This theorem states that every conservation law or integral in Hamiltonian systems comes from a symmetry: The conservation of energy comes from the invariance of the Hamiltonians with respect to time. The invariance of a Hamiltonian with respect to spatial translations implies the conservation of linear momentum, and a rotational invariance implies the conservation of the angular momentum.

Let $F : D \subset \mathbb{R}^{2n} \to \mathbb{R}$ be a Hamiltonian with flow $\psi(t, \mathbf{z})$. It is well known that the flow is a local action of group in $D \subset \mathbb{R}^{2n}$, thus:

– $\psi(0, \mathbf{z}) = \mathbf{z}$.

– $\psi(t + s, \mathbf{z}) = \psi(t, \psi(s, \mathbf{z}))$, for all t and s where this expression is defined.

The flow $\psi(t, \mathbf{z})$ is a *symplectic symmetry* of the Hamiltonian $H : D \subset \mathbb{R}^{2n} \to \mathbb{R}$ if

$$H(\mathbf{z}) = H(\psi(t, \mathbf{z})) \text{ for all } t \in \mathbb{R} \text{ and } \mathbf{z} \in D. \tag{18.8}$$

Theorem 4.3 (Noether). *If $\psi(t, \mathbf{z})$ is a symplectic symmetry of the the Hamiltonian $H : D \subset \mathbb{R}^{2n} \to \mathbb{R}$, then F is an integral of the Hamiltonian system (18.3).*

18.5 The Mathieu Transformation

In this section we study a special type of symplectic transformation introduced by E. Mathieu. It is a generalization of Example 3.7, see [5, 9].

We start with a diffeomorphism $f : D \subset \mathbb{R}^n \to \mathbb{R}^n$, defined in an open and simple connected set D. The Mathieu transformation is

$$\mathbf{Q} = f(\mathbf{q}), \qquad \mathbf{P} = (Df)^{-t}\mathbf{p}, \tag{18.9}$$

where the Jacobian matrix is Df and $(Df)^{-t}$ denotes its inverse transposed matrix. The Mathieu transformation coincides with the original diffeomorphism $f(\mathbf{q})$ in the configuration space, and is extended to the momenta variables with the aim to obtain a symplectic change of coordinates. There are several possible extensions, but the Mathieu transformation is the simpler.

The rest of this section will be devoted to apply the Mathieu transformation to the two-body problem. We start by introducing the change of coordinates

$$\mathbf{Q}_1 = m_1\,\mathbf{q}_1 + m_2\,\mathbf{q}_2, \qquad \mathbf{Q}_2 = \mathbf{q}_2 - \mathbf{q}_1.$$

Hence the Mathieu transformation becomes completed with

$$\mathbf{P}_1 = \mathbf{p}_1 + \mathbf{p}_2, \qquad \mathbf{P}_2 = -m_2\,\mathbf{p}_1 + m_1\,\mathbf{p}_2.$$

The inverse transformation is given by

$$\mathbf{q}_1 = \mathbf{Q}_1 - m_2\,\mathbf{Q}_2, \;\; \mathbf{q}_2 = m_1\,\mathbf{Q}_2 + \mathbf{Q}_1,$$
$$\mathbf{p}_1 = m_1\,\mathbf{P}_1 - \mathbf{P}_2, \;\; \mathbf{p}_2 = \mathbf{P}_2 + m_2\,\mathbf{P}_1.$$

Then the Hamiltonian (18.5) in the new coordinates becomes

$$H(\mathbf{Q}_1, \mathbf{Q}_2, \mathbf{P}_1, \mathbf{P}_2) = \underbrace{\frac{1}{2}\mathbf{P}_1 \cdot \mathbf{P}_1}_{H_1(\mathbf{Q}_1, \mathbf{P}_1)} + \underbrace{\frac{1}{2\,m_1 m_2}\mathbf{P}_2 \cdot \mathbf{P}_2 - \frac{m_1 m_2}{|\mathbf{Q}_2|}}_{H_2(\mathbf{Q}_2, \mathbf{P}_2)}, \tag{18.10}$$

and the system equations yields

$$\dot{\mathbf{Q}}_1 = \mathbf{P}_1, \qquad \dot{\mathbf{P}}_1 = \mathbf{0}, \tag{18.11a}$$

$$\dot{\mathbf{Q}}_2 = \frac{1}{m_1 m_2}\mathbf{P}_2, \qquad \dot{\mathbf{P}}_2 = -m_1 m_2 \frac{\mathbf{Q}_2}{|\mathbf{Q}_2|^3}. \tag{18.11b}$$

The system has been simplified in several ways, first of all the new Hamiltonian is a sum of two independent functions $H_1(\mathbf{Q}_1, \mathbf{P}_1)$ and $H_2(\mathbf{Q}_2, \mathbf{P}_2)$. This fact produces two independent Hamiltonian equations, each one of two degrees of freedom.

Equation (18.11a) has a very simple solution: $\mathbf{Q_1} = \mathbf{c}_1 t + \mathbf{c}_2$, $\mathbf{P_1} = \mathbf{c}_1$. The reason of this simplicity is that we use an integral of the problem as the definition of $\mathbf{Q}_1$.

In order to solve Eq. (2.1) it suffices to solve the Hamiltonian equation (18.11b), called the Kepler problem. Therefore the problem has been reduced from a four degrees of freedom problem into one with only two degrees of freedom. To simplify the notation from here on we omit the subindex "2" in all the variables of the Kepler Hamiltonian. Thus the Hamiltonian associated with system (18.11b) takes the form $H(\mathbf{Q}, \mathbf{P}) = \frac{1}{2m_1 m_2}\mathbf{P}\cdot\mathbf{P} - \frac{m_1 m_2}{|\mathbf{Q}|}$ and has a rotational symmetry. The best way to take advantage of this fact is to use polar coordinates.

Let $\mathbf{Q} = (r\cos\theta, r\sin\theta)$, and using the Mathieu transformation we obtain $\mathbf{P} = R(\cos\theta, \sin\theta) + \frac{\Theta}{r}(-\sin\theta, \cos\theta)$, where Θ is the angular momentum and is conjugate to θ, and R is the conjugate momentum of r. Then the Hamiltonian is

$$H(r, \theta, R, \Theta) = \frac{1}{2m_1 m_2}(R^2 + \frac{\Theta^2}{r^2}) - \frac{m_1 m_2}{r}. \tag{18.12}$$

Let us remark that the variable θ is cyclic, thus $\Theta = c$ is constant. The equations of motion are

$$\dot{r} = R, \qquad \dot{R} = \frac{\Theta^2}{2m_1 m_2 r^3} - \frac{m_1 m_2}{r^2}, \tag{18.13a}$$

$$\dot{\theta} = \frac{\Theta}{r^2}, \qquad \dot{\Theta} = 0. \tag{18.13b}$$

Hence (18.13a) is the following one degree Hamiltonian equation parametrized by the constant c:

$$\ddot{r} = \dot{R} = \frac{c^2}{2m_1 m_2 r^3} - \frac{m_1 m_2}{r^2}. \tag{18.14}$$

If we assume that $c \neq 0$, then the motion is not collinear. Making the change of variable $u = 1/r$ and rescaling the time by $dt = (r^2/c)d\theta$, it becomes

$$u'' + u = \frac{m_1 m_2}{c^2}, \tag{18.15}$$

where $' = d/d\theta$. It is just a nonhomogeneous harmonic oscillator with general solution

$$u = \frac{m_1 m_2}{c^2}(1 + e\cos f), \tag{18.16}$$

where e and g are constants and $f = \theta - g$. Substituting back for r we have

$$r = \frac{c^2/\mu}{1 + e\cos f}, \tag{18.17}$$

which is the general equation of a conic in polar coordinates, with a focus at the origin, $\mu = m_1 m_2$, e is the *eccentricity*, and c^2/μ is the *semilatus rectum*. The four possible conics are: circle if $e = 0$, an ellipse $0 < e < 1$, a parabola $e = 1$, and $e > 1$ for a hyperbola. The angle f is called the *true anomaly*, θ the *true longitude* and g the argument of the perihelio (perigee).

18.6 Recommendations for Further Reading

The object of study of this paper has a long history that goes back to Newton and has the names of Lagrange, Poincaré, and many others attached to it.

Nowadays symplectic coordinates is a wide area of research that that links many disciplines of mathematics and physics. The reference [6] is similar in level and scope to this work and is the natural choice to continue studying.

There is an alternative way to proceed by using calculus of variations. This field of mathematics studies methods for finding functions that minimize functionals defined by integrals. The basic one is analogous in infinite dimensions to the first derivative test and is known as the Euler–Lagrange equation.

Equation (18.3) is the Euler–Lagrange equation of the so-called action functional. Hence the orbits are extremals of it. This is the "principle of least action." Lagrangian mechanics is a formulation of the mechanics using this principle. It is not completely equivalent to the Hamiltonian mechanics, but there are many important facts that intertwine both. The references [1, 2] contain a discussion of this facts.

On the other hand, an important topic is the study of the symmetries of Hamiltonian or Lagrangian systems. It is expressed in the Lie groups theory and the Noether theorem. It can be reviewed in the reference [7].

We strongly suggest readers to look at [3] for a well-written explanation on symplectic geometry. This book assumes the reader has only a general background in analysis and familiarity with linear algebra, and includes extensive appendices which provide background material on vector bundles, on cohomology, and on Lie

groups and Lie algebras and their representations. It starts with the basics of the geometry of symplectic vector spaces, and then, symplectic manifolds are defined and explored.

Finally, the references [4, 8, 9] are classic books but offer a nice presentation of the analytical mechanics, as well as the application of variational methods and dynamic systems in the study of mechanical systems, in particular of the n-body problem.

Acknowledgements The authors were partially supported by the grant: Red de cuerpos académicos Ecuaciones Diferenciales. Proyecto sistemas dinámicos y estabilización. PROMEP 2011-SEP, México.

References

1. Abraham, R., Marsden, J.E.: Foundations of Mechanics. Benjamin/Cummings, Reading (1978)
2. Arnold, V.I.: Mathematical Methods of Classical Mechanics. Springer, New York (1978)
3. Berndt, R.: An Introduction to Symplectic Geometry. Graduate Studies in Mathematics, vol. 26. American Mathematical Society, Providence (2001)
4. Golstein, H., Poole, C., Safko, J.: Classical Mechanics, 3rd edn. Addison Wesley, Reading (2001)
5. Mathieu, E.: Mémoire sur les équations différentielles canoniques de la mécanique. J. Math. **XIX**, 265–306 (1874)
6. Meyer, K.R., Hall, G.R., Offin, D.: Introduction to Hamiltonian Dynamical Systems and the N-Body Problem. Applied Mathematical Sciences, vol. 90. Springer, New York (2009)
7. Olver, P.J.: Applications of Lie Groups to Differential Equations. Graduate Text in Mathematics. Springer, New York (2000)
8. Szebehely, V.: Theory of Orbits. The Restricted Problem of Three Bodies. Academic, New York (1967)
9. Whittaker, E.: A Treatise on the Analytical Dynamics of Particles and Rigid Bodies. Cambridge University Press, Cambridge (1904)

Chapter 19
Parallelized Solution of Banded Linear Systems with an Introduction to p-adic Computation

Anthony A. Ruffa, Michael A. Jandron, and Bourama Toni

Abstract We present an approach that supports a parallelized solution of banded linear systems without communication between processors. We do this *by adding unknowns to the system* equal to the number of superdiagonals q. We then perform r forward substitution processes in parallel (where r is the number of nonzero terms in the right-hand side vector), and superimpose the resulting solution vectors. This leads to the determination of the extra unknowns, and by extension, to the overall solution. However, some systems exhibit exponential growth behavior during the forward substitution process, which prevents the approach from working. We present several modifications to address this, extending the approach (in a modified form) to be used for general systems. We also extend it to block banded systems. Numerical results for well-behaved test systems show a speedup of 20–80 over conventional solvers using only 8 processors. Theoretical estimates assuming q processors demonstrated a speedup of a factor exceeding 300 for 10^5 unknowns when $q = 2000$; for 10^9 unknowns, the speedup exceeds a factor of 10^4 when $q = 45{,}000$. We also introduce some fundamentals of p-adic computation and *modular arithmetic* as the basis of the development and implementation of a fully parallel p-adic linear solver, which allows *error-free* computation over the rational numbers, and is better suited to control coefficient growth.

Keywords Banded systems • Parallel solver • Forward substitution • Backward substitution • p-adic arithmetic • Hensel codes

Mathematics Subject Classification (2010): 15A06, 65F05

A.A. Ruffa (✉) • M.A. Jandron
Naval Undersea Warfare Center, Newport, RI 02841, USA
e-mail: anthony.ruffa@navy.mil; michael.jandron@navy.mil

B. Toni
Department of Mathematics & Economics, Virginia State University, Petersburg, VA 23806, USA
e-mail: btoni@vsu.edu

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_19

19.1 Introduction

There is a great need to solve larger linear systems. Although advancements in computer hardware have supported a tremendous growth in the size of systems that can be solved, scientists and engineers continue to develop larger and higher fidelity models. It is not uncommon to develop models that cannot be solved with existing solvers, or models that can be solved, but the computation time is so large that they become impractical as a design tool.

One potential approach to meet the increasing need for larger models involves the development of methods that can support parallel computing. However, common solvers for tridiagonal systems (the simplest banded systems resulting from physics-based models) use a variant of Gaussian elimination combined with backward substitution (e.g., LU decomposition [1, 2]) that does not support full parallelization [3–6].

We take an alternative approach. Instead of attempting to develop methods to parallelize existing approaches on general matrices, we limit our focus to Hessenberg and banded systems. We show that the simplest such system, a tridiagonal system, can be solved in a fully parallel manner (i.e., without any communication between processors) *by adding an unknown to the system* [7].

We also break up the right-hand side (RHS) vector into r vectors, each limited to one nonzero term. In a further step that seems counterintuitive, we compute a solution vector corresponding to each of the r RHS vectors by assuming a value for the first term and then computing the remaining terms (including the extra added term) via forward substitution. Finally, we superimpose all of the solution vectors, leading to an equation to determine the extra introduced unknown, and by extension, to determine the solution to the overall system.

We can extend this approach to banded systems having an arbitrary number of superdiagonals q by adding q unknowns. For well-behaved test systems, we have shown that this approach can lead to a significant speedup over existing solvers. Theoretical estimates of the number of operations indicate potential speedups of $O(10^2)$ to $O(10^4)$ over present solvers under ideal conditions [8]. However, not all systems are well-behaved. The solution vectors for some systems exhibit exponential growth behavior when computed via forward substitution that requires a modification to the approach. Here we will outline the original approach and modifications that address the observed exponential growth behavior. We also extend the approach to block banded systems that arise from the finite element method.

The next section discusses tridiagonal systems, along with a summary of theoretical concepts supporting the approach and its implementation, considering systems exhibiting, respectively, non-exponential and exponential behavior. In Sect. 19.3, we consider pentadiagonal systems, again for both non-exponential and exponential behavior. That section also discusses the modular solution and its application to active vibration suppression. Section 19.4 presents periodic systems, in particular a

one-dimensional Helmholtz problem with periodic boundary conditions. General banded systems and block banded systems are covered in Sects. 19.5 and 19.6, respectively. In Sect. 19.7, we develop the p-adic method for a parallel linear solver, to include basic notions and examples of p-adic analysis. We finally conclude with Sect. 19.8.

19.2 Tridiagonal Systems

Tridiagonal systems provide a good starting point to demonstrate the approach and illustrate some of the issues that arise when implementing it. We will begin with "well-behaved" tridiagonal systems and then investigate systems that are not as well-behaved, in the sense that the forward substitution process leads to exponential growth behavior. We will then develop variants to the algorithm to solve the more general systems.

19.2.1 Tridiagonal Systems Exhibiting Non-exponential Behavior

19.2.1.1 Theory

Let

$$Ax = d, \tag{19.1}$$

where

$$A = \begin{bmatrix} b_1 & c_1 & 0 & \cdots & 0 \\ a_2 & b_2 & c_2 & \ddots & \vdots \\ 0 & a_3 & b_3 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & \cdots & 0 & a_n & b_n \end{bmatrix} \in \mathbb{R}^{n\times n}$$

is a general nonsingular tridiagonal matrix with $c_k \neq 0 \ \forall \ k \in [1, n-1]$, $x = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$, and RHS vector $d = [d_1, d_2, \ldots, d_n]^T \in \mathbb{R}^n$. The approach here follows that developed by Ruffa [7] and Jandron et al. [8].

We first summarize the theoretical concepts supporting a rigorous validation of the computational approach. In order to prepare for parallelization we consider an auxiliary system whose range includes, as a subset, the range $R(A)$ of system (19.1). For such a general nonsingular system, there exists a unique solution $x \in \mathbb{R}^n$ given by $x = A^{-1}d$, i.e., $d \in R(A)$. We write the auxiliary system as

$$By = \tilde{b}, \tag{19.2}$$

where B is $n \times m,\ m > n$, mapping $\mathbb{R}^m$ into $\mathbb{R}^n$. The auxiliary system is consistent iff $\tilde{b} \in R(B) = span(col(B))$, where $col(B)$ denotes the subspace of the column vectors of B. To recover the original solution from a consistent auxiliary system requires $R(A) \subset R(B)$, i.e., the range $R(A)$ decomposes into

$$R(B) = R(A) + C \tag{19.3}$$

for C a complementary subset to $R(A)$ in $\mathbb{R}^n$, i.e., $\dim R(A) + \dim C = \dim R(B)$. This leads to seeking an appropriate augmented matrix $\tilde{A}$ in the form

$$\tilde{A} = \begin{bmatrix} A\ \tilde{e} \end{bmatrix} \tag{19.4}$$

whose range $R(B) = R(A) + span(\tilde{e})$. The dimension of $span(\tilde{e})$ is ad hoc, e.g., in the tridiagonal case, $\dim span(\tilde{e}) = 1$, with $\tilde{e}$ sets to $\tilde{e} = \tilde{e}_i$ any of the basis vector, without loss of generality. Here we actually set $\tilde{e} = \tilde{e}_n$, the augmented matrix $\tilde{A}$ is $n \times (n+1)$ mapping $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$. That is, the solution vector x of the original system is the projection of the solution

$$y = \begin{bmatrix} x \\ \alpha \end{bmatrix} \tag{19.5}$$

for an arbitrary $\alpha \in \mathbb{R}$. Additional requirements follow in order for the original RHS vector d to be in $R(\tilde{A})$, i.e., we take $\tilde{b} = d + \alpha\tilde{e}$.

19.2.1.2 Implementation

Following [8], we add an unknown $\alpha \in \mathbb{R}$ to x and a column $\bar{e}_n \in \mathbb{R}^n$ to A where $\bar{e}_n = [0, \ldots, 0, 1]^T$ to obtain an equivalent system to (19.1), i.e.,

$$\begin{bmatrix} A\ \bar{e}_n \end{bmatrix} \begin{bmatrix} x \\ \alpha \end{bmatrix} = d + \alpha\bar{e}_n. \tag{19.6}$$

We decompose (19.6) into two independent sub-systems, i.e.,

$$\begin{bmatrix} A\ \bar{e}_n \end{bmatrix} \begin{bmatrix} u \\ \alpha_1 \end{bmatrix} = d, \tag{19.7}$$

$$\begin{bmatrix} A\ \bar{e}_n \end{bmatrix} \begin{bmatrix} v \\ \alpha_2 \end{bmatrix} = \bar{e}_n, \tag{19.8}$$

where $u = [u_1, u_2, \ldots, u_n]^T \in \mathbb{R}^n$, $v = [v_1, v_2, \ldots, v_n]^T \in \mathbb{R}^n$, and $\alpha_1, \alpha_2 \in \mathbb{R}$, and we write the solution of (19.6) in the following form:

$$\begin{bmatrix} x \\ \alpha \end{bmatrix} = \begin{bmatrix} u \\ \alpha_1 \end{bmatrix} + \alpha \begin{bmatrix} v \\ \alpha_2 \end{bmatrix}. \tag{19.9}$$

The solution to (19.7)–(19.8) can be determined with a forward substitution approach [8], and the last row in (19.9) can then be used to determine α, i.e.,

$$\alpha = \frac{\alpha_1}{1 - \alpha_2}, \tag{19.10}$$

where $\alpha_2 \neq 1$. Substitution of α into (19.9) leads to the solution to (19.1).

The number of calculations is bounded by $5n$ (i.e., for optimal parallelization) and $12n$ (i.e., for sequential execution) [8]. Numerical experiments demonstrated a speedup of 20–80 over the PARDISO solver [9–11] (depending on n) for the tridiagonal Toeplitz system having row structure $[1, -2, 1]$.

19.2.2 Tridiagonal Systems Exhibiting Exponential Behavior

This approach will not work for all systems. For example, consider a tridiagonal Toeplitz system having row structure $[1, -3, 1]$. The forward substitution approach will lead to solution vectors dominated by exponential growth behavior so that the magnitude of individual terms can reach 10^{300} when $n = O(10^3)$. This will prevent the approach from working.

This behavior has been previously observed [12–16]. To understand the underlying mechanisms, we will focus on tridiagonal Toeplitz systems, which admit an analytical solution. We can find the analytical solution by observing that the tridiagonal Toeplitz system with row structure $[1, -3, 1]$ can result from the discretization of a continuous system, e.g., one having the form

$$\Phi(z - h) - 3\Phi(z) + \Phi(z + h) = 0, \tag{19.11}$$

when the continuous independent variable z is discretized with equal spacing h. We can assume solutions to (19.11) having the form

$$\Phi(z) = Ae^{\zeta z}. \tag{19.12}$$

This leads to

$$e^{\zeta(z-h)} - 3e^{\zeta z} + e^{\zeta(z+h)} = 0, \tag{19.13}$$

or

$$e^{-\zeta h} - 3 + e^{\zeta h} = 0. \tag{19.14}$$

The roots are

$$\zeta h = \pm 0.9624. \tag{19.15}$$

Since h is arbitrary, we choose $h = 1$ for simplicity. Now consider a tridiagonal Toeplitz system $Ax = d$ with row structure $[1, -3, 1]$ and a single nonzero RHS

term d_m. The solution is

$$x_k = \frac{v_k d_m}{v_{m+1} - 3v_m + v_{m-1}}, \tag{19.16}$$

where

$$v_k = e^{\zeta(k-m)}; k \leq m, \tag{19.17}$$

$$v_k = e^{\zeta(m-k)}; k \geq m. \tag{19.18}$$

This solution is valid unless $m \approx 1$ or $m \approx n$. Since our purpose is to use the analytical solution to understand the exponential behavior exhibited in the numerical solution, we will intentionally avoid those cases. (We could compute them, but they are more complicated.) With these restrictions, v_k depends only on $k-m$, regardless of n or m (Fig. 19.1). The error is $O(10^{-16})$. It is similar to the solution developed by Meek [17] that uses ratios of determinants having terms derived from the difference equation corresponding to the Toeplitz system. However, we use a different notation that is more appropriate for use with a RHS vector that has a single nonzero term.

The forward substitution process cannot capture (19.18), i.e., the exponential decay component of the solution. However, the exponential growth behavior exhibited in the solution vector is in agreement with (19.17) for $k \leq m$, indicating that it is an exact solution. As a result, we can conclude that the exponential behavior is not due to numerical stability issues, but is a consequence of the forward substitution approach.

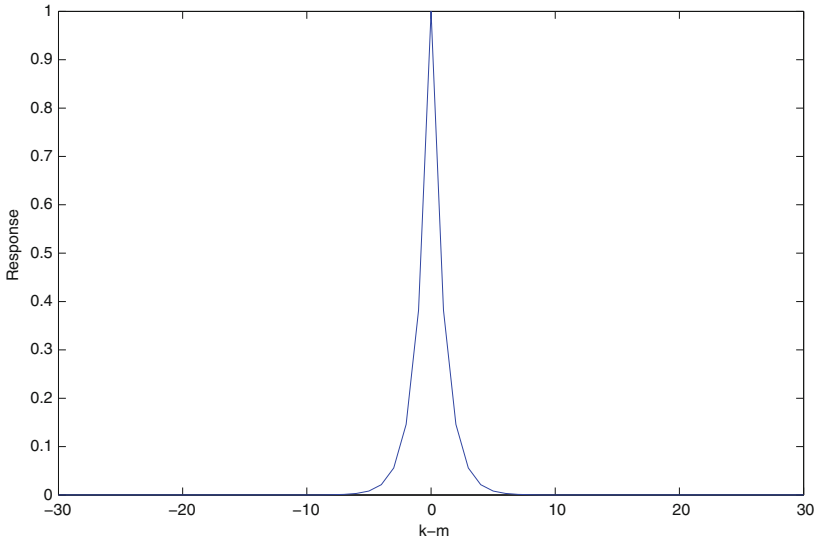


Fig. 19.1 Analytical solution of the $[1, -3, 1]$ Toeplitz system (normalized to 1)

The analytical solution suggests a modification to the numerical approach: forward and backward substitution. Here the forward substitution process is executed for $1 \leq k \leq m-1$ and the backward substitution process assumes that $x_n = 1$ and then is executed for $m+1 \leq k \leq n$. The exponential growth behavior exhibited in the forward substitution process becomes exponential decay behavior when backward substitution is performed, so that the problem resolves itself.

Implementing this involves splitting the system and performing a forward substitution from $k = [1, m]$ and a backward substitution from $k = [m, n]$, where m is the split point. The forward substitution and backward substitution operations can be performed in parallel.

Although this approach will work for any tridiagonal system, it can only accommodate one nonzero RHS term for each solution vector when exponential growth is present. Solving a system with an arbitrary RHS vector would require breaking it up into r RHS vectors (each having one nonzero term), solving them in parallel (ideally with r processors), and then obtaining the solution by superposition.

We can adapt the approach developed by Jandron et al. [8] to decompose (19.1) into two sub-systems at $k = m$ for arbitrary k. We require $c_k \neq 0 \ \forall \ k \in [1, m-1]$ and $a_k \neq 0 \ \forall \ k \in [m+1, n]$ and decompose the problem as follows:

$$A^{(1)} \begin{bmatrix} v^{(1)} \\ \alpha^{(1)} \end{bmatrix} = \mathbf{0}, \tag{19.19}$$

$$A^{(2)} \begin{bmatrix} \alpha^{(2)} \\ v^{(2)} \end{bmatrix} = \mathbf{0}. \tag{19.20}$$

where $v^{(1)} = [v_1^{(1)}, v_2^{(1)}, \ldots, v_{m-1}^{(1)}]^T \in \mathbb{R}^{m-1}$, $v^{(2)} = [v_{m+1}^{(2)}, v_{m+2}^{(2)}, \ldots, v_n^{(2)}]^T \in \mathbb{R}^{n-m}$, $\alpha^{(1)}, \alpha^{(2)} \in \mathbb{R}$, and

$$A^{(1)} = \begin{bmatrix} b_1 & c_1 & 0 & \cdots & 0 & 0 \\ a_2 & b_2 & c_2 & \ddots & \vdots & \vdots \\ 0 & a_3 & b_3 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & c_{m-2} & 0 \\ 0 & \cdots & 0 & a_{m-1} & b_{m-1} & c_{m-1} \end{bmatrix} \in \mathbb{R}^{(m-1)\times m}, \tag{19.21}$$

and

$$A^{(2)} = \begin{bmatrix} a_{m+1} & b_{m+1} & c_{m+1} & 0 & \cdots & 0 \\ 0 & a_{m+2} & b_{m+2} & c_{m+2} & \ddots & \vdots \\ 0 & 0 & a_{m+3} & b_{m+3} & \ddots & 0 \\ 0 & \vdots & \ddots & \ddots & \ddots & c_{n-1} \\ 0 & 0 & \cdots & 0 & a_n & b_n \end{bmatrix} \in \mathbb{R}^{(n-m)\times(n-m+1)}. \tag{19.22}$$

We assume that $v_1^{(1)} = 1$ and $v_n^{(2)} = 1$ and solve for $v^{(1)} = [v_1^{(1)}, v_2^{(1)}, \cdots, v_{m-1}^{(1)}]^T$ and $v^{(2)} = [v_{m+1}^{(2)}, v_{m+2}^{(2)}, \ldots, v_n^{(2)}]^T$ using forward and backward substitution, respectively. For continuity, we normalize $[v^{(1)}, \alpha^{(1)}]^T$ and $[\alpha^{(2)}, v^{(2)}]^T$ to ensure $\alpha^{(1)} = \alpha^{(2)} = 1$. The solution is

$$x = \begin{bmatrix} \dfrac{v^{(1)}}{\alpha^{(1)}} \cdot \dfrac{d_m}{\alpha_3} \\ \dfrac{d_m}{\alpha_3} \\ \dfrac{v^{(2)}}{\alpha^{(2)}} \cdot \dfrac{d_m}{\alpha_3} \end{bmatrix}, \tag{19.23}$$

where $\alpha_3 = a_m \cdot \dfrac{v_{m-1}^{(1)}}{\alpha^{(1)}} + b_m + c_m \cdot \dfrac{v_{m+1}^{(2)}}{\alpha^{(2)}}$.

19.3 Pentadiagonal Systems

Extending the algorithm to pentadiagonal systems is straightforward when exponential behavior is not present [8]. Pentadiagonal systems exhibiting exponential behavior will require additional modifications to the algorithm.

19.3.1 Pentadiagonal Systems Exhibiting Non-exponential Behavior

Consider the pentadiagonal linear system

$$Ax = f, \tag{19.24}$$

where $A \in \mathbb{R}^{n \times n}$ is a general nonsingular pentadiagonal system, $x, f \in \mathbb{R}^n$, and

$$A = \begin{bmatrix} c_1 & d_1 & e_1 & 0 & \cdots & 0 \\ b_2 & c_2 & d_2 & e_2 & \ddots & \vdots \\ a_3 & b_3 & c_3 & d_3 & \ddots & 0 \\ 0 & a_4 & b_4 & \ddots & \ddots & e_{n-2} \\ \vdots & \ddots & \ddots & \ddots & c_{n-1} & d_{n-1} \\ 0 & \cdots & 0 & a_n & b_n & c_n \end{bmatrix}, \tag{19.25}$$

where $e_k \neq 0$. Following [8], we add *two* unknowns, $\alpha_1, \alpha_2 \in \mathbb{R}$, to x, i.e.,

$$\begin{bmatrix} A\,\bar{e}_{n-1}\,\bar{e}_n \end{bmatrix} \begin{bmatrix} x \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = f + \alpha_1\,\bar{e}_{n-1} + \alpha_2\,\bar{e}_n, \tag{19.26}$$

where $\bar{e}_{n-1}, \bar{e}_n \in \mathbb{R}^n$ represent the $n-1$ and n basis vectors, respectively. We then decompose (19.26) into *three* linear independent sub-systems, i.e.,

$$\begin{bmatrix} A\,\bar{e}_{n-1}\,\bar{e}_n \end{bmatrix} \begin{bmatrix} v^{(1)} \\ \alpha_1^{(1)} \\ \alpha_2^{(1)} \end{bmatrix} = f, \tag{19.27}$$

$$\begin{bmatrix} A\,\bar{e}_{n-1}\,\bar{e}_n \end{bmatrix} \begin{bmatrix} v^{(2)} \\ \alpha_1^{(2)} \\ \alpha_2^{(2)} \end{bmatrix} = \bar{e}_{n-1},$$

$$\begin{bmatrix} A\,\bar{e}_{n-1}\,\bar{e}_n \end{bmatrix} \begin{bmatrix} v^{(3)} \\ \alpha_1^{(3)} \\ \alpha_2^{(3)} \end{bmatrix} = \bar{e}_n.$$

where $v^{(1)} = [v_1^{(1)}, v_2^{(1)}, \ldots, v_n^{(1)}]^T \in \mathbb{R}^n$, $v^{(2)} = [v_1^{(2)}, v_2^{(2)}, \ldots, v_n^{(2)}]^T \in \mathbb{R}^n$, and $v^{(3)} = [v_1^{(3)}, v_2^{(3)}, \ldots, v_n^{(3)}]^T \in \mathbb{R}^n$, and all α terms $\in \mathbb{R}$. Superposition leads to

$$\begin{bmatrix} x \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} v^{(1)} \\ \alpha_1^{(1)} \\ \alpha_2^{(1)} \end{bmatrix} + \alpha_1 \begin{bmatrix} v^{(2)} \\ \alpha_1^{(2)} \\ \alpha_2^{(2)} \end{bmatrix} + \alpha_2 \begin{bmatrix} v^{(3)} \\ \alpha_1^{(3)} \\ \alpha_2^{(3)} \end{bmatrix}. \tag{19.28}$$

Note that the last two equations in (19.28) support the determination of α_1 and α_2, leading to the solution to (19.24).

19.3.2 Pentadiagonal Systems Exhibiting Exponential Behavior

Pentadiagonal systems can also exhibit exponential growth behavior. However, splitting the system and performing forward substitution from $k = [1, m]$ and backward substitution from $k = [m, n]$ will lead to solution vectors but will not satisfy three of the equations, i.e., those corresponding to $m-1 \leq k \leq m+1$. These equations can be satisfied by allowing the associated RHS terms to remain unconstrained; however, some means must then be developed to remove two of the three RHS terms in order to make the approach general.

To illustrate such an approach, we will focus on a pentadiagonal system that exhibits exponential behavior, specifically, a system resulting from a finite difference analysis of the beam vibration problem. This system has two roots that lead to exponential behavior and two that lead to oscillatory behavior. As a result, the forward/backward substitution approach developed for tridiagonal systems will emphasize the exponential behavior at the expense of the oscillatory behavior, preventing it from being applicable to general pentadiagonal systems. In addition to the problem of the three unconstrained RHS terms, this is the primary complication that can arise when extending the approach from tridiagonal to pentadiagonal systems, so methods that successfully address these complications will be valid for general pentadiagonal systems.

The beam vibration problem is governed by the Euler–Bernoulli equation, i.e.,

$$EI\frac{\partial^4 Y}{\partial s^4}+\rho\chi\frac{\partial^2 Y}{\partial t^2}=W. \tag{19.29}$$

In this example, $E = 2\times10^{11}\,\text{GPa}$ is the Young's modulus, I is the moment of inertia, $\rho = 8000\,\text{kg/m}^2$ is the density, χ is the cross-sectional area, W is the applied load, s and t are the independent spatial and temporal coordinates, respectively, and Y is the transverse displacement. A harmonic time dependence is assumed, i.e., $Y(s,t) = Y(s)e^{i\omega t}$, and a finite difference approximation is used, i.e.,

$$\frac{d^4Y}{ds^4}\approx\frac{x_{n-2}-4x_{n-1}+6x_n-4x_{n+1}+x_{n+2}}{\Delta s^4}. \tag{19.30}$$

This leads to

$$\frac{x_{n-2}-4x_{n-1}+6x_n-4x_{n+1}+x_{n+2}}{\Delta s^4}-\frac{\omega^2\rho\chi}{EI}x_n=0. \tag{19.31}$$

The beam is pinned on both ends, leading to the following boundary conditions:

$$x_0=0; \tag{19.32}$$

$$\frac{d^2x_0}{ds^2}\approx\frac{2x_0-5x_1+4x_2-x_3}{\Delta s^2}=0; \tag{19.33}$$

$$x_{n+1}=0; \tag{19.34}$$

$$\frac{d^2x_{n+1}}{ds^2}\approx\frac{2x_{n+1}-5x_n+4x_{n-1}-x_{n-2}}{\Delta s^2}=0. \tag{19.35}$$

Consider a square beam with a length of 0.1 m on each side (with $\Delta s = 0.0301\,\text{m}$) and vibrating at 100 Hz. The finite difference discretization length is $\lambda/100$ (where λ is the wavelength), and $n = 3901$, so the beam effectively contains 39 wavelengths. The RHS vector is zero except for the term f_m, where $m = 1951$ (i.e., at the center of the beam).

Equations (19.31)–(19.35) the boundary conditions are also part of the system represents a pentadiagonal Toeplitz system (except for the first and last equations) with a row structure of $[1, -4, 6+\gamma, -4, 1]$, where $\gamma = -\omega^2\rho\chi\Delta s^4/(EI)$. An analytical solution can be developed by finding the roots of the characteristic polynomial, leading to

$$x_k = C_1 e^{\sigma k} + C_2 e^{-\sigma k} + C_3 cos(\sigma k) + C_4 sin(\sigma k), \tag{19.36}$$

where $\sigma = 0.0628$ based on the parameters given. Denoting (19.31)–(19.35) as $Ax = f$, the coefficient matrix $A \in \mathbb{R}^{n\times n}$ is

$$\begin{bmatrix}
-5 & 4 & -1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
-4 & (6+\gamma) & -4 & 1 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\
1 & -4 & (6+\gamma) & -4 & 1 & 0 & \ddots & \ddots & \ddots & \vdots \\
0 & 1 & -4 & (6+\gamma) & -4 & 1 & 0 & \ddots & \ddots & \vdots \\
\vdots & 0 & 1 & -4 & (6+\gamma) & -4 & 1 & 0 & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \ddots & 1 & -4 & (6+\gamma) & -4 & 1 & 0 \\
\vdots & \ddots & \ddots & \ddots & 0 & 1 & -4 & (6+\gamma) & -4 & 1 \\
\vdots & \ddots & \ddots & \ddots & \ddots & 0 & 1 & -4 & (6+\gamma) & -4 \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & -1 & 4 & -5
\end{bmatrix}. \tag{19.37}$$

The forward and backward substitution approach applied to the pentadiagonal beam problem leads to the result shown in Fig. 19.2. The first two terms in (19.36) dominate the solution for large systems. The three remaining equations are as follows:

$$x_{1948} - 4x_{1949} + (6+\gamma)x_{1950} - 4x_{1951} + x_{1952} = f_{1950}; \tag{19.38}$$

$$x_{1949} - 4x_{1950} + (6+\gamma)x_{1951} - 4x_{1952} + x_{1953} = f_{1951}; \tag{19.39}$$

$$x_{1950} - 4x_{1951} + (6+\gamma)x_{1952} - 4x_{1953} + x_{1954} = f_{1952}. \tag{19.40}$$

Equations (19.38)–(19.40) lead to $f_{1951} = 0.2510$ and $f_{1950} = f_{1952} = -0.1257$ (Fig. 19.3). The error is $O(10^{-15})$. However, an approach is needed to remove f_{1950} and f_{1952}.

One such approach involves the use of adjacent solutions that have overlapping RHS terms. For example, consider two additional solutions, centered at $m = 1950$ and $m = 1952$ (these can be computed in parallel).

These solutions can be appropriately superimposed to obtain a new solution with $f_{1950} = f_{1952} = 0$ by taking advantage of the overlapping RHS terms. This leaves f_{1949}, f_{1951}, and f_{1953} as the only nonzero RHS terms. This strategy can be repeated,

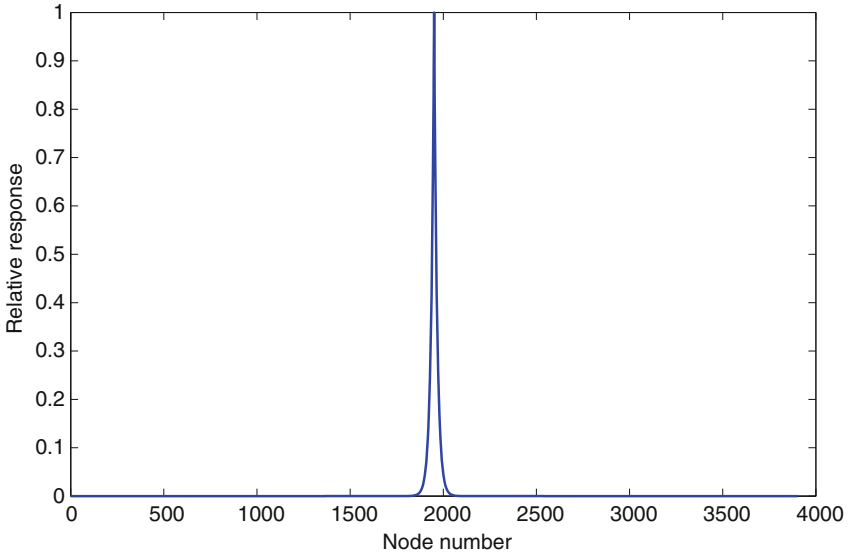


Fig. 19.2 Initial forward/backward substitution solution to the beam vibration problem. Note that the exponential behavior dominates and the oscillatory behavior is not evident

each time with three solutions, appropriately centered, and with the same spacing as the central solution, so that two RHS terms on either side overlap with those of the central solution. This would in effect double the separation between the central RHS term and that on each side for each iteration. The idea is to progressively move the extra RHS terms towards $m = 1$ and $m = n$, finally removing them. For example, the RHS terms for tenth iteration (Fig. 19.4) have a spacing of 1024 terms, and the solution (Fig. 19.5) is dominated by the oscillatory component. This was unexpected because the solution in Fig. 19.5 is the sum of evanescent responses (Fig. 19.2), just centered at different values of m.

The procedure to remove the RHS terms as they approach $m = 1$ and $m = n$ is doable but tedious, but fortunately there is a better way. The approach to subtract out the unwanted RHS terms can be generalized by computing n solutions in parallel (one centered at each value of m) and then performing a weighted sum so that all of the RHS terms sum to zero except for f_{1951}. There are three RHS terms for each individual solution (two for the solutions centered at $m = 1$ and $m = n$), so this leads to a tridiagonal system to compute the weights. The weights have an oscillatory nature (Fig. 19.6) resembling the solution itself.

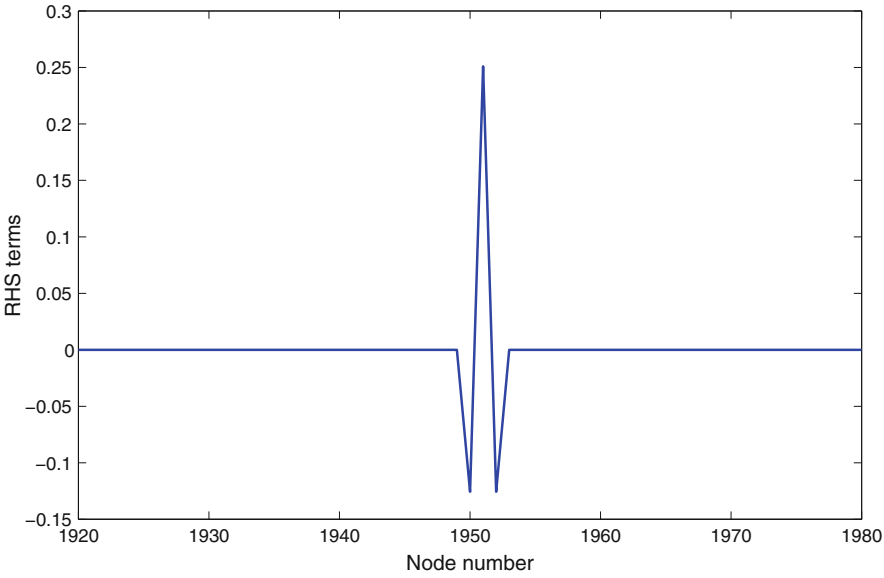


Fig. 19.3 The three source terms generated by the initial forward/back substitution solution

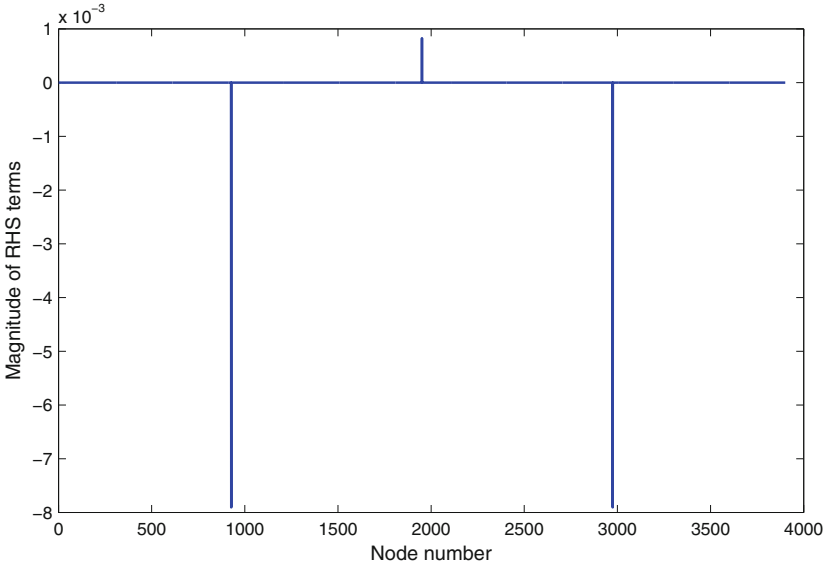


Fig. 19.4 RHS terms after the tenth iteration

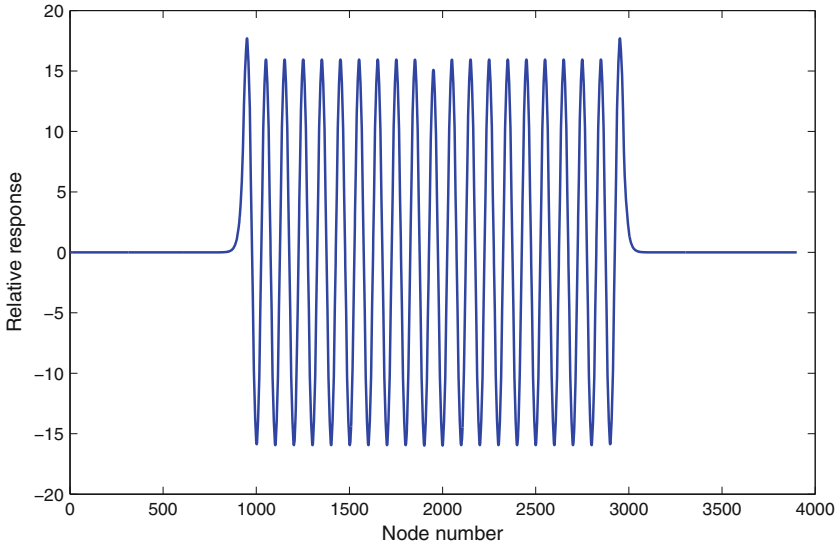


Fig. 19.5 Solution after the tenth iteration

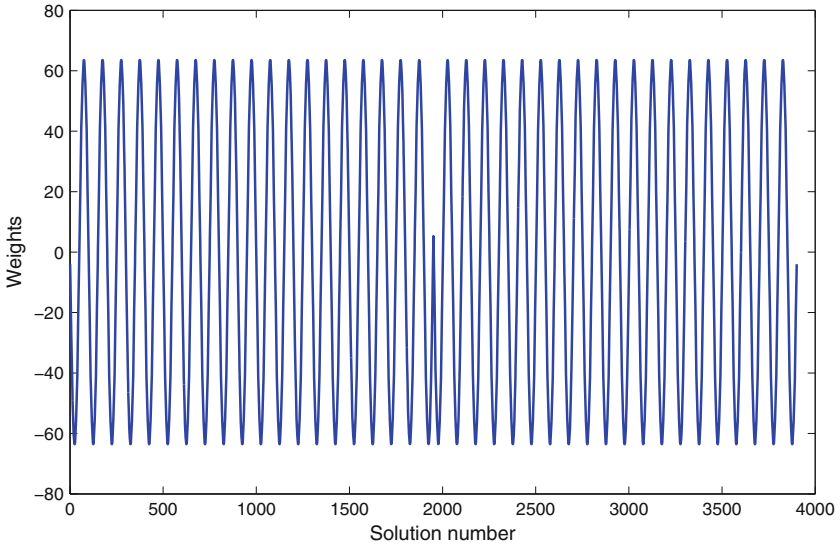


Fig. 19.6 Weights to obtain the required RHS vector for the beam vibration problem

Specifically, we compute $\hat{A}w = f$, where $w = [w_1, w_2, \ldots, w_n]^T \in \mathbb{R}^n$ and

$$\hat{A} = \begin{bmatrix} f_1^{(1)} & f_1^{(2)} & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ f_2^{(1)} & f_2^{(2)} & f_2^{(3)} & 0 & 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & f_3^{(2)} & f_3^{(3)} & f_3^{(4)} & 0 & 0 & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & f_4^{(3)} & f_4^{(4)} & f_4^{(5)} & 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & 0 & 0 & f_5^{(4)} & f_5^{(5)} & f_5^{(6)} & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 & f_{n-3}^{(n-4)} & f_{n-3}^{(n-3)} & f_{n-3}^{(n-2)} & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 & 0 & f_{n-2}^{(n-3)} & f_{n-2}^{(n-2)} & f_{n-2}^{(n-1)} & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 0 & f_{n-1}^{(n-2)} & f_{n-1}^{(n-1)} & f_{n-1}^{(n)} \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 0 & f_n^{(n-1)} & f_n^{(n)} \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{19.41}$$

where the $f^{(m)}$ terms represent the unconstrained RHS terms resulting from solving the problem via forward and backward substitution processes that meet at $k = m$. The solution to this system leads to the weights w that support a superposition process leading to any specified RHS vector f and the associated solution x, i.e.,

$$f_k = \sum_{m=1}^{n} w_m f_k^{(m)}, \tag{19.42}$$

$$x_k = \sum_{m=1}^{n} w_m x_k^{(m)}. \tag{19.43}$$

19.3.2.1 Modular Solution

We can also create a solution so that $x_k \neq 0$ for a limited range of values of k, and $x_k = 0$ otherwise. For example, introducing fictitious RHS terms at $k = 1751$ and $k = 2151$ in the beam vibration example can lead to $x_k \neq 0$ for $1753 \leq k \leq 2149$, and $x_k = 0$ for all other values of k. Methods can then be developed to remove the introduced RHS terms to solve the original posed problem. Consider the following equation:

$$x_{1749} - 4x_{1750} + (6+\gamma)x_{1751} - 4x_{1752} + x_{1753} = f_{1751}. \tag{19.44}$$

Introducing f_{1751} allows us to set $x_k = 0$ for $k < 1753$ by setting $x_{1753} = f_{1751}$. We can then continue with the forward substitution process with $-4x_{1753} + x_{1754} = 0$; $(6+\gamma)x_{1753} - 4x_{1754} + x_{1755} = 0$; $-4x_{1753} + (6+\gamma)x_{1754} - 4x_{1755} + x_{1756} = 0$, etc.

Likewise, we can set $x_{2149} = f_{2151}$, and by extension, set $x_k = 0$ for $k > 2149$ using the following equation:

$$x_{2149} - 4x_{2150} + (6+\gamma)x_{2151} - 4x_{2152} + x_{2153} = f_{2151}. \tag{19.45}$$

In the same way, we can perform the backward substitution process with $x_{2148} - 4x_{2149} = 0$; $x_{2147} - 4x_{2148} + (6+\gamma)x_{2149} = 0$; $x_{2146} - 4x_{2147} + (6+\gamma)x_{2148} - 4x_{2149} = 0$, etc.

Within the "module," forward substitution is then performed for $1753 \leq k \leq 1951$ and backward substitution is performed for $1951 \leq k \leq 2149$. These two solutions are then matched at $k = m = 1951$, which requires two independent solutions for both $1753 \leq k \leq m$ and $m \leq k \leq 2149$. For the second solution, we can use the evanescent solution shown in Fig. 19.2.

A set of four equations can then be developed to match the forward and backward solutions with four unknowns, i.e., ξ_1, ξ_2, η_1, and η_2, which govern the contributions of the forward and backward solution vectors u^f, v^f, and u^b and v^b. The equations are as follows:

$$\xi_1 u^f_{1951} + \xi_2 v^f_{1951} = \eta_1 u^b_{1951} + \eta_2 v^b_{1951}; \tag{19.46}$$

$$\begin{aligned} &\xi_1 u^f_{1948} + \xi_2 v^f_{1948} - 4(\xi_1 u^f_{1949} + \xi_2 v^f_{1949}) + (6+\gamma)(\xi_1 u^f_{1950} + \xi_2 v^f_{1950}) \\ &-4(\eta_1 u^b_{1951} + \eta_2 v^b_{1951}) + \eta_1 u^b_{1952} + \eta_2 v^b_{1952} = 0; \end{aligned} \tag{19.47}$$

$$\begin{aligned} &\xi_1 u^f_{1949} + \xi_2 v^f_{1949} - 4(\xi_1 u^f_{1950} + \xi_2 v^f_{1950}) + (6+\gamma)(\xi_1 u^f_{1951} + \xi_2 v^f_{1951}) \\ &-4(\eta_1 u^b_{1952} + \eta_2 v^b_{1952}) + \eta_1 u^b_{1953} + \eta_2 v^b_{1953} = f_{1951}; \end{aligned} \tag{19.48}$$

$$\begin{aligned} &\xi_1 u^f_{1950} + \xi_2 v^f_{1950} - 4(\xi_1 u^f_{1951} + \xi_2 v^f_{1951}) + (6+\gamma)(\eta_1 u^b_{1952} + \eta_2 v^b_{1952}) \\ &-4(\eta_1 u^b_{1953} + \eta_2 v^b_{1953}) + \eta_1 u^b_{1954} + \eta_2 v^b_{1953} = 0. \end{aligned} \tag{19.49}$$

This leads to the solution (Fig. 19.7) and the associated source terms (Fig. 19.8). The error is $O(10^{-12})$. We can thus perform the solution in sections (or "modules") in parallel and assemble them to obtain the overall solution. We can make each module sufficiently small to prevent the exponential behavior from dominating the solution. Finally, we can develop a system of equations to solve for the weights for each solution to remove the introduced RHS terms, similar to the procedure in the last section. The size of this system to determine the weights will be smaller than the original problem size by a factor equal to the module length.

19.3.3 Active Vibration Suppression

The modular solution leads to a novel approach for active vibration suppression. The introduced RHS terms in the modular solution (Fig. 19.8) can be interpreted

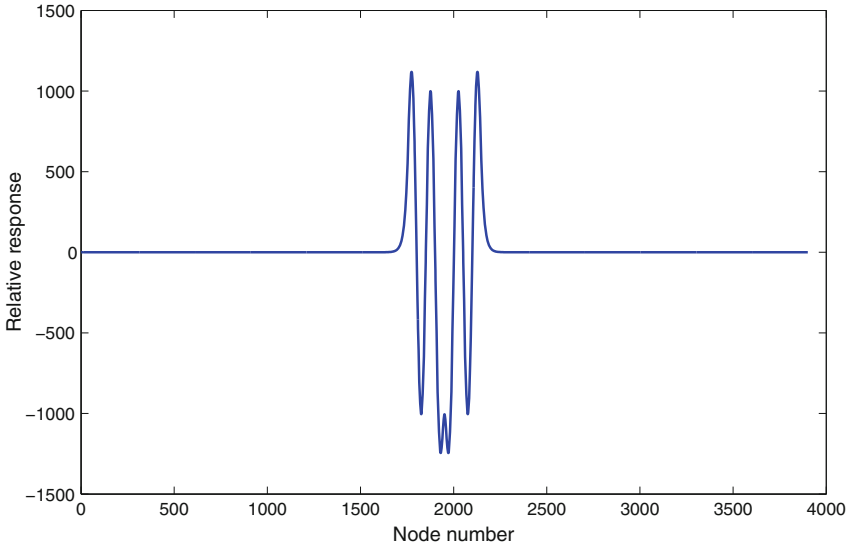


Fig. 19.7 Solution confined to $1751 \leq k \leq 2151$ for the beam vibration problem

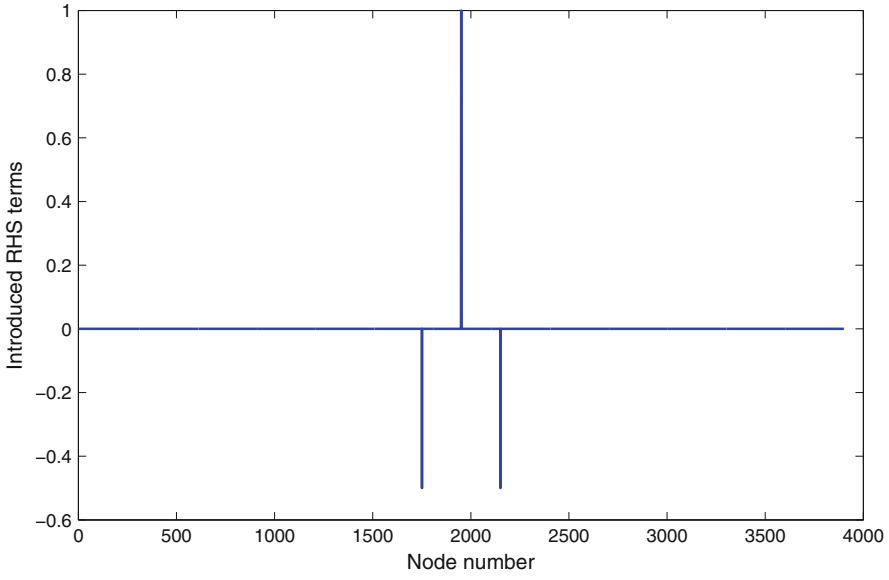


Fig. 19.8 Source terms for the confined solution

as applied harmonic forces that can confine the vibrational energy to a region of an arbitrary size (Fig. 19.7). Figure 19.9 shows the solution without the introduced RHS terms. Since the RHS source terms can be introduced anywhere in the system,

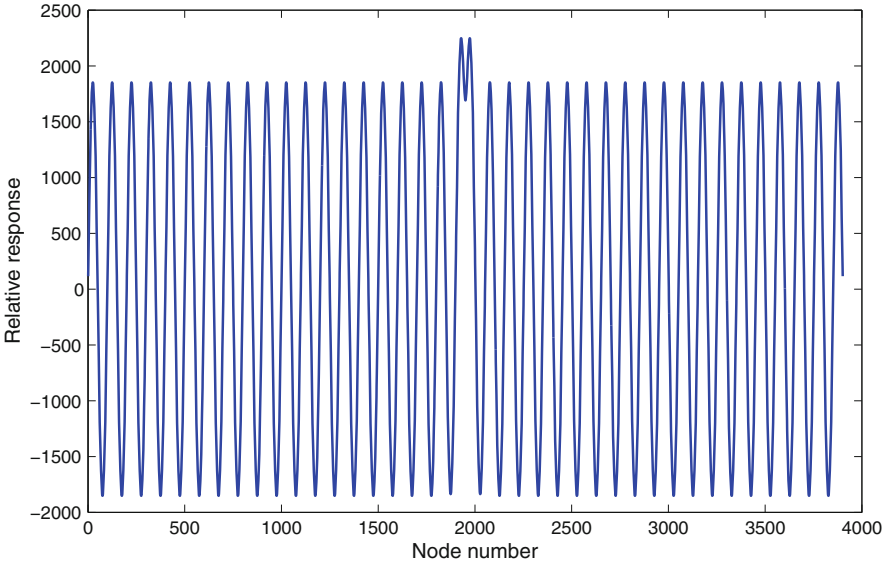


Fig. 19.9 Solution without the introduced RHS terms f_{1751} and f_{2151}

it follows that vibrational energy can be confined to any region on the beam. Outside of the applied forces, the oscillatory vibrational energy is exactly zero; only the evanescent energy remains.

In practice, sensors on the beam (e.g., accelerometers) can detect a single noise source (or multiple noise sources) that will be confined. Based on the input data for these noise sources, a model of the beam is run with RHS terms simulating those noise sources. Other sources that surround the noise sources are then introduced on the beam. The model will determine the magnitude and phase of the introduced sources that will isolate the remainder of the beam from the noise sources. Finally, actuators mounted on the beam at the locations corresponding to those in the model can be driven to the prescribed levels to cancel the noise.

In this example the noise sources all had the same phase, so all of the RHS terms were real. This will not be true in general, which will mean that the introduced RHS terms will be complex, meaning that each will have a computed amplitude and phase.

Note that each set of noise sources can be isolated individually or all of them can be isolated with just two introduced sources, as long as the vibrations are linear (so that superposition holds). Although just using two introduced sources is simpler, multiple sets of introduced sources may lead to better results when the noise sources that need to be suppressed are distributed over most of the beam.

19.4 Periodic Systems

Periodic systems are often employed as an idealized model for some physics-based systems, e.g., for wave propagation problems. For example, a one-dimensional Helmholtz problem with periodic boundary conditions is governed by the following equations:

$$\frac{d^2P}{ds^2}+\kappa^2 P = 0, \tag{19.50}$$

$$P(L) = P(0)e^{i\kappa L}; \frac{dP(L)}{ds} = \frac{dP(0)}{ds}e^{i\kappa L}. \tag{19.51}$$

This will lead to a tridiagonal system with off-diagonal terms as follows:

$$\begin{bmatrix}
e^{i\kappa L} & 0 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & 0 & -1 \\
e^{i\kappa L} & -e^{i\kappa L} & 0 & 0 & 0 & \ddots & \ddots & \ddots & \Delta s & -\Delta s \\
1 & (-2+\delta) & 1 & 0 & 0 & 0 & \ddots & \ddots & \ddots & \vdots \\
0 & 1 & (-2+\delta) & 1 & 0 & 0 & 0 & \ddots & \ddots & \vdots \\
\vdots & 0 & 1 & (-2+\delta) & 1 & 0 & 0 & 0 & \ddots & \vdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\vdots & \ddots & \ddots & \ddots & 1 & (-2+\delta) & 1 & 0 & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & 0 & 1 & (-2+\delta) & 1 & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \ddots & 0 & 1 & (-2+\delta) & 1 & 0 \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & -1 & (-2+\delta) & 1
\end{bmatrix}. \tag{19.52}$$

Here $\delta = \kappa^2 \Delta s^2$ (where κ is the wavenumber), Δs is the finite difference discretization length along the s coordinate, and L is the length of the problem domain. We can solve this system by assuming $x_1 = x_2 = 1$, which leads to the determination of x_{n-1} and x_n. We can then find the remaining unknowns via forward substitution, leaving the last two equations unsatisfied. Adding two unknowns in the same way that they are added in pentadiagonal systems allows these equations to be satisfied. So even though the diagonals have spacings of $n-1$ and $n-2$ from the three central diagonals, the procedure is similar to that for pentadiagonal systems.

19.5 Banded Systems

Jandron et al. [8] extended the algorithm to general banded systems. For systems that do not exhibit exponential behavior, the optimal computational expense scales with $2np+2nq+n+2q^3+4q^2$, where p is the lower bandwidth and q is the upper bandwidth when solved using q processors and where vectorization can be completed across n processors to have a negligible expense [8]. In contrast, sequential LU decomposition requires $2npq+2np+2nq$ calculations when $n \gg 1$. When $p = q$, the speedup scales with [8]

$$\frac{n\cdot(2q^2+4q)}{n\cdot(4q+1)+2q^3+4q^2}. \tag{19.53}$$

Figure 19.10 shows the speedup for different cases of processor availability. The optimal case, i.e., $q|n$-proc, refers to q-processors for the q asynchronous forward

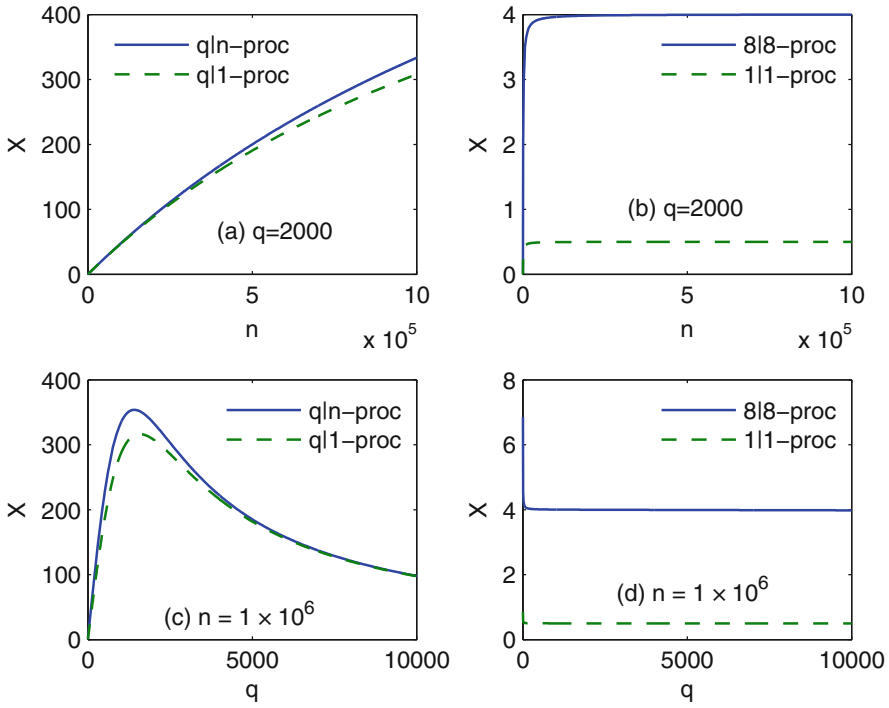


Fig. 19.10 Expected speedup (X) relative to sequential banded LU decomposition from [8]. (**a**) and (**c**) use optimal q-processors for the asynchronous forward substitution calls with n and 1 processor used, respectively, for the final vectorized superposition. Note that in (**c**) the limit as $q \to n$ is $X \to 1$ where the $q\times q$ system dominates the expense. (**b**) and (**d**) use non-optimal cases of 1 and 8 processors which could represent solver performance on a standard workstation

Table 19.1 Theoretical optimal speedups over sequential banded LU decomposition from [8]

n	q	$q/n\,(\%)$	Speedup
1×10^1	3	30	1.364
1×10^2	13	13	3.764
1×10^3	43	4.3	11.40
1×10^4	140	1.4	35.58
1×10^5	446	0.446	112.0
1×10^6	1413	0.1413	353.8
1×10^7	4470	0.0447	1118
1×10^8	14,140	0.0141	3536
1×10^9	44,720	0.0045	11,180

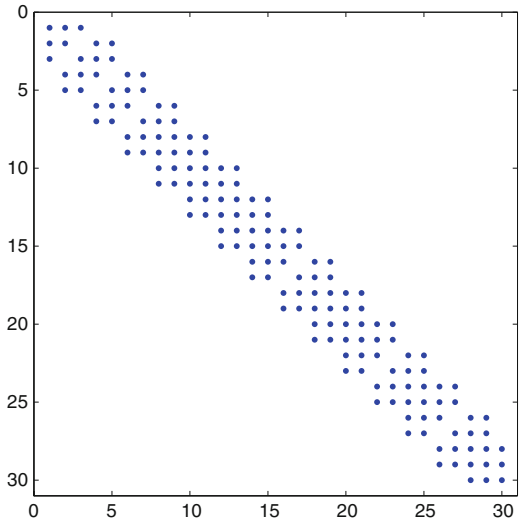


Fig. 19.11 The structure of the nonzero terms of the upper portion of the coefficient matrix resulting from a finite element analysis of the beam vibration problem

substitution steps and n-processors for the final vectorized superposition. The $q|1$-proc refers to the same except only one processor is used for the superposition step.

Table 19.1 shows the theoretical optimal speedups when $p = q$. In particular, a system with $n = 1\times10^9$ and $q = p = 44{,}720$ (i.e., a bandwidth of 89,441) shows a speedup of 11,180 times over the sequential banded LU decomposition algorithm. This assumes that the algorithm can asynchronously execute 44,720 threads with negligible latency. However, we do not yet have speedup results for systems exhibiting exponential behavior.

19.6 Block Banded Systems

Finite element approaches typically lead to block banded systems. Figure 19.11 shows the structure of the nonzero terms in the upper portion of the coefficient matrix for the beam vibration problem resulting from a finite element mesh.

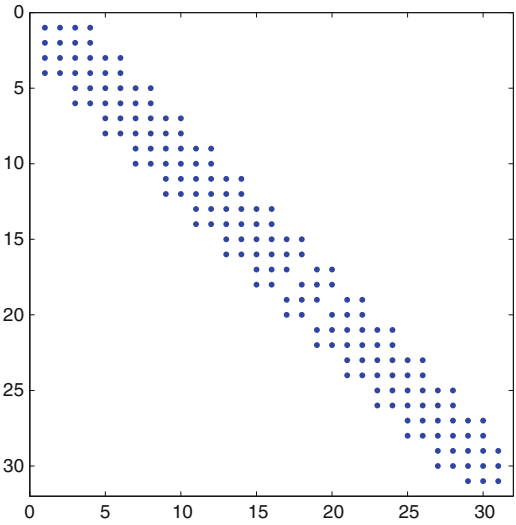


Fig. 19.12 The structure of the nonzero terms of the lower portion of the coefficient matrix resulting from a finite element analysis of the beam vibration problem

The forward substitution process can be performed on this problem, as it does not exhibit exponential growth behavior (this is a surprise: the finite difference version leads to significant exponential growth behavior). We assume $v_1^{(1)} = v_2^{(1)} = 1$ and otherwise follow the procedure outlined for tridiagonal or pentadiagonal systems, except that a series of 2×2 systems are solved to obtain the terms in the solution vector.

Figure 19.12 shows the lower portion of the coefficient matrix. After completing the forward substitution process by solving 2×2 systems, there is one remaining equation to compute x_n, with two equations not satisfied. We can add two column vectors just as in the pentadiagonal problem, or add unknowns that will lead to a 2×2 matrix (Table 19.2).

19.7 Towards a p-adic Parallel Linear Solver

The purpose of this section is to provide an introduction to the theory of p-adic matrix computation and related parallel p-adic algorithms in the form of a survey of existing methods and their evolution. The ultimate goal is to fully develop an exact p-adic parallel solver for very large banded linear systems, bypassing, in the process, the so-called Bellman's (1961) *Curse of Dimensionality* [18].

The idea is to use *p-adic arithmetic* which operates over the field of rational numbers, to perform exact computations. The p-adic representation of these rational numbers is given by the so-called *Hensel Codes* [19–23]. The section is self-contained in terms of notations. That is, for the sake of completeness and clarity,

Table 19.2 Nomenclature for Sects. 19.1–19.6

Symbol	Usage
A	Coefficient matrix
$\hat{A}$	Coefficient matrix used to compute weights
$\widetilde{A}$	Augmented matrix
$A^{(1)}$	Upper part of the tridiagonal coefficient matrix
$A^{(2)}$	Lower part of the tridiagonal coefficient matrix
a_i	Diagonal term
α	Added unknown
$\alpha^{(1)}$	Added unknown
$\alpha^{(2)}$	Added unknown
$\alpha^{(3)}$	Added unknown
α_1	Scale factor
α_2	Scale factor
α_3	Scale factor
$\widetilde{b}$	RHS vector
b_i	Diagonal term
B	Coefficient matrix
c_i	Diagonal term
C_1	Arbitrary constant
C_2	Arbitrary constant
C_3	Arbitrary constant
C_4	Arbitrary constant
χ	Cross-sectional area
d	Tridiagonal RHS vector
d_i	Diagonal term
δ	$\kappa^2 \Delta s^2$
e_i	Diagonal term
$\bar{e}$	Added column vector
$\widetilde{e}$	Added column vector
E	Young's modulus
η_1	Unknown governing the contribution of u^b
η_2	Unknown governing the contribution of v^b
f	Pentadiagonal RHS vector
γ	$-\omega^2 \rho \chi \Delta s^4/(EI)$
k	Index associated with the forward/backward substitution process
κ	Wavenumber
λ	Wavelength
h	Discretization length
i	$\sqrt{-1}$
I	Moment of inertia
L	Periodic length
m	Index of the single nonzero RHS term

(continued)

Table 19.2 (continued)

Symbol	Usage
n	Number of unknowns
Φ	Continuous dependent variable
ω	Angular frequency
P	Continuous dependent variable in Helmholtz equation
p	Number of subdiagonals
q	Number of superdiagonals
r	Number or nonzero RHS terms
ρ	Density
s	Continuous independent spatial variable
σ	Roots of the characteristic polynomial
t	Continuous independent temporal variable
u	Solution vector
u^f	Forward solution vector
u^b	Backward solution vector
v	Solution vector
v^f	Forward solution vector
v^b	Backward solution vector
$v^{(1)}$	Solution vector
$v^{(2)}$	Solution vector
$v^{(3)}$	Solution vector
W	Applied load
w	Weights used for superposition
x	Solution vector
X	Expected speedup
y	Solution vector
Y	Transverse displacement in the beam vibration problem
ξ_1	Unknown governing the contribution of u^f
ξ_2	Unknown governing the contribution of v^f
z	Continuous independent variable
ζ	Roots of the characteristic polynomial

most of the nomenclature is specific to the topic, and not carrying any additional meaning from previous sections.

19.7.1 p-adic Numbers

p-adic numbers, ultrametric spaces, non-Archimedean numbers, and isosceles spaces all express the same idea [20, 24]. Kurt Hensel, who initiated the p-adic analysis in 1898, considered number as analytic functions on some affine scheme

Spec($\mathbb{Z}$) whose points are the prime ideals $p\mathbb{Z}$, for $p = 0$ or p a prime, acting as a "local coordinate," so that a number n is "locally" represented as a unique power series expansion

$$\mathbb{Z} \ni n = \sum_{k=0}^{\infty} n_k p^k \tag{19.54}$$

finite for natural numbers, and with coefficients $n_k \in \{0, 1, 2, \ldots, p-1\}$. The length of the common initial sequence defined the p-adic metric, i.e.,

$$|n-m|_p = p^{-d}, \tag{19.55}$$

when $m = n_0+\cdots+n_{d-1}p^{d-1}+m_d p^d+\cdots$ and $m_d \neq n_d$. Such a metric is an *ultrametric* in the sense it satisfies the strong inequality

$$|x+y|_p \leq max(|x|_p, |y|_p) \tag{19.56}$$

with equality for $|x|_p \neq |y|_p$. That is, all triangles are isosceles under such a metric. Equivalently, the underlying valuation is given by

$$|0|_p = 0; \quad |n|_p = p^{-k} \tag{19.57}$$

from the factorization $n = n'p^k$ with n' and p relatively prime. And for a rational number in the *normalized form* $r = \frac{a}{b}p^k$, $b \neq 0$ with $gcd(a, b) = 1 = gcd(a, p) = gcd(b, p)$, we have $|r|_p = p^{-k}$. This actually defined the *p-adic* norm on $\mathbb{Q}$, with the associated metric $\rho(x, y) = |x-y|_p$.

Moreover the allowed infinite expansion leads to completion with respect to the ultrametric with the completed space denoted $\mathbb{Z}_p$, the space of *p-adic integers* with the usual integers set $\mathbb{Z}$ as a dense subset, and with no zero divisors, p being prime. The field of fractions which contains densely the usual set $\mathbb{Q}$ of rational numbers is denoted $\mathbb{Q}_p$ the space of p-adic numbers described as

$$\mathbb{Q}_p = \{r = \sum_{k=-N}^{\infty} r_k p^k | \quad r_k \in \{0, 1, \ldots, p-1\}\} \tag{19.58}$$

with $\mathbb{Z}_p = \{x \in \mathbb{Q}_p; |x|_p \leq 1\}$ the compact p-adic unit disk. $\mathbb{Q}_p$ is locally compact, totally disconnected, allowing calculus to be performed but no "reasonable" analyticity is expected. The real numbers line gives a natural geometric ordering for the usual Euclidean metric, whereas a *hierarchical tree* offers a natural ordering in the ultrametric case. The completion of the algebraic closure $\mathbb{Q}_p^a$ of $\mathbb{Q}_p$ is the space $\mathbb{C}_p$ of p-adic complex numbers, also totally disconnected, and not even locally compact, and its unit disk and projective line are not compact. The value groups are, respectively, $|\mathbb{Q}_p^*| = p^{\mathbb{Z}}$ and $|\mathbb{C}_p^*| = p^{\mathbb{Q}}$. A p-adic number $x \in \mathbb{Q}_p^a$ is said

to be *ramified* if $|x|_p \notin p^{\mathbb{Z}}$, and *unramified* otherwise. Notably the above properties suggest that any "reasonable" analytic work is better done in some larger space such as Berkovich spaces; see [25, 26].

Given an infinite p-adic expansion of $x = x_0 + \cdots + x_{n-1}p^{n-1} +$higher order terms, or written as

$$x = x_0 + \cdots + x_{n-1}p^{n-1} + O(p^n),$$

its p-adic approximation is given by $\bar{x} = x_0 + \cdots + x_{n-1}p^{n-1}$, i.e., $x \equiv \bar{x} \bmod p^n$. Indeed we have

$$|x - \bar{x}| \leq p^{-n}$$

ensuring convergence. n is the *order* or *absolute precision* of $\bar{x}$. The *relative precision* is given by $n - min\{i \in \mathbb{Z}, a_i \neq 0\}$. The p-adic precision arithmetic goes as follows:

1. $(a + O(p^{k_1})) + (b + O(p^{k_1})) = a + b + O(p^{min(k_1,k_2)})$. That is, *p-adic errors do not add*, a great advantage over real precision with roundoff errors accumulating.
2. $(a + O(p^{k_1})) \times (b + O(p^{k_1})) = a \times b + O(p^{min(k_1+v_p(b),k_2+v_p(a))})$, where $v_p(.)$ indicates the corresponding p-adic valuation. That is, $v_p(x) := \max\{r \in \mathbb{Z} : p^r | x\}$.

Note also that $\mathbb{Q}_p$ carries a measure, the *Haar measure dx*, normalized to give a volume of 1 to the unit disk $\mathbb{Z}_p$.

p-adic matrix computation involves the so-called *Hensel codes* as defined by Krishnamurty, originated from the following Hensel Lemma: [21, 22, 27, 28].

Lemma 1 (Basic Hensel's Lifting Lemma). *Let $f(x) \in \mathbb{Z}_p[x]$ be an n-tuple polynomial in the variables $x = (x_1, \ldots, x_n)$ with coefficients in $\mathbb{Z}_p$. Let $a \in \mathbb{Z}_p^n$ such that* (1) *$f(a) \equiv 0$ mod p and* (2) *$det(\frac{\partial f}{\partial x}(a)) \not\equiv 0$ mod p. (Invertible Jacobian) Then there exists a unique root b of f "near a," i.e., $f(b) = 0$ and $b \equiv a$ mod p.*

Note the similarity with Newton's Method in Real Analysis.

Example 1. Take $f(x) = x^2 - 7$ and $p = 3$. We get $f(1) = -6 \equiv 0 \bmod 3$, and $f'(1) = 2 \not\equiv 0 \bmod 3$. So, by Hensel Lemma, there exists a unique 3-adic integer n such that $n^2 - 7 = 0$, and $n \equiv 1 \bmod 3$. There are many approximations to n such as

$$\begin{aligned} n &\equiv 1 + 3 \quad mod \quad 3^2 \\ &\equiv 1 + 3 + 3^2 \quad mod \quad 3^3 \\ &\equiv 1 + 3 + 3^2 + 2.3^4 \quad mod \quad 3^5 \end{aligned} \tag{19.59}$$

That is, a root 1 of $f(x)$ mod 3 is lifted to a root in $\mathbb{Z}_3$.

Note, however, that not all solutions modulo p lift to p-adic solutions, For instance, the equation $x^2 - 5$ converts to $x^2 + 1 = 0$ modulo 2, and this has the

solution $x = 1$ modulo 2. But due to the fact that any 2-adic unit (i.e., $u \equiv 1 \bmod 2$) is a square iff $u \equiv 1 \bmod 8$, (see [29, 30]), the equation has no solution in $\mathbb{Z}_2$.

19.7.1.1 Some Examples

1. $199 = 1244$ *in 5-ary (base 5)* but $199 = 0.4421$ *in 5-adic*.
2. For $p = 5$, we have
   (a) $13.41 = 1\times5^{-2}+3\times5^{-1}+4\times5^{0}+1\times5^{1} = \frac{241}{25}$
   (b) $0.1341 = 1\times5^{0}+3\times5^{1}+4\times5^{2}+1\times5^{3} = 241$
   (c) $0.01341 = 0\times5^{0}+1\times5^{1}+3\times5^{2}+4\times5^{3}+1\times5^{4} = 1205$
3. $x = \sum_0^\infty p^n = \frac{1}{1-p} = \frac{-1}{p-1}$. Consequently $-1 = (p-1)\sum_0^\infty p^n$.
   For instance, $-1 = (3-1)\sum_{i=0}^\infty 3^i \cdots 2222$ in $\mathbb{Z}_3$.
4. For $x = \sum_{i=n}^\infty x_i p^i$, we have $-x = \sum_{i=n}^\infty y_i p^i$, where $y_n = p-x_n$, $y_i = (p-1)-x_i$, for $i > n$.
   For instance: $\frac{1}{3} = 0.2313131\cdots$ and $-\frac{1}{3} = 0.3131313\cdots$

19.7.2 p-adic Solver Preliminary

Matrix computations include solving a linear system, finding the inverse or generalized inverse of a matrix, reducing a matrix to a specific canonical form (e.g., triangular), and determining the characteristic equation [19, 22, 31]. In the conventional *n-ary* or *floating-point arithmetic* roundoff errors are allowed to accumulate rending the results oftentimes totally unreliable. For rational matrices one could use either of the so-called *Rational or Residue Arithmetics.* The former is very expensive and laborious with the rational add/subtract and multiplication/division operations with reduction to lowest form. (See Knuth.) In the latter, also called *Modular Arithmetic,* a rational $r = \frac{a}{b}$ with $b \neq 0$, $0 \leq a, b \leq p-1$, is uniquely written as $r = a\times b^{-1}$ modulo p, and the multiplicative inverses in the Galois finite field $\mathbb{F}_p = \{0, 1, 2, \ldots, p-1\}$ under the binary operations "addition and multiplication modulo p." Here computational complexity may increases proportionally to p, e.g., multiplication complexity of order $O(\frac{N^2+N}{2})$ for p with N binary digits (its precision). Reduction of the complexity by using multiple prime moduli was suggested; see Young and Gregory, Rao et al. in [21, 23, 28, 31].

For an integer linear system, using the *Chinese Remainder Theorem* (CRT) in the modular approach allows to avoid the growth of the coefficients. Combining the p-adic method with linear lifting allows to avoid both the growth of the coefficients as well as the growth of the number of modules. It has been known that the p-adic method initiated by Dixon in the 1980s is the best method both theoretically and practically, in particular for large-scale systems, in part because the procedure to reconstruct a quotient is more complicated than that of the numerator and denominator separately.

p-adic methods have been considered many times to obtain the exact rational solution to a nonsingular system $Ax = d$ of linear equations, mostly with integer coefficients; see Krishnamurty et al, Dixon, Sjogren et al. For example, exact rational solutions are sought out for systems so ill-conditioned that the usual floating-point calculations are inadequate. Exact solutions could be obtained by direct methods as well as by congruence techniques.

Gaussian elimination, as the premier direct method, combined with multiple-precision arithmetic is often used to find exact solutions to a system of integer or rational linear equations. However, generating truncation error is the major drawback of any algorithm based on the floating-point system, making computational results for large systems unreliable and unacceptable.

It is therefore necessary to design an *error-free* rational computation. Over the years residue and p-adic-based efficient sequential algorithms and software for solving linear equations have been designed and implemented, along with, recently related parallel algorithms for exact solution.

The fact that computing each modulus can be done separately makes multiple moduli congruence and p-adic expansion algorithms more adequate for parallelism. This step is completely parallel and no communication is required among the processors. The single-radix or mixed-radix conversion algorithms, both parallelizable to some degree, are then used to combine the solutions for each of the moduli.

A brief comparison of direct methods utilizing the multi-precision arithmetic versus the p-adic methods is given in [28, 30, 32], where the execution times with respect to the upper bound on the matrix entries are also analyzed.

19.7.3 Hensel Code Arithmetic

Consider the infinite p-adic expansion of a rational number

$$x = \sum_{k=-m}^{\infty} x_k p^k$$

represented symbolically as

$$x = x_{-m} \ldots x_{-1} \bullet x_0 x_1 \ldots x_n \ldots$$

The p-adic approximation $\bar{x}$ is given by

$$\bar{x} = \sum_{k=-m}^{n} x_k p^k = x_{-m} \ldots x_{-1} \bullet x_0 x_1 \ldots x_n$$

This finite sequence defines the so-called *Hensel code* of x denoted $H(p, r, x)$ where $r = m+n+1$ is the number of digits with the radix point as in the expansion. In the

mantissa-exponent form, we write $H(p, r, x) = (m_x, e_x)$. Using the rational x in the form

$$x = \frac{a}{b} = \frac{c}{d}p^m, \quad gcd(c, d) = 1 = gcd(c, p) = gcd(d, p),$$

$m_x := mantissa = cd^{-1} \bmod p^r$ and $e_x := exponent = -m$.

For example, take $x = \frac{7}{15}$, that is, $x = 7\times 3^{-1}\times 5^{-1}$. For $p = 5,\ r = 4,\ H(5, 4, x)$ is obtained as follows, modulo $5^4 = 625$

$$\begin{aligned} m_x &= 7\times 3^{-1} \\ &= 7\times 417 \quad (417 = 3^{-1} \mod 625) \\ &= 419 = 0.4313 \\ e_x &= -1 \end{aligned}$$

Importantly we have: let $N = \lfloor \sqrt{\frac{p^r-1}{2}} \rfloor$. Then every rational number $x = \frac{a}{b}, b \neq 0$ such that $0 \leq |a|, b \leq N$ has a unique Hensel code $H(p, r, x)$. The Hensel codes are closed with respect to the basic arithmetic operations (add/subtract/multiply/divide) within the range condition. See more details in Krishnamurty, including the basic arithmetic operations. The following conversion process codes any rational number $x = \frac{a}{b} = \frac{c}{d}p^m$ into an infinite expansion; see Sjogren et al. in [32]

Conversion Process

Step 1: x mod p $= x_0$

Step 2: $x = (x-1)/p$, *go to Step 1 to get* x_1

Continue Step 1 and Step 2 to get x_i

Finally, $x = p^n \sum_{i=0}^{\infty} x_i p^i = \sum_{i=n}^{\infty} x_{i-n} p^i$.

19.7.4 Single Modulus: Dixon's Algorithm and Its Parallelization

Given a linear system

$$Ax = d \tag{19.60}$$

with an integer matrix $A \in \mathbb{Z}^{n\times n}$ nonsingular modulo p prime, an integer vector $d \in \mathbb{Z}^{n\times 1}$, computation of the exact rational solution $x \in \mathbb{Q}^{n\times 1}$ using Dixon's Algorithm consists in three main steps: see [23, 27, 33, 34].

1. *Inversion* Step : Compute the inverse of the matrix A modulo p.
2. *Iteration* Step: Obtain the solution $\bar{x}$ of $A\bar{x} = b \bmod p^m$ is by iteration.

3. *Euclidean* Step: Recover the rational solution using the Extended Euclidean Algorithm.

From the p-adic expansion $x = \sum_{i=1}^{\infty}$, we write $x = \bar{x} + O(p^m)$, where $\bar{x} = \sum_{i=1}^{m-1}$, and m the so-called *absolute precision* of x. Algorithmically these steps break down as: (See [33].)

Dixon's Algorithm (A; b; p;m)
Step 1 (Inversion): $C = A^{-1}$ *mod p ;*
Step 2 (Iteration): $b_0 = b$ *;*
for $i = 1$ *to m*
$x_i = Cb_i$ *mod p*
$b_{i+1} = p^{-1}(b_i - Axi)$
end
$\bar{x} = \sum_{i=1}^{m-1} x_i p^i$
Step 3 (Euclidean): for $j = 1$ *to n ;*
$u_{-1}(j) = p^m,\ u_0(j) = \bar{x}(j)$
$v_{-1}(j) = 0,\ v_0(j) = 1$
while $u_i(j) < p^{m/2}$
$q_i(j) = \lfloor u_{i-1}(j)/u_i(j) \rfloor$
$u_{i+1}(j) = u_{i-1}(j) - q_i(j)u_i(j)$
$v_{i+1}(j) = v_{i-1}(j) + q_i(j)v_i(j)$
end
end
$x(j) = ((-1)^{-i}u_i(j)/v_i(j)); 1 \leq j \leq n$; *(Rational solution)*

19.7.5 Parallel Chinese Algorithm

19.7.5.1 Chinese Remainder Theorem

Theorem 1. *(CRT) Let* $r_1, r_2, \ldots, r_s$ *be a sequence of residues of an integer n with respect to the moduli* $p_1, p_2, \ldots, p_s$ *where* $gdc(p_i, p_j) = 1$ *for* $i \neq j$. *Define* $p = \prod_{i=1}^{s} p^i$, *and* $\tilde{p}_i$ *by* $\frac{p}{p_i}\tilde{p}_i \equiv 1$ *mod* p_i. *Then n is given by*

$$n \equiv \sum_{i=1}^{s} \frac{p}{p_i}\tilde{p}_i r_i \quad mod \quad p$$

To wit we use the following example from Sjogren et al.:

$$n \equiv 2 \quad mod \quad 3$$

$$n \equiv 3 \quad mod \quad 4$$

$$n \equiv 4 \quad mod \quad 5$$

Then $p = 60$, and $\tilde{p}_1 = 2,\ \tilde{p}_2 = 3,\ \tilde{p}_3 = 3$ leading to $n \equiv 59(\text{mod } 60)$. If $|n| < \frac{1}{2}p$ is desired, then $n = -1$.

One notes then that the CRT convert large integers into sequence of small integers. An extended version, denoted E-CRT, as given in Kornerup et al. convert a fractional number with large numerator and/or denominator into a sequence of small integers as well [30, 32, 34].

19.7.5.2 Decoding Algorithm

We present an algorithm proposed in Sjogren et al. to decode from the E-CRT, based also on Dixon's work, in which the following is proved: For a rational $\frac{a}{b}$ with $gcd(a, b) = 1$ and $\delta = \max(a, b)$. If $\delta \leq \lambda\sqrt{p}$, with $\lambda^2+\lambda-1 = 0$, that is, $\lambda = 0.618\cdots$ the decoding algorithm below gets the rational back.

Decoding Algorithm

Step 1: Chinese remainder theorem
$p = \prod_{i=1}^{s} p^i$
For $i = 1$ *to* s
Using extended Euclidean Algorithm to find
$\tilde{p}$ *by* $\frac{p}{p_i}\tilde{p}_i \equiv 1$ *(mod* p_i*.)*
End
$\bar{x} = n \equiv \sum_{i=1}^{s} \frac{p}{p_i}\tilde{p}_i r_i \ mod\ p$
Step 2: Euclidean Algorithm
$u_1 = p,\ u_0 = \bar{x}$
$v_1 = 0,\ v_0 = 1$
$i = -1$
While $u_i < \sqrt{p}$
$q_i = \lfloor u_{i-1}/u_i \rfloor$
$u_{i+1} = u_{i-1} - q_i u_i$
$v_{i+1} = v_{i-1} + q_i v_i$
i++
End
Rational solution
$x = ((-1)^i u_i / v_i)$

19.7.6 Multiple p-adic Arithmetic

The combination of the Extended Chinese Remainder Theorem (E-CRT) with p-adic arithmetic results in an algorithm called the *Multiple p-adic Algorithm (MPAA)* in Morrison [30].

19.7.6.1 Application of the MPAA

Consider the rational matrix

$$A = \begin{bmatrix} 1 & 2 \\ \frac{1}{3} & \frac{1}{4} \\ 5 & 6 \end{bmatrix}$$

and compute the generalized inverse A^g of A on a 64-bit CPU architecture, with a p-adic length $r = 2$. Choose a prime number set of the largest primes smaller than $\sqrt{2^{64}}$ to get

$$p_1 = 2147483647; \quad p_2 = 2147483629; \quad p_3 = 2147483587.$$

Upon parallel calculation of each p_i under p-adic arithmetic and decoding from the E-CRT, we obtain

$$A^g = \begin{bmatrix} -\frac{3}{5} & \frac{24}{5} & 0 \\ \frac{4}{5} & -\frac{12}{5} & 0 \end{bmatrix}$$

That is,

$$AA^g = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Remark 1. Parallel computation is a natural medium to perform the *Multiple p-adic Arithmetic* by the mere structure of the Extended Chinese Theorem.

To ensure overflow protection, it has been recommended to choose the prime numbers p_i such that $p_i \leq 46337$ for a 32-bit CPU architecture and $p_i \leq 2147483647$ on a 64-bit architecture [31–33].

19.8 Conclusion

We have introduced a solver for banded linear systems that involves adding a number of unknowns p equal to the number of superdiagonals. This allows us to compute solution vectors in parallel via a forward substitution process without any communication between processors. For the ideal case (i.e., $q = p$ processors), the theoretical speedup can exceed a factor of 10,000 when $n = 10^9$ for well-behaved systems. This represents the upper limit of performance. More realistic systems are not always well-behaved, in the sense that they exhibit exponential growth behavior as a result of the forward substitution process so that the magnitude of individual

terms can reach 10^{300} for $n = O(10^3)$. This prevents the approach from working, but there are several ways to address it, e.g., a forward/backward substitution approach for tridiagonal systems, and a modular solution approach. We have also introduced an approach based on p-adic analysis to be applied to Banded Linear Systems with the goal to develop a related p-adic parallel solver with Matlab and/or SAGE implementation, taking advantage of its potential for "Error-free Computation."

References

1. Demmel, J.W.: Applied Numerical Linear Algebra. SIAM, Philadelphia (1997)
2. Golub, G., Van Loan, C.: Matrix Computations, 4th edn. John Hopkins University Press, Baltimore (2013)
3. Demmel, J.W., Gilbert, J.R., Li, X.S.: An asynchronous parallel supernodal algorithm for sparse Gaussian elimination. SIAM. J. Matrix Anal. Appl. **20**, 915–952 (1999)
4. Donfack, S., Dongarra, J., Faverge, M., Gates, M., Kurzak, J., Luszczek, P., Yamazaki, I.: A survey of recent developments in parallel implementations of Gaussian elimination. Concurr. Comput. Pract. Exp. (2014). doi:10.1002/cpe.3306
5. Stone, H.S.: An efficient parallel algorithm for the solution of a tridiagonal linear system of equations. J. ACM **20**, 27–38 (1973)
6. Van der Vorst, H.A.: Analysis of a parallel solution method for tridiagonal linear systems. Parallel Comput. **5**, 303–311 (1987)
7. Ruffa, A.A.: A solution approach for lower Hessenberg linear systems. ISRN Appl. Math. **2011**, 236727 (2011)
8. Jandron, M.A., Ruffa, A.A., Baglama, J.: An asynchronous direct solver for banded linear systems (under review)
9. Luisier, M., Schenk, O., et al.: Fast methods for computing selected elements of the Green's function in massively parallel nanoelectronic device simulations. In: Wolf, F., Mohr, B., an Mey, D. (eds.) Euro-Par 2013. Lecture Notes in Computer Science, vol. 8097, pp. 533–544. Springer, Berlin/Heidelberg (2013)
10. Schenk, O., Bollhoefer, M., Roemer, R.: On large-scale diagonalization techniques for the Anderson model of localization. SIAM Rev. **50**, 91–112 (2008)
11. Schenk, O., Waechter, A., Hagemann, M.: Matching-based preprocessing algorithms to the solution of saddle-point problems in large-scale nonconvex interior-point optimization. J. Comput. Optim. Appl. **36**, 321–341 (2007)
12. Levinson, N.: The Wiener RMS (root mean square) error criterion in filter design and prediction. In: Wiener, N. (ed.) Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications, Appendix B, pp. 129–148. Wiley, New York (1949)
13. Trench, W.F.: An algorithm for the inversion of finite Toeplitz matrices. J. Soc. Ind. Appl. Math. **12**, 515 (1964)
14. Zohar, S.: The solution of a Toeplitz set of linear equations. J. ACM **21**, 272 (1974)
15. Gavel, D.T.: Solution to the problem of instability in banded Toeplitz solvers. IEEE Trans. Signal Process. **40**, 464 (1992)
16. MacLeod, A.J.: Instability in the solution of banded Toeplitz systems. IEEE Trans. Acoust. Speech Signal Process. **37**, 1449 (1989)
17. Meek, D.S.: The inverses of Toeplitz band matrices. Linear Algebra Appl. **49**, 117 (1983)

18. Bellman, R.E.: Dynamic Programming. Princeton University Press, Princeton (1957)
19. Gregory, R.T.: Error-Free Computation. Krieger, Huntington (1980)
20. Gregory, R.T., Krishnamurty, E.V.: Methods and Applications of Error-Free Computation. Springer, Berlin (1984)
21. Krishnamurty, E.V., Rao, T.M., Subramanian, K.: P-adic arithmetic procedures for exact matrix computations. Proc. Indian Acad. Sci. **82A**(5), 165–175 (1975)
22. Krishnamurty, E.V.: Matrix processors using p-adic arithmetic for exact linear computations. IEEE Trans. Comput. **26**(7), 633–639 (1977)
23. Limongelli, C.: On an efficient algorithm for big rational number computations by parallel p-adics. J. Symb. Comput. **15**(2) (1993)
24. Katok, S.: P-adic Analysis Compared with Real. AMS Student Math Library, vol. 37. American Mathematical Society, Providence (2007)
25. Berkovich, V.: Spectral Theory and Analytic Geometry over Non-achimedean Fields. Mathematical Surveys and Monographs, vol. 33. American Mathematical Society, Providence (1990)
26. Silverman, J.: The Arithmetic of Dynamical Systems. Springer, New York (2007)
27. Dixon, J.D.: Exact solution of linear equations using p-adic expansion. Numer. Math. **40**, 137–141 (1982)
28. Villard, G.: Parallel general solution of rational linear systems using p-adic expansions. In: Barton, M., Cosnard, M., Vanneschi, M. (eds.) Proceedings of the IFIP WG 10.3 Working Conference on Parallel Processing. Elsevier, Pisa (1988)
29. Gouvêa, F.Q.: P-Adic Numbers. An Introduction. Universitext, 2nd edn. Springer, Heidelberg (2003)
30. Morrison, J.: Parallel p-adic computation. Inf. Process. Lett. **28**(3) (1988)
31. Young, D.M., Gregory, R.T.: A Survey of Numerical Mathematics. Addison Wesley, Readin (1973)
32. Li, X., Lu, C., Sjogren, J.A.: Parallel implementation of exact matrix computation using multiple p-adic arithmetic. Int. J. Netw. Distrib. Comput. **1**(3), 124–133 (2013)
33. Koc, C.K.: Parallel p-adic method for solving linear systems of equations. Parallel Comput. **23**(13) (1997)
34. Kornerup, P., Gregory, R.T.: Mapping integers and Hensel codes onto Farey fractions. BIT **23**, 9–23 (1983)

Chapter 20
Genetic Vulnerability and Crop Loss: The Case for Research on Underutilized and Alternative Crops

Laban K. Rutto, Vitalis W. Temu, and Myong-Sook Ansari

Abstract The confluence of global climate change and population growth has brought to greater focus the question of how to satisfy future demand for food and fiber necessary to sustain current standards of living. Inevitably, agriculture will be called upon to do more at a time when established crops and cropping systems must confront new environmental and socio-economic challenges. Current efforts to preserve and characterize crop wild relatives and other genetic resources that could help crops meet future biotic and abiotic challenges are a direct response to the question. This chapter not only reiterates the importance of in situ and ex situ genetic conservation, it draws attention to the urgent need for investment in research on underutilized and alternative crops. The urgency relates directly to the fact that most of these crops are found in global biodiversity hotspots that are currently undergoing rapid environmental and socio-economic change.

Keywords Plant breeding and selection • Genetic vulnerability • Underutilized species • Crop wild relatives

20.1 Introduction

The concentration of modern agriculture on a few high yielding and universally accepted crops has resulted in two distinct phenomena:

(i) Hybrid crop varieties with narrow genetic bases relative to their wild relatives.
(ii) Declining food diversity among different cultures and communities as indigenous foods and practices are abandoned in favor of alternatives from (i) above.

Potential for worsening genetic vulnerability associated with a narrowing of the genetic base of economically important crops has been acknowledged and

L.K. Rutto (✉) • V.W. Temu • M.-S. Ansari
Agricultural Research Station, College of Agriculture, Virginia State University, Petersburg, VA 23806-0001, USA
e-mail: lrutto@vsu.edu; vtemu@vsu.edu; sookansari@yahoo.ca

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_20

gene banks in developed and middle income countries are working to broaden the diversity of genetic material in their repositories. However, it is only recently that the erosion of cultures and loss of indigenous knowledge have been recognized as threats to genetic diversity and remedial measures proposed.

Globally, the conveniences of vertically integrated food systems are driving consumers to neglect local foods, some of which were traditionally sourced from the wild [1]. This chapter will address the question of genetic loss, and erosion of indigenous food cultures as a preamble to making a case for investment in research on underutilized and alternative crops.

20.2 Genetic Loss Through Selection and Breeding

The transition by the human race from hunter-gatherer to a sedentary lifestyle marked the advent of managed crop production. These early beginnings of modern agriculture were characterized by selection and domestication of wild plants with desirable traits, e.g., precocity, high yield, compact growth, and other positive attributes. Through repeated selection from this initial population of domesticated species, early agriculturalists isolated the genetic pools that now represent crops of economic importance. Table 20.1 summarizes the chronology of events marking this process and the fate of plant genetic diversity thereafter [2].

Although the initial isolation of a few species through selection and domestication may have resulted in the preservation of certain traits at the expense of others, the most significant genetic loss can be traced to the pressures imposed on plant genetic diversity by modern plant breeding. As demonstrated by Tanksley and McCouch [3] in Fig. 20.1, domestication and breeding have served as bottlenecks to genetic transfer that in the long term have resulted in a drastic narrowing of the genetic base of most crops.

Modern plant breeding has contributed to genetic loss by selecting parental lines from a small number of highly productive varieties to which a majority of improved

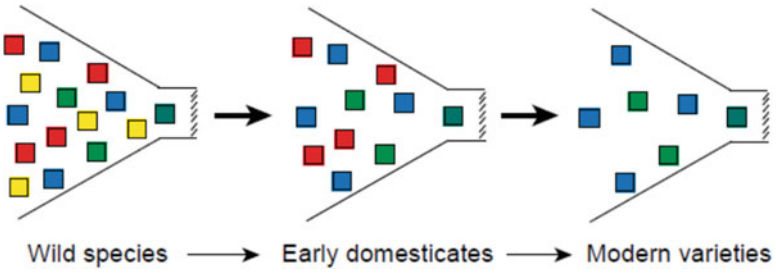


Fig. 20.1 Genetic bottlenecks imposed on crop plants during domestication and through modern plant-breeding practices. *Boxes* represent allelic variations of genes originally found in the wild, but gradually lost through domestication and breeding. Such lost alleles can be recovered only by going back to the wild ancestors of our crop species (*Source*: Tanksley and McCough, 1997)

Table 20.1 A history of the development and exchange of plant genetic resources

Era	Year	Milestone
Prehistoric era Domestication and geographical spread of crops	c. 8000 BC	• Humans start their transition from nomad hunters to sedentary farmers
	c. 4000 BC	• Cultural contacts and interactions result in crop diffusion and global transfer of plant genetic resources for food and agriculture (PGRFA)
The past 5 millennia Development of agriculture and agricultural biodiversity	c. 3000 BC	• Sumerians and Egyptians actively collect PGRFA
	1400	• The discovery of America boosts intercontinental exchange of PGRFA
The nineteenth and early twentieth centuries Science realizes the value and potential of genetic diversity	1845	• The Great Famine in Ireland dramatically demonstrates the need for genetic diversity in agriculture
	1850	• Charles Darwin and Gregor Mendel prove the importance of genetic diversity for biological evolution and adaptation
	1920	• Nikolai Vavilov identifies the main areas of crop origin and their genetic diversity
	1960	• The Green Revolution boosts productivity, but contributes to the loss of genetic diversity
The 1960s to mid-1970s Scientific and institutional developments take place, but concerns remain about genetic erosion and vulnerability	1965	• The Food and Agricultural Organization (FAO) of the United Nations (UN) starts technical work on PGRFA collection and conservation, including a series of international technical conferences
	1972	• The UN Stockholm Conference on the Human Environment calls for strengthening of PGRFA conservation activities The US National Academy of Sciences raises concern over the genetic vulnerability of crops after a major maize epidemic
	1974	• What is now the International Plant Genetic Resources Institute was established to support and catalyze collection and conservation efforts
Late 1970s to early 1990s The first major policy developments	1978	• The International Union for the Protection of New Varieties of Plants was established in 1978 and revised in 1991. National legislation in many countries restricts access to PGRFA, including through intellectual property rights

(continued)

Table 20.1 (continued)

Era	Year	Milestone
	1979	• FAO member countries start policy and legal discussions, leading to the first permanent intergovernmental forum on PGRFA in 1983—the Commission on Genetic Resources for Food and Agriculture (CGRFA)—and the adoption of the non-binding International Undertaking (IU) on Plant Genetic Resources
	1989	• Non-governmental organizations promote an international dialogue on PGRFA, reaching common understandings that feed into the CGRFA's negotiations
1992 to the present day An era of global instruments and legally binding agreements	1992	• The first international binding agreement on biological diversity in general, the Convention on Biological Diversity, is adopted. Its members recognize the special nature of agricultural biodiversity and support the negotiations of the FAO
	1993	• The CGRFA agrees to renegotiate the IU, resulting in the adoption in 2001 of the legally binding International Treaty for Plant Genetic Resources for Food and Agriculture (ITPGRFA) • The CGRFA develops the Leipzig Global Plan of Action on plant genetic resources and the first report on the state of the world's PGRFA
	1994	• The Marrakech Agreement on Trade-Related Aspects of Intellectual Property Rights is adopted
	1995	• The CGRFA broadens its mandate from only crops to all components of biodiversity for food and agriculture, including farm animals, forestry, and fisheries
	2004	• The ITPGRFA enters into force on 29 June
Agreed future steps	2006	• The first meeting of the governing body of the ITPGRFA will take place in June 2006 in Spain • The CGRFA is to adopt its Multi-Year Program of Work, covering all agricultural biodiversity sectors

Adapted by permission from Macmillan Publishers Ltd: Nature Reviews Genetics (Esquinas-Alcázar, J.: Protecting crop genetic diversity for food security: political, ethical and technical challenges, copyright 2005)

lines among cultivated crops can be traced. This narrowing of the genetic base has rendered most crops vulnerable to biotic and abiotic pressures and jeopardized the potential for future genetic improvement of economically important crops. A number of studies on the extent of genetic loss in crops of economic importance have been carried out with varying results. For example, Fu and Somers [4] link widespread allelic gene reduction in Canadian hard spring wheat starting from the 1930s to pressures exerted by modern plant breeding, while Duvick [5] noting an improvement in the genetic base of field crops in the USA in 1981 relative to 1970 concluded at the time that genetic vulnerability was not a major threat to US field crops.

In the case of soybean, it has been found that although the genetic bottlenecks imposed by breeding and selection have negatively impacted the genetic base, the most significant reduction in genetic diversity occurred at domestication when the low sequence diversity present in wild species was halved and 81 % of rare alleles lost [6]. Jordan et al. [7] report that selection for resistance to sorghum midge (*Stenodiplosis sorghicola*) in Australia had been achieved at the expense of diversity, a situation that may lead to genetic vulnerability, and in future, impact the rate of progress in breeding for yield. It is for the same reasons that Hammons [8] has proposed breeding strategies for widening the genetic base and increasing genotypic diversity of economically dominant peanut (*Arachis hypogea*) cultivars in the USA.

As a general theme, the various studies emphasize the dangers of genetic loss and identify mechanized monoculture as a major contributor to genetic erosion. In a review of the Indian green revolution, Safeeulla [9] associates the shift from traditional crops and age-old practices to high yielding varieties with increased vulnerability and potential for future crop epidemics. Jacques and Jacques [10] echo the same concerns about the dangers of mechanized monoculture both from the perspective of genetic loss and in terms of its contribution to the erosion of social and cultural diversity.

20.3 Genetic Loss Through Crop Loss

Most of what are considered lost crops comprise of species native to tropical and sub-tropical regions of Africa and Latin America that were the pre-colonial staple foods of indigenous communities. One of the pervasive consequences of colonialism was the introduction of food crops from the northern hemisphere and sequestration of land for large-scale production of industrial crops like rubber, sugarcane, and cotton. These activities diminished the stature of native crops and began a process of neglect and genetic loss from which native food crops are yet to recover. An excerpt from the introduction to the "Lost Crops of the Incas" notes how at the time of Spanish conquest, the locals cultivated a large number of diverse crops:

> On mountainsides up to 4 km high, along the spine of a whole continent, and in climates varying tropical to polar, they grew a wealth of roots, grains, legumes, vegetables, fruits and nuts [11].

The book identifies a significant number of Inca root and tuber crops, fruits, and some grains, legumes, vegetables, and nuts that would benefit from research and development to facilitate introduction and commercial production in regions beyond their centers of origin. The US National Research Council has also published three volumes on the Lost Crops of Africa covering grains, vegetables, and fruits. Among grains, the authors single out fonio (*Digitaria exilis*) and tef (*Eragrostis tef*) as consumed solely by Africans. Other species addressed include African rice (*Oryza glaberrima*), a number of millets and sorghums, and other wild and cultivated grains [12]. The edition on vegetables discusses 18 species of which some, e.g., Okra (*Abelmoschus esculentus*) and Cowpea (*Vigna unguiculata*) are cultivated worldwide but originated from Africa where there remains a large selection of uncharacterized germplasm [13]. The companion volume on fruits lists 10 cultivated and 14 wild species that according to the authors have virtually not been touched by science [14].

Although the most important indigenous crops have been identified and their cultural and economic value acknowledged, the threat of extinction still exists. In the current era, the loss of native species including indigenous crops can be linked to environmental pressures exerted by population growth as documented by Cincotta et al. [15] and others like Maurer [16] who has demonstrated a direct relationship between the increasing fraction of solar energy consumed by humans and loss in biodiversity. Furthermore, human activities common to many developing countries, e.g., uncontrolled logging, unregulated mining, slash and burn agriculture, overgrazing, and commercial hunting can lead to ecosystem degradation and species loss even in the absence of widespread human settlement.

Climate change is another factor that is contributing to habitat loss and a shift in species composition. For example, climate models predict a 51–65 % loss of the Fynbos biome in South Africa by 2050, an event that would result in complete dislocation of up to 10 % of Proteaceae endemic to the region [17]. This observation underlines the need for urgent intervention measures particularly in developing countries with limited infrastructure for tracking and reporting changes in climate and biodiversity.

20.4 Global Response to Genetic Vulnerability and Crop Loss

In the nineteenth century Vavilov identified the main areas of origin and genetic diversity of cultivated plants and highlighted the potential of wild relatives as sources of genetic material for improving modern crops [18]. His work and other events including the Southern corn leaf blight epidemic of 1970–1971 [19] motivated the collection of races and species related to cultivated plants and the establishment of gene banks.

Currently, collection, cataloging, and preservation of germplasm are widespread and well-coordinated with most nations maintaining repositories for economically important crop accessions and landraces. In the USA, the system established by congress after the 2nd World War to maintain and distribute plant genetic resources has grown into a National Plant Germplasm System consisting of 26 repositories with more than half a million individual collections [20]. Another notable initiative in the effort to preserve plant genetic diversity is the Svalbard Global Seed Vault in Norway. Established by Cary Fowler jointly with the Consultative Group on International Agricultural Research (CGIAR), the vault is serving as a repository for duplicate copies of seeds held in gene banks worldwide with more than 840,000 samples from 4000 distinct species secured by the year 2015.

The science of conservation spearheaded largely by international agencies like the Food and Agriculture Organization (FAO) of the United Nations, Bioversity International, and the International Plant Genetic Resources Institute (IPGRI) has progressed apace. Together with other agencies and institutions of higher learning, they have generated considerable knowledge and information on the science and practice of conservation, characterization, and sharing of genetic material. Good examples include a manual for in situ conservation of crop wild relatives released by Bioversity International in 2011 [21], and the recently published revision to gene bank standards for plant genetic resources [22].

Hammer [23] examines what he finds to be defining shifts in the field of plant genetic resources driven largely by scientific advancements starting from the 1990s. These changes include:

(i) Increasing emphasis on in situ as opposed to ex situ conservation, against which he recommends a judicious and balanced approach.
(ii) A shift in priority from major cultivated species to underutilized and neglected crops as a means of maintaining species diversity. This change is motivated by the realization that of the more than 7000 cultivated plants, only 100 account for a majority of holdings in gene banks worldwide.
(iii) Utilization of emerging tools including genetic analysis to expand on existing collection strategies by examining landrace populations and their potential productive components in order to optimally conserve genetic diversity.
(iv) Exploration of approaches for decreasing gene erosion by increasing participation in maintenance and use beyond the gene bank. Others including Tanksley and McCouch [3] and Wright [24] have previously called for policy and technical changes to optimize the utilization of genetic material already at the disposal of gene banks.
(v) Greater emphasis on molecular tools for the evaluation of gene bank material. Here too Hammer [23] warns against wholesale neglect of traditional methods and recommends measured application of both new and old technologies.
(vi) Recognition of the need for a strategy on large-scale reproduction and replacement of ex situ accessions before loss of viability.

In conclusion, Hammer [23] puts forward the idea of an integrated gene bank that combines the new aspects of molecular biology and biodiversity with classical

conservation, evaluation, and utilization tasks while staying abreast of in situ and other dynamic conservation strategies.

However, as observed by Ledig [25], the technical aspects of gene preservation ex situ in seed banks and arboreta or in situ in reserves or special management areas are fairly simple. It is ecosystem level processes, e.g., the roles that co-evolution and adapted gene complexes play in the preservation of genetic diversity that still pose difficult research questions. Furthermore, as noted in his examination of strategies for conserving forest genetic resources, socio-economic factors associated with land as an economic resource may be the greatest challenge to genetic conservation in situ, an observation supported by Zimmer [26] who found that the value of native crops in the southern Peruvian highlands was impacted by shifts in access to land, labor, and capital, the socio-cultural value of the crop, and the biogeographic patterning of cultivars. For this reason, he recommends that in situ conservation programs be pursued only after fully understanding conditions upon which continued production of target crops is contingent.

Genetic erosion through neglect or loss of indigenous crops and knowledge has not received as much attention as genetic loss due to selection and breeding. Interest in indigenous plants with economic or cultural value started in the 1950s and continues to increase but the many works on ethnobotany and ethnomedicine (Fig. 20.2) largely document novel and historical usage. It is only in the 1980s that serious effort to characterize, conserve, and improve indigenous crops and other neglected species started with the formation of the International Center for

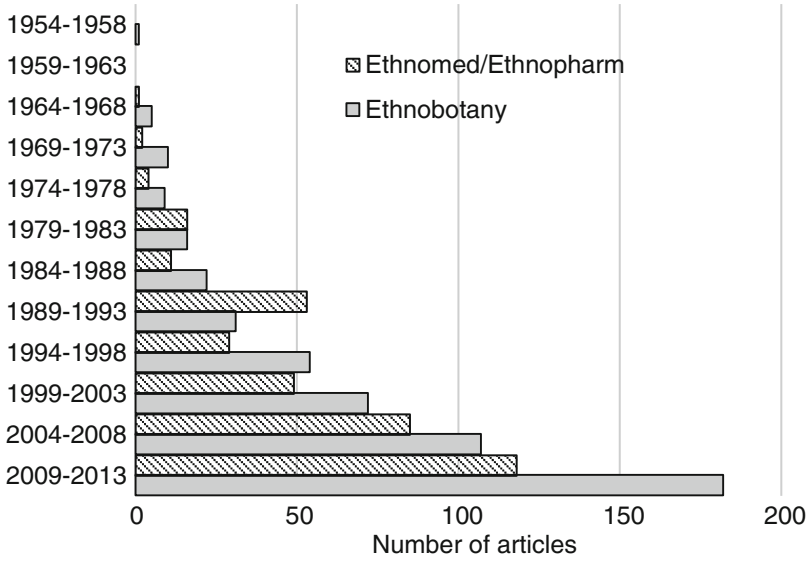


Fig. 20.2 Number of articles with ethnobotany, or either ethnomedicine or ethnopharmacology, in the title that were published in peer-reviewed periodicals between 1954 and 2013 (*Source*: Scopus, http://www.scopus.com/home.url)

Underutilized Crops (ICUC), an independent research institute that supported and coordinated research programs to increase productivity and use of neglected and underutilized crops. Renamed Crops for the Future (CFF) after merging with the Global Facilitation Unit for Underutilized Species (GFU) in 2009, the organization continues to coordinate and conduct research on underutilized crops. Other organizations including FAO, Bioversity International, IPGRI, and international development agencies like the Department for International Development-DFID (United Kingdom) and the German Deutsche Gesellschaft für Technische Zusammenarbeit (GTZ) are also involved in research and development on indigenous crops and other underutilized species.

Review of literature suggests that the effort is starting to bear fruit. A sampling from an emerging stream of peer-reviewed work on indigenous crops and other underutilized species research shows activity in diverse areas including conservation and use [27–29], climate change [15, 17, 30], agronomy [31], genotyping [32], breeding [33], human nutrition [34–37], and market research [38–41]. Meetings dedicated solely to underutilized species have also been convened, e.g., by the American Society for Horticultural Science [42], the International Atomic Energy Agency [43], the International Society for Horticultural Science [44–46], and Bioversity International [47]. However, there remains a lot of ground to be covered before a fully functional research and development framework for underutilized plant species can be realized [48, 49].

20.5 Research on Underutilized and Alternative Crops

As mentioned in the introduction, limited research and development on neglected and underutilized species is a contributor to gene erosion and food insecurity. Most of such species are to be found among indigenous communities where they are grown or harvested for food, medicine, or cultural purposes. For example, van Andel [50] found that descendants of enslaved Africans in Suriname still grow African rice (*Oryza glaberrima*) for food and ritual. The same applies to most cultivated/harvested species native to Africa that are discussed in the volumes on lost grains, vegetables, and fruits published by the US National Research Council [11–14].

A defining quality of these species that arises directly from limited genetic manipulation is strong adaptability and resilience. A majority of the plants are produced in low-intensity, low input systems, and some survive in the wild. Most of them are also highly efficient in resource utilization and show a high tolerance for regional biotic and abiotic stresses. The trend to direct attention away from major cultivated crops to underutilized species reported by Hammer [23] as among paradigm shifts informing the field of plant genetic resource conservation is driven in part by the realization that some underutilized species possess genes and traits that may be useful for future crop improvement and global food security.

The fact that a large number of indigenous crops and other plants of economic importance native to the tropical and sub-tropical belt have not received commensurate

scientific attention is a major obstacle to their preservation and commercialization. They are not yet optimized for large-scale production and postharvest processing, and are vulnerable to displacement by introduced species including newly developed genetically modified strains. Limited scientific testing also means that the responses of most native germplasm to ecological shifts associated with global climate change remain unknown. In their totality, these circumstances preface potential widespread decline in genetic diversity and crop loss that in Africa and other regions presently witnessing rapid urbanization and other social change will have serious implications for food security and socio-economic sustainability.

Furthermore, it must not be assumed that neglect of species with economic potential is limited to the regions addressed in the four cited volumes published by the US National Research Council. The highly evolved food production and distribution systems common to the USA and other developed countries ensure that consumers have year-round access to a limited variety of fresh produce and no longer have to eat according to season, or live off the land. For this reason, a number of edible plants originally grown as niche crops or foraged from the wild are no longer consumed and have fallen into neglect. For example, in a survey of Virginia flora, we found more than 500 wild and cultivated species with an edibility or medicinal rating higher than 3 (max 5) as ranked by the Plants for a Future database (Fig. 20.3). We also found that about half of species rated 5 either for edibility or medicinal value were introduced (Table 20.2), confirming their importance to earlier settlers [51]. Similarly, Stamp et al. [52] observe that a majority of underutilized

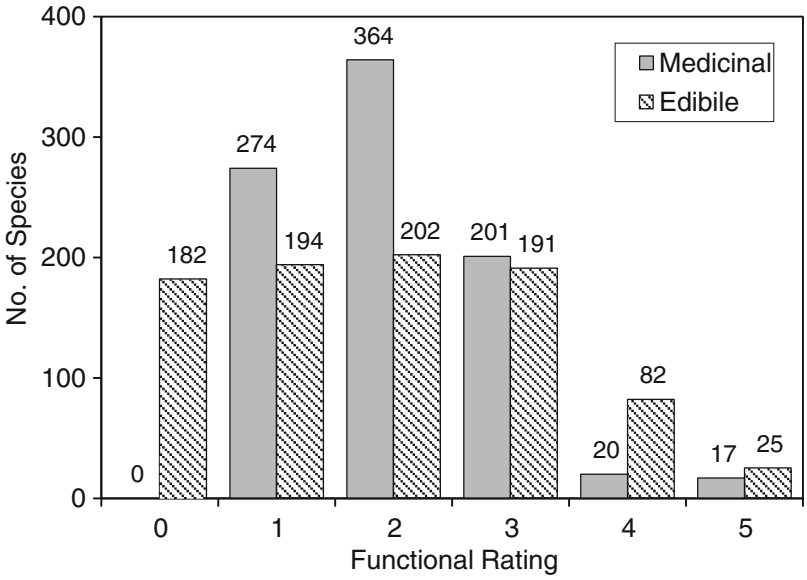


Fig. 20.3 Number of wild and cultivated edible and medicinal plants found in Virginia (Digital Atlas of the Virginia Flora: http://vaplantatlas.org/). Functional ratings (0–5) courtesy of Plants for a Future (http://pfaf.org/)

Table 20.2 Wild and cultivated edible and medicinal plants found in Virginia

Common name	Botanical name	Edible[a]	Medicinal[a]	Origin[b]
Marsh Mallow	*Althaea officinalis*	5	5	Not specified
Stinging Nettle	*Urtica dioica*	5	5	Introduced
Wild Leek	*Allium ampeloprasum*	5	3	Introduced
Winter Squash	*Cucurbita maxima*	5	3	Introduced
Squash	*Cucurbita moschata*	5	3	Introduced
Fennel	*Foeniculum vulgare*	5	3	Introduced
Peach	*Prunus persica*	5	3	Introduced
Raspberry	*Rubus idaeus*	5	3	Native
Sassafras	*Sassafras albidum*	5	3	Native
Small Reed Mace	*Typha angustifolia*	5	3	Native
Reedmace	*Typha latifolia*	5	3	Native
Sweet Violet	*Viola odorata*	5	3	Introduced
Nodding Onion	*Allium cernuum*	5	2	Native
Chives	*Allium schoenoprasum*	5	2	Introduced
Hawthorn Hybrid	*Crataegus missouriensis*	5	2	Native
Goumi	*Elaeagnus multiflora*	5	2	Introduced
Elaeagnus	*Elaeagnus pungens*	5	2	Introduced
Sunflower	*Helianthus annuus*	5	2	Introduced
Common Day Lily	*Hemerocallis fulva*	5	2	Introduced
Musk Mallow	*Malva moschata*	5	2	Introduced
Common Reed	*Phragmites australis*	5	2	Not specified
Plum	*Prunus domestica*	5	2	Introduced
Ramanas Rose	*Rosa rugosa*	5	2	Introduced
American Persimmon	*Diospyros virginiana*	5	1	Native
Duck Potato	*Sagittaria latifolia*	5	1	Native
Hop	*Humulus lupulus*	4	5	Introduced
Balsam Fir	*Abies balsamea*	3	5	Introduced
Lesser Burdock	*Arctium minus*	3	5	Introduced
Shatavari	*Asparagus racemosus*	3	5	Introduced
Lemon Balm	*Melissa officinalis*	3	5	Introduced
Evening Primrose	*Oenothera biennis*	3	5	Native
Sage	*Salvia officinalis*	3	5	Introduced
Milk Thistle	*Silybum marianum*	3	5	Introduced
Comfrey	*Symphytum officinale*	3	5	Introduced
Slippery Elm	*Ulmus rubra*	2	5	Native
Agnus-Castus	*Vitex agnus-castus*	2	5	Introduced
Echinacea	*Echinacea purpurea*	1	5	Introduced
Witch Hazel	*Hamamelis virginiana*	1	5	Native
German Camomile	*Matricaria recutita*	1	5	Introduced

[a]Edibility and medicinal ratings courtesy of Plants for a Future (http://pfaf.org/)
[b]Information on origin obtained from the Digital Atlas of the Virginia Flora (http://vaplantatlas.org/)

crops that were once widely grown in Europe are no longer suitable for today's agriculture and call for long-term breeding programs and other initiatives to stem the erosion of biodiversity.

20.6 Conclusion

Conservation, improvement, and commercialization of indigenous and underutilized crops demand urgent and coordinated effort on a scale similar to that invested in ex situ preservation of genetic diversity. This is necessitated by global climate change, human encroachment, and other socio-economic factors that cumulatively threaten biodiversity. According to Williams and Haq [47], there is still a lot of work that remains to be done. In their assessment of the state of global research on underutilized crops, they found that despite resurgent interest in underutilized crops as well as recognition of the interconnection between agriculture and the environment, policies supportive of underutilized crops remain underdeveloped.

Furthermore, this field of research is unlikely to attract private investment because underutilized crops are generally unsuited to modern agriculture [52]. As observed by Rubenstein et al. [53], private investment in conservation often falls far short of public objectives even for established crops because plant genetic resources are considered a public good. This puts the onus on governments to enact favorable policies, and to dedicate resources in support of work on underutilized and other alternative crops.

References

1. Khoury, C.K., Bjorkman, A.D., Dempewolf, H., et al.: Increasing homogeneity in global food supplies and the implications for food security. Proc. Natl. Acad. Sci. U. S. A. **111**(11), 4001–4006 (2014)
2. Esquinas-Alcázar, J.: Protecting crop genetic diversity for food security: political, ethical and technical challenges. Nat. Rev. Genet. **6**, 946–953 (2005)
3. Tanksley, S.D., McCouch, R.S.: Seed banks and molecular maps: unlocking genetic potential from the wild. Science **277**, 1063–1066 (1997)
4. Fu, Y.-B., Somers, D.J.: Genome-wide reduction of genetic diversity in wheat breeding. Crop. Sci. **49**, 161–168 (2009)
5. Duvick, D.N.: Genetic diversity in major farm crops on the farm and reserve. Econ. Bot. **38**(2), 161–178 (1984)
6. Hyten, D.L., Song, Q., Zhu, Y., Choi, I.-Y., Nelson, R.L., Costa, J.M., Specht, J.E., Shoemaker, R.C., Cregan, P.B.: Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl. Acad. Sci. U. S. A. **103**(45), 16666–16671 (2006)
7. Jordan, D.R., Tao, Y.Z., Godwin, I.D., Henzell, R.G., Cooper, M., McIntyre, C.L.: Loss of genetic diversity associated with selection for resistance to sorghum midge in Australian sorghum. Euphytica **102**, 1–7 (1998)

8. Hammons, R.O.: Peanuts: genetic vulnerability and breeding strategy. Crop. Sci. **16**, 527–530 (1976)
9. Safeeulla, K.M.: Genetic vulnerability: the basis of recent epidemics in India. Ann. N. Y. Acad. Sci. **287**, 72–85 (1977)
10. Jacques, P.J., Jacques, J.R.: Monocropping cultures into ruin: the loss of food varieties and cultural diversity. Sustainability **4**, 2970–2997 (2012)
11. National Research Council: Lost Crops of the Incas: Little Known Plants of the Andes with Promise for Worldwide Cultivation. National Academy Press, Washington, DC (1989)
12. National Research Council: Lost Crops of Africa. Volume I: Grains. National Academy Press, Washington, DC (1996)
13. National Research Council: Lost Crops of Africa. Volume II: Vegetables. National Academy Press, Washington, DC (2006)
14. National Research Council: Lost Crops of Africa. Volume III: Fruits. National Academy Press, Washington, DC (2008)
15. Cincotta, R.P., Wisnewski, J., Engelman, R.: Human population in the biodiversity hotspots. Nature **404**, 990–992 (2000)
16. Maurer, B.A.: Relating human population growth to loss of biodiversity. Biodivers. Lett. **3**(1), 1–5 (1996)
17. Midgley, G.F., Hannah, L., Millar, D., Rutherford, M.C., Powrie, L.W.: Assessing the vulnerability of species richness to anthropogenic climate change in a biodiversity hotspot. Global Ecol. Biogeogr. **11**, 445–451 (2002)
18. Vavilov, N.I.: The new systematic of cultivated plants. In: Huxley, J. (ed.) The New Systematics, pp. 549–566. Clareon Press, Oxford (1940)
19. Ullstrup, A.J.: Impacts of Southern corn leaf blight epidemics of 1970–1971. Annu. Rev. Phytopathol. **10**(1), 37 (1972)
20. Johnson, R.C.: Gene banks pay big dividends to agriculture, the environment, and human welfare. PLoS Biol. **6**(6), e148 (2008)
21. Hunter, D., Heywood, V. (eds.): Crop Wild Relatives: A Manual of In Situ Conservation, 1st edn. Bioversity International, New York, NY (2011)
22. FAO: Genebank Standards for Plant Genetic Resources for Food and Agriculture. Rev. Ed., Rome (2014)
23. Hammer, K.: A paradigm shift in the discipline of plant genetic resources. Genet. Resour. Crop. Evol. **50**, 3–10 (2003)
24. Wright, B.D.: Crop genetic resource policy: the role of ex situ genebanks. Aust. J. Agr. Resour. Econ. **41**(1), 81–115 (1997)
25. Ledig, F.T.: Conservation strategies for forest gene resources. For. Ecol. Manage. **14**, 77–90 (1986)
26. Zimmer, K.S.: The loss and maintenance of native crops in mountain agriculture. GeoJournal **27**(1), 61–72 (1992)
27. Padulosi, S., Eyzaquirre, P., Hodgkin, T.: Challenges and strategies in promoting conservation and use of neglected and underutilized crop species. In: Janick, J. (ed.) Perspectives on New Crops and Uses. ASHS Press, Alexandria, VA (1999)
28. Brussaard, L., Caron, P., Campbell, B., Lipper, L., Mainka, S., Rabbinge, R., Babin, D., Pulleman, M.: Reconciling biodiversity conservation and food security: scientific challenges for a new agriculture. Curr. Opin. Environ. Sustain. **2**, 34–42 (2010)
29. Sthapit, B., Padulosi, S., Mal, B.: Role of on-farm/in situ conservation on underutilized crops in the wake of climate change. Indian J. Plant Genet. Resour. **23**(2), 145–156 (2010)

30. Padulosi, S., Heywood, V., Hunter, D., Jarvis, A.: Underutilized species and climate change: current status and outlook. In: Yadav, S.S., Redden, R.J., Hatfield, J.L., Lotze-Campen, H., Hall, A.E. (eds.) Crop Adaptation to Climate Change, pp. 507–521. Blackwell Publishing Ltd. (2011)

31. Mal, B., Padulosi, S., Ravi, S.B. (eds.): Minor Millets in South Asia: Learnings from IFAD-NUS Project in India and Nepal. Bioversity International, Maccarese, Rome, Italy and the M.S. Swaminathan Research Foundation, Chennai, India (2010)

32. Pearl, S.A., Burke, J.M.: Genetic diversity in *Carthamus tinctorius* (Asteraceae; Safflower), an underutilized oilseed crop. Am. J. Bot. **101**(10), 1640–1650 (2014)

33. Ochatt, S., Jain, S.M. (eds.): Breeding of Neglected and Under-Utilized Crops, Spices and Herbs. Science Publishers, Enfield, NH (2007)

34. Vuong, L.T.: Underutilized β-carotene-rich crops of Vietnam. Food Nutr. Bull. **21**(2), 173–181 (2000)

35. Flyman, M.V., Afolayan, A.J.: The suitability of wild vegetables for alleviating human dietary deficiencies. S. Afr. J. Bot. **72**, 492–497 (2006)

36. Pande, G., Akoh, C.C.: Organic acids, antioxidant capacity, phenolic content and lipid characterization of Georgia-grown underutilized fruit crops. Food Chem. **120**, 1067–1075 (2010)

37. Schönfeldt, H.C., Pretorius, B.: The nutrient content of five traditional South African dark green leafy vegetables-a preliminary study. J. Food Compos. Anal. **24**, 1141–1146 (2011)

38. Mwangi, S., Kimathi, M.: African leafy vegetables evolve from underutilized species to commercial cash crops. Research Workshop on Collective Action and Market Access for Smallholders. Cali, Colombia, October 2–5 (2006)

39. Gruère, G., Giuliani, A., Smale, M.: Marketing underutilized plant species for the benefit of the poor: a conceptual framework. EPT Discussion paper 154. IFPRI, Washington, DC (2006)

40. Horna, D., Timpo, S., Gruère, G.: Marketing Underutilized Crops: The Case of the African Garden Egg (*Solanum aethiopicum*) in Ghana. Global Facilitation Unit for Underutilized Species (GFU), Rome (2007)

41. Gruère, G., Nagarajan, L., King, E.D.I.O.: The role of collective action in the marketing of underutilized plant species: lessons from a case study on minor millets in South India. Food Policy **34**, 39–45 (2009)

42. Janick, J. (ed.): Perspectives on New Crops and Uses. ASHS Press, Alexandria, VA (1999)

43. IAEA: Genetic improvement of under-utilized and neglected crops in low income food deficit countries through irradiation and related techniques. In: Proceedings of a final Research Coordination Meeting organized by the joint FAO/IAEA Division of Nuclear Techniques in Food and Agriculture, Pretoria, South Africa, May 19–23, 2003 (2004)

44. Oh, D.-G., Kubota, C.: Proceedings of the International symposium on cultivation and utilization of Asian, sub-tropical, and underutilized horticultural crops. Acta Hortic. **770** (2008). 212 pp

45. Jaenicke, H., Ganry, J., Hoeschle-Zeledon, I., Kahane, R.: Proceedings of the International symposium on underutilized plants for food, nutrition, income and sustainable development. Acta Hortic. **806**(1–2) (2008). 739 pp

46. Massawe, F., Mayes, S., Alderson, P.: Proceedings of the second international symposium on underutilized plant species: crops for the future - beyond food security. Acta Hortic. **979**(1–2) (2011). 806 pp

47. Padulosi, S., Bergamini, N., Lawrence, T. (eds.): On-farm conservation of neglected and underutilized species: status, trends and novel approaches to cope with climate change. In: Proceedings of an International Conference, Frankfurt, 14–16 June, 2011. Bioversity International, Rome (2012)

48. Williams, J.T., Haq, N.: Global Research on Underutilized Crops. An Assessment of Current Activities and Proposals for Enhanced Cooperation. ICUC, Southampton, UK (2002)
49. Jaenicke, H., Höschle-Zeledon, I. (eds.): Strategic Framework for Underutilized Plant Species Research and Development, with Special Reference to Asia and the Pacific, and to Sub-Saharan Africa. International Centre for Underutilized Crops, Colombo, Sri Lanka and Global Facilitation Unit for Underutilized Species, Rome, Italy (2006)
50. Van Andel, T.: African rice (*Oryza glaberrima* Steud.): lost crop of the enslaved Africans discovered in Suriname. Econ. Bot. **64**, 1–10 (2009)
51. The Flora of Virginia Project. Digital atlas of the Virginia flora. http://www.vaplantatlas.org/. Accessed March 2014
52. Stamp, P., Messmer, R., Walter, A.: Competitive underutilized crops will depend on state funding of breeding programmes: an opinion on the example of Europe. Plant Breed. **131**, 461–464 (2012)
53. Rubenstein, K.D., Heisey, P., Shoemaker, R. Sullivan, J., Frisvold, G.: Crop Genetic Resources: An Economic Appraisal. EIB-2, U.S. Department of Agriculture, Economic Research Service (2005)

Chapter 21
Uncovering Cluster Structure and Group-Specific Associations: Variable Selection in Multivariate Mixture Regression Models

Mahlet G. Tadesse, Frédéric Mortier, and Stefano Monni

Abstract Variable selection for mixture of regression models has been the focus of much research in recent years. These models combine the ideas of mixture models, regression models, and variable selection to uncover group structures and key relationships between data sets. The objective is to identify homogeneous groups of objects and determine the cluster-specific subsets of covariates modulating the outcomes. In this chapter we review frequentist and Bayesian methods we have proposed to address in a unified manner the problems of cluster identification and cluster-specific variable selection in the context of mixture of regression models. These methods have a wide range of applications, in particular in the context of high-dimensional data analysis. We illustrate their performance in two diverse areas: one in ecology for modeling species-rich ecosystems and the other in genomics for integrating data from different genomic sources.

Keywords Adaptive lasso • Markov chain Monte Carlo • Mixture regression models • Multivariate outcomes • Stochastic partitioning • Variable selection

M.G. Tadesse (✉)
Department of Mathematics and Statistics, Georgetown University, 37th & O Streets NW, Washington, DC 20057, USA
e-mail: mgt26@georgetown.edu

F. Mortier
UPR Biens et Services des Ecosystémes Forestiers Tropicaux (B&SEF), CIRAD, TA C-105/D Campus International de Baillarguet, 34398 Montpellier Cedex 5, France

S. Monni
Department of Mathematics, American University of Beirut, P.O. Box 11-0236, Riad El Solh, Beirut 1107 2020, Lebanon

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_21

21.1 Introduction

Mixtures of regression models provide an effective tool to understand structures and relationships in complex systems, by identifying clusters of objects that behave similarly and fitting regression models specific to each cluster. When faced with a large number of covariates, as often happens in practice, it is also necessary to select the relevant predictors for each cluster. Such selection improves the identification of homogeneous groups and provides a better understanding of the underlying processes generating the data. A unified method that simultaneously uncovers clusters and selects group-specific relevant covariates has several advantages over a two-stage approach that first clusters the objects and then performs variable selection within each cluster. In particular, the latter would ignore the uncertainty in estimating the cluster allocations, thus introducing bias in the variable selection and estimation procedures.

In mixture of multivariate regression models, the data consist of n independent samples with p covariates, $X_1, \ldots, X_p$, and q outcomes, $Y_1, \ldots, Y_q$. That is, the observations $(\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1,\ldots,n}$ are realizations of the pair of random variables $(\boldsymbol{X}, \boldsymbol{Y})$ with $\boldsymbol{X} \in \mathbb{R}^p$ and $\boldsymbol{Y} \in \mathbb{R}^q$. The goal is to cluster objects with similar profiles conditional on their identical dependence on a subset of covariates. In our applications of interest, we want to cluster the outcome variables rather than the subjects. For example, in the ecological application we consider in Sect. 21.4.1, the goal is to cluster q species for each of which n trees are measured. In the genomic application of Sect. 21.4.2, we want to cluster the expression phenotypes of q genes with each gene expression measured on n independent samples. The mixture regression model for outcome j ($j = 1, \ldots, q$) is then given by

$$f(\boldsymbol{y}_j|\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{y}_j|\boldsymbol{x}, \boldsymbol{\theta}_k) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{n} f(y_{ji}|\boldsymbol{x}_i, \boldsymbol{\theta}_k) \tag{21.1}$$

where the number of components K is unknown and needs to be determined, π_k corresponds to the mixture weights, $f(.)$ denotes the probability density/mass function (Gaussian, Poisson, binomial, etc.) defined for the k-th component in terms of the relevant covariates and component parameters $\boldsymbol{\theta}_k$, and n corresponds to the sample size for each object j such that $\boldsymbol{y}_j = (y_{j1}, \ldots, y_{jn})$. In Gaussian mixture regressions, $\boldsymbol{\theta}_k = (\boldsymbol{\beta}_k, \sigma_k)$, while in other mixtures of generalized linear models (GLM) $\boldsymbol{\theta}_k = \boldsymbol{\beta}_k$, the component-specific regression parameters. Various frequentist and Bayesian methods have been proposed to fit this model. The former mostly rely on penalized maximum likelihood estimation using the expectation–maximization (EM) algorithm. The latter use stochastic search techniques within a Markov chain Monte Carlo (MCMC) framework.

This chapter is organized as follows. In Sect. 21.2, we present penalized mixture of regression models with an emphasis on a method we have proposed for variable selection in mixtures of multivariate GLM in the context of Gaussian, Poisson, and Bernoulli outcomes. Section 21.3 focuses on Bayesian methods and, in particular,

a computationally efficient stochastic partitioning method we have proposed for uncovering clusters and identifying cluster-specific relevant covariates. Section 21.4 presents applications of these models to two real data examples, one in ecology and the other in genomics. We conclude the paper with some discussion in Sect. 21.5.

21.2 Penalized Mixture Regressions

In the frequentist setting, variable selection and estimation of the mixture components' regression parameters are accomplished using penalized maximum likelihood methods. Khalili and Chen [5] proposed an EM algorithm for variable selection using lasso penalty in mixtures of univariate linear regression models under the standard assumption that the dimension p of the feature space is smaller than the sample size. Städler et al. [10] further studied the properties of lasso for these models in the context of high-dimensional data ($p \gg n$).

In Mortier et al. [9], we extended these penalized methods to mixtures of multivariate generalized linear regression models. This allows us to consider multiple outcomes simultaneously, as well as to move beyond Gaussian models and handle categorical outcomes (binary or count data). Taking the product of the mixture distribution in Eq. (21.1) over the q objects, the incomplete data likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta}\,|\boldsymbol{Y}) = \prod_{j=1}^{q}\left[\sum_{k=1}^{K}\pi_k \prod_{i=1}^{n} f(y_{ji}|\boldsymbol{x}_i, \boldsymbol{\theta}_k)\right]$$

and the incomplete data log-likelihood is given by

$$l(\boldsymbol{\theta}\,|\boldsymbol{Y}) = \sum_{j=1}^{q}\log\left[\sum_{k=1}^{K}\pi_k \prod_{i=1}^{N} f(y_{ji}|\boldsymbol{x}_i, \boldsymbol{\theta}_k)\right] \tag{21.2}$$

where $f(\cdot)$ is the appropriate Gaussian, Poisson, or Bernoulli probability density/mass function. In order to identify the component-specific relevant covariates, we used the adaptive lasso approach [11]. The model parameters $\boldsymbol{\theta}$ are estimated by maximizing the penalized log-likelihood function

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\left\{l(\boldsymbol{\theta}\,|\boldsymbol{Y}) - \mathcal{P}(\boldsymbol{\theta})\right\}, \qquad \mathcal{P}(\boldsymbol{\theta}) = \sum_{k=1}^{K}\pi_k\lambda_k \sum_{r=1}^{p}\frac{|\theta_{kr}|}{|\hat{\theta}_{kr}|} \tag{21.3}$$

where $\mathcal{P}(\boldsymbol{\theta})$ corresponds to the adaptive lasso penalty, θ_{kr} is the r-th element of $\boldsymbol{\theta}_k$, $|\hat{\theta}_{kr}|$ is the maximum likelihood estimator of θ_{kr}, and λ_k is a tuning parameter selected via cross-validation.

The maximization is accomplished using an EM algorithm. The data are augmented by introducing cluster allocation indicator variables z_{jk} $(j = 1, \ldots, q)$, such that $z_{jk} = 1$ if outcome j is from component k. The mixture distribution for $\boldsymbol{y}_j$ arising from cluster k can then be written as

$$f(\boldsymbol{y}_j|\boldsymbol{x}, \boldsymbol{\theta}) = \pi_k \cdot f(\boldsymbol{y}_j|\boldsymbol{x}, \boldsymbol{\theta}_k, z_{jk} = 1) = \prod_{k=1}^{K} \left[\pi_k f(\boldsymbol{y}_j|\boldsymbol{x}, \boldsymbol{\theta}_k, z_{jk})\right]^{z_{jk}}$$

and the complete-data likelihood becomes

$$\mathcal{L}(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{Z}) = \prod_{j=1}^{q} \prod_{k=1}^{K} \left[\pi_k f(\boldsymbol{y}_j|\boldsymbol{x}, \boldsymbol{\theta}_k, z_{jk})\right]^{z_{jk}}.$$

The complete-data log-likelihood is therefore given by

$$l(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{Z}) = \sum_{j=1}^{q} \sum_{k=1}^{K} z_{jk} \log \left[\pi_k \prod_{i=1}^{n} f(y_{ji}|\boldsymbol{x}_i, \boldsymbol{\theta}_k)\right]. \tag{21.4}$$

The E-step consists in taking the expectation of the complete-data log-likelihood and provides the posterior probability of assigning outcome j to component k. At iteration t of the algorithm, we estimate

$$w_{jk}^{(t)} = P\left(z_{jk} = 1|\boldsymbol{y}_j, \boldsymbol{x}_i, \boldsymbol{\theta}_k^{(t-1)}\right) = \frac{\pi_k^{(t-1)} \prod_{i=1}^{n} f\left(y_{ji}|\boldsymbol{x}_i, \boldsymbol{\beta}_k^{(t-1)}\right)}{\sum_{l=1}^{K} \pi_l^{(t-1)} \prod_{i=1}^{n} f\left(y_{ji}|\boldsymbol{x}_i, \boldsymbol{\beta}_l^{(t-1)}\right)} \tag{21.5}$$

and we adopt the approximation used in [5] to update the mixing proportions π_k

$$\pi_k^{(t)} = \frac{1}{nq} \sum_{j=1}^{q} \sum_{i=1}^{n} w_{jk}^{(t)}.$$

In the M-step, the expectation of the penalized complete log-likelihood is maximized for each component separately using the posterior allocation probabilities as weights

$$\boldsymbol{\beta}_k^{(t)} = \underset{\boldsymbol{\beta}_k}{\text{argmax}} \left\{ \sum_{j=1}^{q} \sum_{i=1}^{n} w_{jk}^{(t)} \log f(y_{ji}|\boldsymbol{x}_i, \boldsymbol{\beta}_k) - \pi_k^{(t)} \lambda_k \frac{|\boldsymbol{\beta}_k|}{|\hat{\boldsymbol{\beta}}_k|} \right\}. \tag{21.6}$$

We determine the number of components K using the integrated completed likelihood (ICL) criterion [1]. This criterion is similar in spirit to AIC or BIC and consists of introducing in the maximized log-likelihood a penalty term for

the number of parameters estimated in a model. It has the advantage of being specifically developed for mixture models and takes into account the quality of classifications. The proposed multivariate mixture of GLM with variable selection is fit for varying number of components, K, and the value that minimizes the ICL is chosen.

21.3 Bayesian Variable Selection in Mixture Regressions

In the Bayesian framework, variable selection is typically performed by introducing a latent binary indicator, γ_r $(r = 1, \ldots, p)$, taking value 1 if the corresponding covariate X_r is included in the model and 0 otherwise [2]. This latent vector is used to search the model space. In the context of mixture of regressions, this latent vector can be introduced for each component k $(k = 1, \ldots, K)$, resulting in a $K \times p$ indicator matrix. The problem of variable selection is much more challenging in this context compared to standard regression model settings because the membership of objects to the different components is not known and needs to be learned simultaneously with the search of component-specific predictors. Another complication is that the number of components K is unknown. Gupta and Ibrahim [4] proposed, for fixed K, an MCMC algorithm that iterates between the following steps: (1) the binary variable selection indicator matrix, $\boldsymbol{\gamma}$, is updated using an evolutionary Monte Carlo method; (2) the cluster allocation vector $\boldsymbol{z}$ is updated from its full conditional distribution via Gibbs sampling; (3) the component parameters, $\sigma_k^2, \boldsymbol{\beta}_k, \pi_k$, are updated from their posterior distributions. This MCMC procedure is repeated for varying values of K and the best value is determined by comparing the different models using Bayes factors evaluated via importance sampling procedures.

When there is a large number of objects to cluster, updating the cluster allocation for each object using Gibbs sampling becomes computationally burdensome. In addition, the update of the variable selection indicator vector $\boldsymbol{\gamma}_k$ for each of the K components can be quite expensive when K is large. Finally, fitting the model with different values of K to determine the number of components can be computationally expensive and fails to capture the uncertainty on the number of clusters. In Monni and Tadesse [7], we proposed a stochastic partitioning method that overcomes these limitations by constructing a Markov chain in the space of pairwise partitions of the set of regressors, $\boldsymbol{X}$, into possibly non-disjoint subsets and of the set of responses, $\boldsymbol{Y}$, into disjoint subsets. Each response Y_j is assigned to exactly one component, whereas a predictor X_r may belong to many components if it has a differential association with the outcomes in several components or may belong to no component if it is associated with no outcome.

21.3.1 Stochastic Partitioning Method

We denote a partition of the data into K components by

$$\mathcal{S}_1 \oplus \ldots \oplus \mathcal{S}_K = (\boldsymbol{X}_{I_1}, \boldsymbol{Y}_{J_1}) \oplus \ldots \oplus (\boldsymbol{X}_{I_K}, \boldsymbol{Y}_{J_K}) \tag{21.7}$$

where $I_k \subset \{1, \ldots, p\}$ with $0 \leq |I_k| = p_k \leq p$, $J_k \subset \{1, \ldots, q\}$ with $0 \leq |J_k| = q_k \leq q$ and $\sum_{k=1}^{K} q_k = q$. The $\oplus$ symbol is used to indicate that the union of variables is disjoint for the $\boldsymbol{Y}$ and not necessarily so for the $\boldsymbol{X}$ variables. The distribution for each element of the q_k outcomes Y_{J_k} of component $\mathcal{S}_k$ is assumed to be

$$Y_{ji}|\mathcal{S}_k \overset{iid}{\sim} \mathcal{N}(\alpha_j + \mu_k, \sigma_k^2), \qquad j \in J_k, \quad |J_k| = q_k, \quad i = 1, \ldots, n \tag{21.8}$$

where $\mu_k = g_k(X_{s_1}, \ldots, X_{s_{p_k}}) = \boldsymbol{X}_{I_k}\boldsymbol{\beta}_k$ captures the association between the p_k covariates and q_k outcomes in component $\mathcal{S}_k$.

We consider conjugate priors for the component parameters and exploit the conjugacy for computational efficiency by integrating them out. For the $(q_k + p_k)$-vector of regression coefficients $\boldsymbol{\theta}_k^T = (\alpha_{t_1}, \ldots, \alpha_{t_{q_k}}, \beta_{s_1}, \ldots, \beta_{s_{p_k}})$ we take

$$\boldsymbol{\theta}_k \sim \mathcal{N}(\boldsymbol{\theta}_{0k}, H_0\sigma_k^2) \tag{21.9}$$

where $H_0 = \text{diag}(h_0\mathbf{1}_{q_k}, h\mathbf{1}_{p_k})$ with $\mathbf{1}_n$ an n-vector with all components equal to one. We specify an inverse-gamma prior for the component variances

$$\sigma_k^2 \sim \mathcal{IG}(\sigma_0^2, \nu). \tag{21.10}$$

H_0 controls the strength of the prior information on the regression coefficients with larger values of h_0 and h corresponding to a wider spread around $\boldsymbol{\theta}_{0k}$. After integrating out the model parameters, the marginalized likelihood for a component with p_k covariates and q_k outcomes reduces to a multivariate t-distribution of dimension nq_k. Finally, to each configuration we assign a prior that penalizes large components with stronger penalty for smaller values of ρ $(0 < \rho < 1)$

$$\pi(\mathcal{S}_1 \oplus \ldots \oplus \mathcal{S}_K) \propto \prod_{k=1}^{K} \rho^{p_k \cdot q_k}. \tag{21.11}$$

We sample from the posterior probability distribution using an MCMC algorithm that moves in the space of possible configurations by merging two components or splitting one component into two. In order to ensure adequate mixing of the sampler among both the regressors and the response variables, we designed the MCMC algorithm to iterate between two steps, each focusing on splitting/merging different types of components:

(1) add/remove a regressor in a component: we randomly pick a component and we propose to either add to it a covariate uniformly selected among those not already in the component or remove one of its X variables;

(2) split/merge components both in the X and Y spaces: for the split move, we randomly pick a component and randomly split its outcomes into two new components. A subset of its X variables is assigned to the two new components to account for covariates shared by components while the remaining covariates are placed in one or the other component. For the merge move, two randomly selected components are combined into one.

We use the Metropolis acceptance function [6] to determine the acceptance probability of a proposed move. Furthermore, in order to increase the mixing of the sampler and limit the possibility for it to be trapped in local modes of the posterior density, we implement a parallel tempering scheme [3].

For posterior inference, we average over the configurations visited by the MCMC sampler and consider the $p \times q$ matrix of posterior probabilities of association between a covariate X_r and an outcome Y_j, the $p \times p$ matrix of pairwise posterior probabilities that the pair $(X_r, X_{r'})$ of covariates is selected in the same component, and the $q \times q$ matrix of posterior pairwise probabilities that the pair $(Y_j, Y_{j'})$ of outcomes is allocated to the same component.

21.4 Applications

In this section, we illustrate applications of the methods in two different areas. We apply the frequentist method to an ecology problem and the Bayesian method to a genomic study.

21.4.1 Modeling Dynamic Processes in Species-Rich Ecosystems

In ecological applications, understanding how environmental factors impact population dynamics is of primary importance for animal and plant species conservation. One challenge in modeling species-rich ecosystems, such as tropical rain forests, is their high biodiversity resulting in many species having limited number of samples, which hinders the development of species-specific models. This can be circumvented by identifying species with similar dynamics and modeling them at the cluster level. In addition, species respond differently to environmental stress or human pressure and there is interest in determining the cluster-specific relevant predictors.

We illustrate the multivariate penalized mixture GLM models on three different demographic processes—growth (Gaussian), mortality (Bernoulli), and recruitment (Poisson)—using data from the M'Baïki experimental site established in 1982 in the Central African Republic tropical rainforest in partnership with the Centre de coopération Internationale en Recherche Agronomique pour le Développement

(CIRAD). The site consists of permanent sample plots covering an area of 40 hectares with varying disturbances applied to different blocks: some plots were left as controls, others were selectively logged in 1984 by harvesting commercial trees with diameter at breast height greater than 80 cm, and a subset of the logged plots were further thinned in 1986 by poison girdling and by removing all lianas to increase light penetration. The M'Baïki site has been inventoried annually since 1982, thus providing a large amount of data to explore the dynamics of species demographic processes across a wide range of disturbances. The data consist of $q = 230$ species with more than $\sum_{j=1}^{q} n_j = 37{,}000$ trees monitored over a period of 18 years. Several environmental variables are considered. In order to fit the model and evaluate its predictive performance, we split the data into a training and a validation sets.

The models were fit for each demographic process using $K = 1, \ldots, 10$ groups. This was repeated ten times with different starting points for each K and the model with smallest ICL was chosen. Group structures for the different processes were successfully identified: six groups were uncovered for the growth process nested within four recruitment groups, and there were three mortality groups. The crossing of these classifications resulted in 15 non-empty clusters, each containing between one and 24 species. As shown in Fig. 21.1, the species groups plotted along

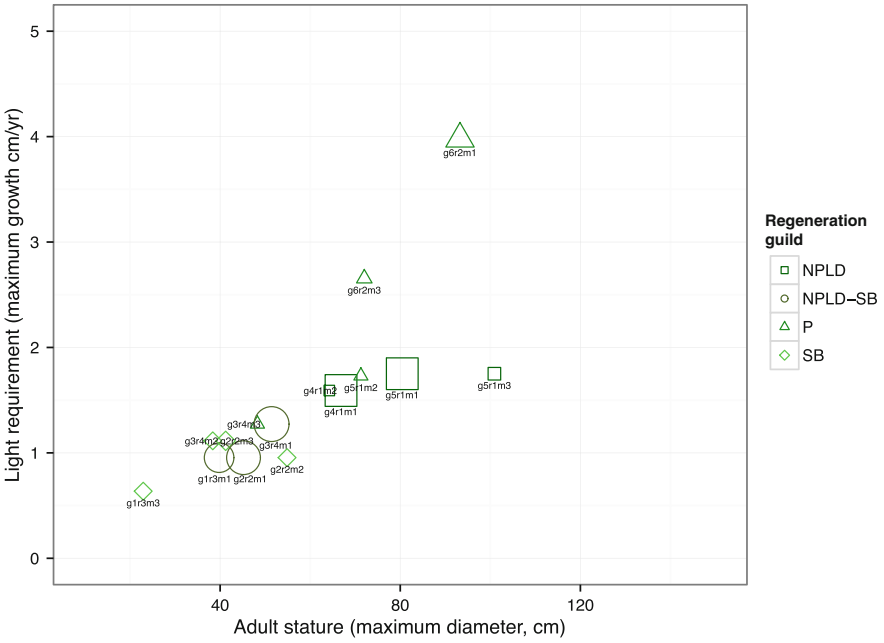


Fig. 21.1 Ecologic application: projection of the uncovered species clusters on the two axes corresponding to the maximum growth rate and the maximum diameter. The labels $g_x r_y m_z$ correspond to the identified species groups. Each *symbol* corresponds to the dominant regeneration guild of each group and its size is proportional to the number of species in the group. *NPLD* nonpioneer light demander, *SB* shade bearer, *P* pioneer

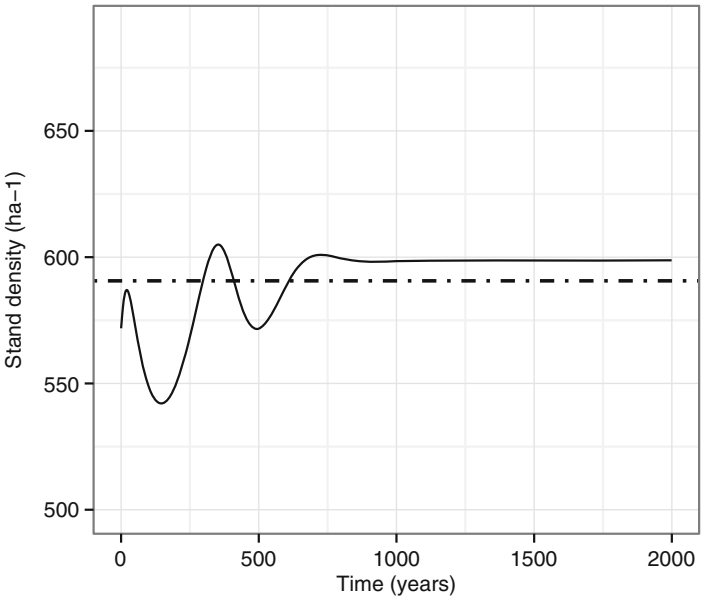


Fig. 21.2 Ecologic application: density (number of trees per hectare) in the simulated/predicted forest (*solid line*) and observed stand in 2012 (*dashed line*) for the validation data

their maximum growth rate and their maximum diameter represent biologically meaningful groupings in terms of regeneration guild. Furthermore, the predicted states for the validation set matched the observed measurements. Figure 21.2 shows the predicted asymptotic tree density (solid line) and the observed stand in 2012 (dashed line). The model was also successful in predicting the reconstitution rate of the basal area after a simulated disturbance designed to replicate the one realized in 1984 at M'Baïki in terms of lost basal area. The predicted dynamics following a 28-year wait after disturbance of the asymptotic state in the logged plots of the validation data matched the observed dynamics between 1982 and 2012 (Fig. 21.3).

21.4.2 Integrative Genomic Analysis

In genomic applications, there is a growing interest in relating data sets from various genome-wide technologies to better understand molecular processes underlying various phenotypes. Here, we illustrate the method using genotype and gene expression data collected on the same individuals. The goal is to identify genes with similar expression patterns and determine DNA sequence variations that modulate the clustered expression profiles. Genes with similar expression patterns are believed to share similar regulatory mechanisms. Thus, co-expressed genes would be co-regulated and share the same regression relationships, whereas genes in different clusters would have different regression models.

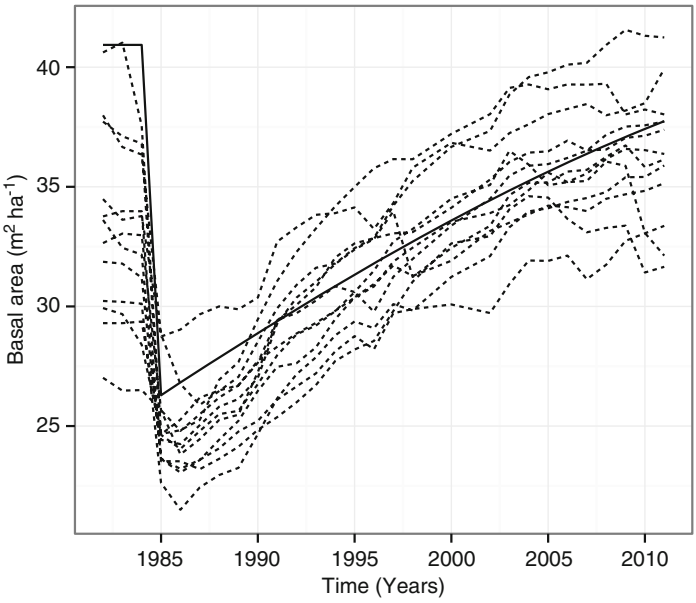


Fig. 21.3 Ecologic application: predicted dynamics of the basal area after disturbance of the asymptotic state depicted in Fig. 21.2 (*dotted lines*) and observed dynamics (*solid line*) between 1982 and 2012 in the validation set

We illustrate the stochastic partitioning method on an expression quantitative trait loci (eQTL) application, where the goal is to relate gene expression and genotype data. We used the data presented in Morley et al. [8], which consist of single nucleotide polymorphism (SNP) markers and gene expression levels collected on $n = 56$ unrelated individuals from 14 CEPH (Centre d'Etude du Polymorphisme Humain) families. We removed markers with minor allele frequencies less than 5 % and missing genotypes in more than a quarter of the individuals, leaving $p = 2455$ SNPs for analysis. For the gene expression data, $q = 3554$ of the most variable probe sets were considered for analysis. The MCMC sampler was run for 50 million iterations with ten steps for the parallel tempering implementation. For posterior inference, we focused on the last 15 million iterations and subsampled configurations every 20,000 scans. At each MCMC iteration, the stochastic partitioning method yields a partition of the data in the form of Eq. (21.7). We average over these visited configurations and estimate the pairwise posterior probability of two gene expression phenotypes being allocated to the same component by the proportion of configurations that have these outcomes in the same component. Similarly, to estimate the posterior probability of association between a SNP marker and a gene expression outcome, we calculate the proportion of visited configurations having the corresponding variables in the same component.

The sampler mixed well over the number of components and visited models with 250–270 components with stronger support for configurations with around

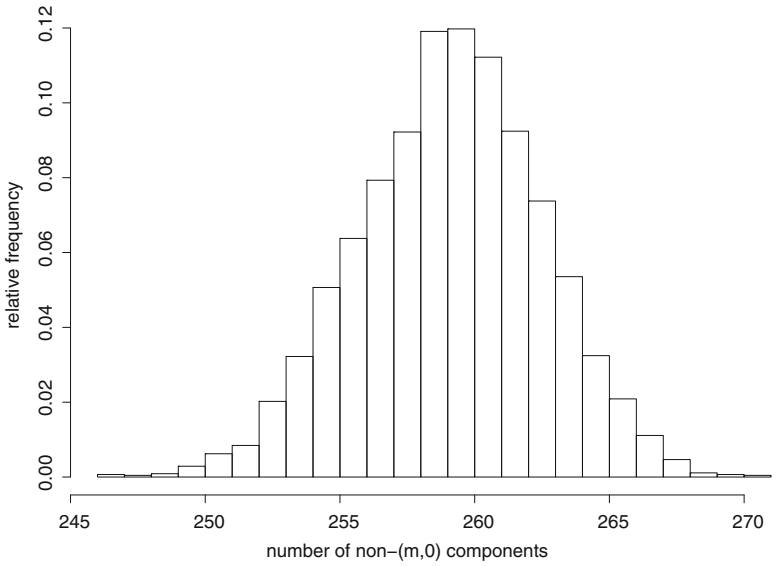


Fig. 21.4 Genomic application: histogram of number of components in visited configurations

260 components (Fig. 21.4). We computed the 2455 × 3554 matrix of posterior probabilities that each SNP be associated with each probe set. Several SNP markers appeared to be associated with the mRNA transcript abundance of multiple genes. There were 363 SNP markers associated to gene expression changes with at least one pairwise posterior probability greater than 0.7. Table 21.1 lists examples of markers that exhibit strong association with a few expression phenotypes. In some cases, such as the set of genes from chromosome X found to be related to rs1859674, changes in expression levels can be explained by the variation of markers located on the same chromosome. Markers can also be associated with genes mapping to a different chromosome but localized in a specific region. For instance, rs533569 located on chromosome 11 appears to be related to several genes found in the small histone gene cluster on chromosome 6p21.3. This is also the case for marker rs1429309 located on chromosome 2, which is strongly associated with several genes mapping to chromosome Yq11. Rather than focusing on specific markers, one could consider particular gene expression phenotypes and examine the corresponding columns of the $p \times q$ matrix of marginal probabilities to locate its related markers. Figure 21.5 gives the marginal posterior probability plots for some of the genes listed in Table 21.1. For instance, HDHD1A which maps to chromosome X appears to be most strongly associated with a marker located in the same region. Gene expression changes in HIST1H3H, on the other hand, are found to be related to variations of several markers scattered on various chromosomes.

The model captures correlated outcomes through their dependence on the same set of regressors. Genes with similar expression profiles can thus be located based on their association with the same set of markers. As an alternative, one could

Table 21.1 Genomic application: example of markers and associated gene expressions with marginal posterior probability greater than 0.7

SNP		Gene expression	
Marker	Location (Mbp)	Name	Location
rs1859674	Chr X (116.29)	HDHD1A	Xp22.32
		UTX	Xp11.2
		U2AF1L2	Xp22.1
		XIST	Xq13.2
rs1429309	Chr 2 (57.18)	RPS4Y1	Yp11.3
		EIF1AY	Yq11.222
		DDX3Y	Yq11
		USP9Y	Yq11.2
		SMCY	Yq11
rs127503	Chr 6 (108.59)	SLC4A2	7q35
		CDK10	16q24
		LCAT	16q22.1
		CYP4F12	19p13.1
rs533569	Chr 11 (93.70)	HIST1H3H	6p21.3
		HIST1H2BF	6p21.3
		HIST1H2BE	6p21.3
		H2BFS	21q22.3
		HIST1H2BC	6p21.3
		HIST1H2AC	6p21.3

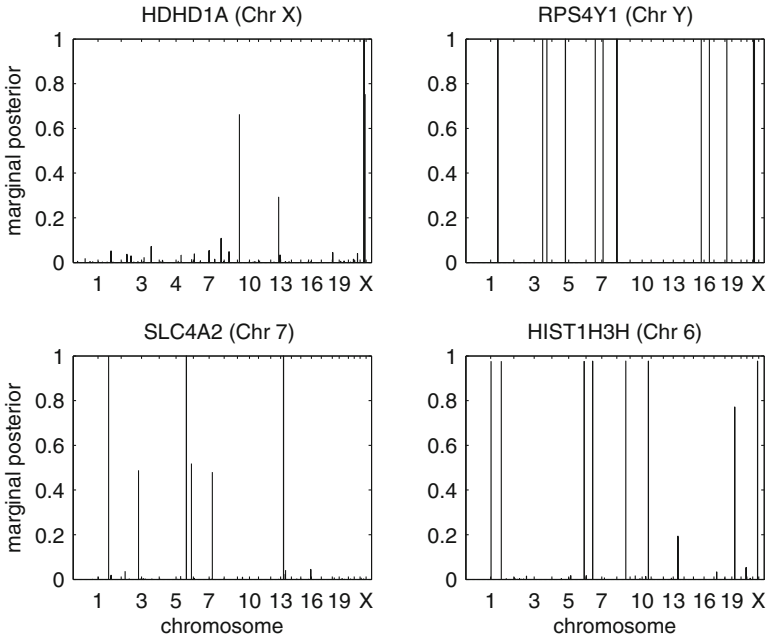


Fig. 21.5 Genomic application: marginal posterior probabilities of association across markers for four expression phenotypes

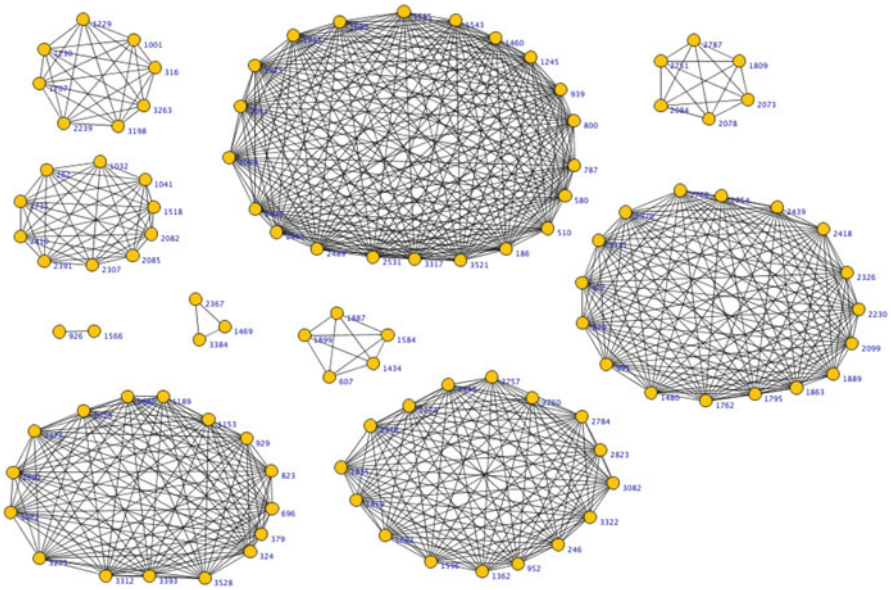


Fig. 21.6 Genomic application: network representation of some gene expressions with pairwise posterior probability of occurring in same components $P(\boldsymbol{y}_j = \boldsymbol{y}_{j'}) \geq 0.7$

examine the 3554×3554 matrix of pairwise posterior probabilities that two probe sets be allocated to the same component and view the entries as similarity metrics for grouping expression phenotypes. One possible display of the results is in the form of a network representation, where gene expressions define the nodes and the pairwise probabilities correspond to edge weights. Figure 21.6 shows a sample of expression phenotypes that occur in the same component with pairwise posterior probability greater than 0.7. The expression levels across the 56 individuals for the genes in the smaller cliques are presented in Fig. 21.7. We note that genes with similar expression profiles are successfully identified in separate groups. For example, the set of genes in Fig. 21.7a corresponds to a cluster of genes mapping to chromosome Y. As expected, they have high transcript abundance among males and lower expression in females. The genes in Fig. 21.7b belong to the small histone gene cluster mentioned above.

21.5 Conclusion

Mixtures of multivariate regressions combined with variable selection provide a flexible method to model complex data by uncovering cluster structures and identifying relevant covariates for each group. In this chapter, we have presented frequentist and Bayesian methods to fit these models in a unified framework. The

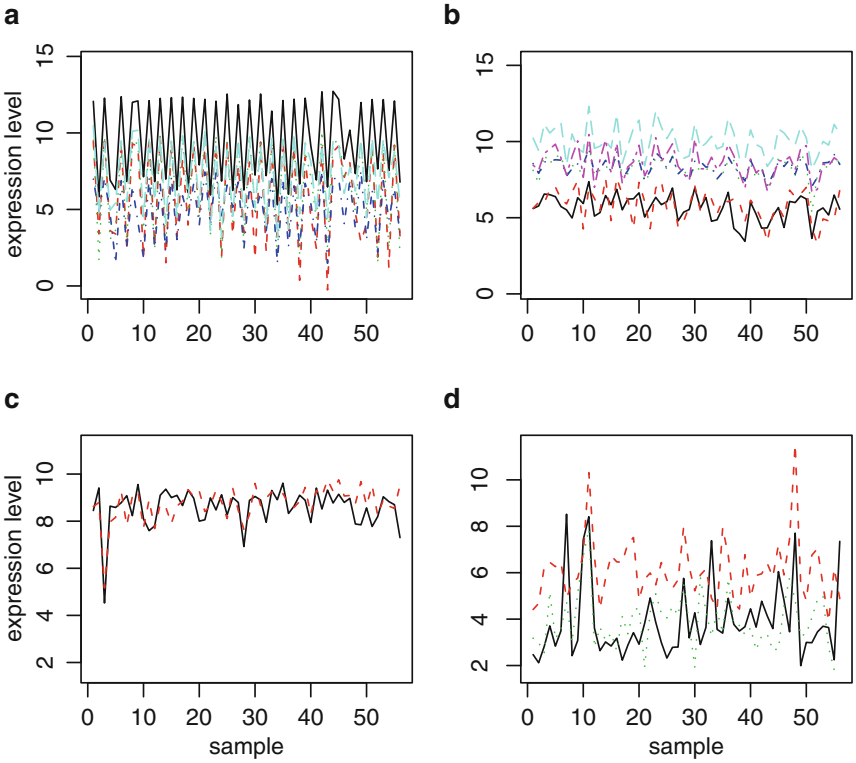


Fig. 21.7 Genomic application: expression profiles for genes forming smaller cliques in the network representation of Fig. 21.6. (**a**) RPS4Y1-chrYp11.3; EIF1AY-chrYq11.222; DDX3Y chrYq11; USP9Y-chrYq11.2; SMCY-chrYq11. (**b**) HIST1H2BC-chr6p21.3; HIST1H3Hchr6p22-p21.3; HIST1H2BF-chr6p21.3; HIST1H2BE-chr6p21.3; H2BFS-chr21q22.3; HIST1H2AC-chr6p21.3. (**c**) GBP2-chr1p22.2; MNDA-chr1q22. (**d**) EYA2-chr20q13.1; EEF1A2-chr20q13.3; GPR25-chr1q32.

frequentist method has the advantage of being computationally less intensive and not requiring the specification of tuning parameters. The Bayesian method, on the other hand, is computationally more expensive and requires the elicitation of prior distributions, which guide the exploration of the posterior model space. However, the Bayesian method has the advantage of visiting a larger portion of the configuration space and providing a posterior distribution over the entire space of partitions. This allows the uncertainties in the cluster structure and in the association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ variables to be captured.

There are a number of possible future directions to extend these models to overcome some assumptions. In the models we have described, the correlation between outcomes in the same component is captured through their identical dependence on the same predictors. Conditional on the component-specific relevant covariates, the outcomes in a component are assumed to be independent. However,

it is unlikely that all important covariates are considered and thus there may still be dependence between the outcomes that needs to be taken into account. Another extension would be to model temporal dependence explicitly by including random effects when dealing with observations that are measured over time as in the ecological application we considered. When working with count data, zero-inflated distributions are often used to account for the large number of zeroes than can be accommodated with the Poisson distribution. The challenge of using zero-inflated models in the context of model-based clustering is the complexity of nesting two levels of mixtures: one corresponding to the mixture of a point mass at zero and a Poisson distribution and the other corresponding to the mixture of distributions used to identify groups of objects. In the stochastic partitioning method, the marginalization over the model parameters provides a substantial gain in computational speed and efficiency. If one chooses to use non-conjugate priors or considers non-Gaussian outcomes, the model parameters would need to be updated in the MCMC procedure and appropriate reallocations would need to be devised at each split/merge moves. Finally, the Bayesian approach presents challenges in effectively summarizing the results. In the genomic application, we used pairwise posterior probabilities to identify SNPs associated with gene expressions and to locate expression phenotypes allocated to the same component. An alternative would be to report the maximum *a posteriori* configuration, which provides important information on higher order relationships between variables, but it neglects additional information coming from potentially very different configurations with similar posterior probabilities. Yet another possibility would be to consider the most likely models by locating different modes of the posterior distribution.

References

1. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 719–725 (2000)
2. George, E., McCulloch, R.: Approaches for Bayesian variable selection. Stat. Sin. **7**, 339–373 (1997)
3. Geyer, C.: Markov chain Monte Carlo maximum likelihood. In: Keramigas, E. (ed.) Computing Science and Statistics, pp. 156–163. Interface Foundation, Fairfax (1991)
4. Gupta, M., Ibrahim, J.G.: Variable selection in mixture modeling for the discovery of gene regulatory networks. J. Am. Stat. Assoc. **102**, 867–880 (2007)
5. Khalili, A., Chen, J.: Variable selection in finite mixture of regression models. J. Am. Stat. Assoc. **102**, 1025–1038 (2007)
6. Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., Teller, E.: Equations of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087–1091 (1953)
7. Monni, S., Tadesse, M.G.: A stochastic partitioning method to associate high-dimensional responses and covariates (with discussion). Bayesian Anal. **4**, 413–464 (2009)
8. Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., Spielman, R.S., Cheung, V.G.: Genetic analysis of genome-wide variation in human gene expression. Nature **430**, 743–747 (2004)

9. Mortier, F., Ouédraogo, D.-Y., Claeys, F., Tadesse, M.G., Cornu, G., Baya, F., Benedet, F., Freycon, V., Gourlet-Fleury, S., Picard, N.: Mixture of inhomogeneous matrix models for species-rich ecosystems. Environmetrics **26**, 39–51 (2015)
10. Städler, N., Bühlmann, P., van de Geer, S.: $\ell 1$-penalization for mixture regression models. Test **19**, 209–256 (2010)
11. Zou, H.: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418–1429 (2006)

Chapter 22
Coalescing Complex Planar Stationary Points

Loïc Teyssier

Abstract Among all bifurcation behaviors of analytic parametric families of real planar vector fields, those that stand out most prominently are confluences of distinct stationary points. The qualitative change is so drastic that in some classes of families (e.g., fold-like bifurcations) the stationary points leave the real plane altogether and slip into the complex plane. Although they disappear from the real domain they continue to organize the dynamics, and studying *complex* planar vector fields becomes a necessity even for real bifurcations. Our main concern is to describe *à la* Martinet-Ramis the analytical classification of generic holomorphic families unfolding a saddle-node vector field, and to relate this classification both to the dynamics of individual members of the family and to analytic properties of the saddle-node. For instance the problem of the existence of an analytic center-manifold for the saddle-node is characterized in terms of persistence (as the parameter tends to the bifurcation value) of heteroclinic connections between stationary points. We emphasize the geometric aspect of the classification. Complex trajectories are connected real surfaces allowing for richer geometric constructions as compared to 1-dimensional real trajectories. The trajectories are split by a finite collection of open "fibred squid sectors," attached by spirals to stationary points within their adherence. The sectors are carved in such a way that one can construct an analytic and bounded conjugacy between the vector field and its formal normal form. The invariants of classification are obtained as transition maps of overlapping such normalization charts. Since we can perform this sectorial normalization analytically in the parameter, by restricting its values to "cells" covering the parameter space minus the bifurcation value, the resulting finite collection of functional invariants is analytic on parameter cells and continuous on their adherence. In that sense it "unfolds" Martinet-Ramis invariant of the saddle-node. The inverse problem (or realization) is addressed in the case of a persistent heteroclinic connections and provides unique normal forms (universal family for the analytic classification). We particularly show that in general the invariant cannot depend holomorphically on the parameter over a full neighborhood of the bifurcation value.

L. Teyssier (✉)
U.F.R. de Mathématiques et d'Informatique & Institut de Recherche Mathématique Avancée, Université de Strasbourg, 7 rue René-Descartes, 67084 Strasbourg Cedex, France
e-mail: teyssier@math.unistra.fr

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_22

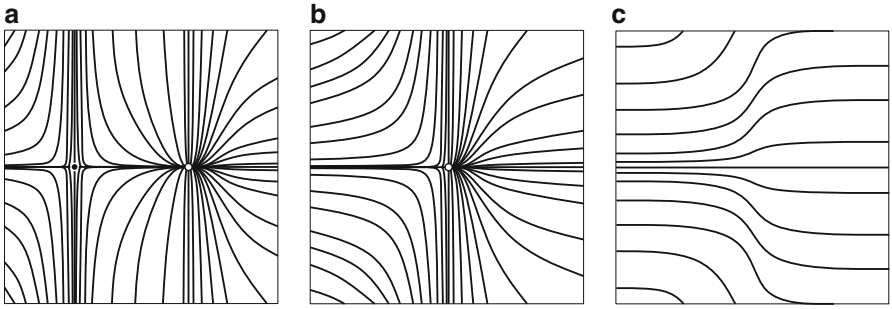


Fig. 22.1 Typical members of the simplest saddle-node bifurcation. (**a**) $\lambda < 0$; (**b**) $\lambda = 0$; (**c**) $\lambda > 0$

Keywords Saddle-node bifurcation • Normal forms • Holomorphic vector fields • Unfolding of singularities • Modulus space • Inverse problem

Among all bifurcation behaviors of parametric families of real planar vector fields $Z_\bullet = (Z_\lambda)_{\lambda\in\Lambda}$, those that stand out most prominently are confluences of distinct stationary points. The qualitative change is so drastic that in some classes of families (e.g., fold-like bifurcations) the stationary points generically annihilate each other in the process (Sotomayor's theorem) (Fig. 22.1).

The simplest example of such a behavior, an instance of saddle-node bifurcation, is the polynomial family $\left(\mathrm{X}_\lambda^\infty\right)_{\lambda\in\mathbb{R}}$ given in the canonical basis of $\mathbb{R}^2$ by

$$\mathrm{X}_\lambda^\infty(x,y) := \begin{bmatrix} x^2+\lambda \\ y \end{bmatrix} . \tag{22.1}$$

The bifurcation value occurs at $\lambda = 0$: for negative λ, the system has two stationary points located at $\left(\pm\sqrt{-\lambda}, 0\right)$ which collide as λ reaches 0, while none remain for $\lambda > 0$. The stationary points have left the real plane, true enough, but only to slip into the complex domain. Let us elaborate a bit on this observation in order to motivate the need for complexifying the whole setting, even in the context of real dynamics.

The trajectories $t \mapsto (x(t), y(t))$ of $\mathrm{X}_\lambda^\infty$ appear naturally as solutions of the autonomous flow-system of $\mathrm{X}_\lambda^\infty$:

$$\begin{cases} \dot{x}(t) &= x(t)^2 + \lambda \\ \dot{y}(t) &= y(t) \end{cases}$$

and can be implicitly expressed by solving the associated non-autonomous differential equation. This equation is obtained by eliminating the time in the

flow-system using the rule $\frac{\dot{y}}{\dot{x}} = \frac{\mathrm{d}y}{\mathrm{d}x}$:

$$\left(x^2 + \lambda\right) y'(x) = y(x) \ .$$

Separation of variables yields multivalued complex solutions

$$y_\lambda \ : \ z \longmapsto c \left(\frac{z - \mathrm{i}\sqrt{\lambda}}{z + \mathrm{i}\sqrt{\lambda}} \right)^{\frac{1}{2\mathrm{i}\sqrt{\lambda}}} , \quad c \in \mathbb{C} \, . \tag{22.2}$$

On the one hand, if $\lambda < 0$ real solutions are given on appropriate intervals by

$$y_\lambda \ : \ x \longmapsto c \left| \frac{x - \sqrt{-\lambda}}{x + \sqrt{-\lambda}} \right|^{\frac{1}{2\sqrt{-\lambda}}} , \ c \in \mathbb{R} \, , \tag{22.3}$$

from which we deduce that $\left(-\sqrt{-\lambda}, 0\right)$ is a saddle-point and $\left(\sqrt{-\lambda}, 0\right)$ a node-point. On the other hand, for $\lambda > 0$ we have

$$y_\lambda \ : \ x \longmapsto c \exp\left(\frac{1}{\sqrt{\lambda}} \arctan \frac{x}{\sqrt{\lambda}} \right) , \ c \in \mathbb{R} \, . \tag{22.4}$$

The latter is a perfectly honest real-analytic function on $\mathbb{R}$. One might wonder why, despite the fact of being so regular a function, its Taylor expansion at 0 does not have infinite convergence radius instead of $\sqrt{\lambda}$. One can *explain* the discrepancy by, say, direct use of Cauchy–Hadamard formula, although one cannot *understand* its source without noticing the imaginary singularities $\pm\mathrm{i}\sqrt{\lambda}$ quietly sitting on the boundary of the disk of convergence. Also it is hard to understand why, when playing the movie backwards starting form positive values of λ and reaching negative ones, a stationary point somehow pops out of nowhere. One can see the singularity coming only when looking along the imaginary axis.

At a less commonplace level, when $\lambda < 0$ both stationary points organize the dynamics of $\mathrm{X}_\lambda^\infty$ and there is no reason why they should stop when $\lambda > 0$, or even when λ is not real, and we will present how.

22.1 Class of Parametric Families

The present chapter deals with germs of a parametric family of vector fields in the complex plane, enjoying a saddle-node bifurcation of codimension 1 and corresponding to first-order *non-linear* differential equations. The basic examples addressed by the text are affine families perturbing $\mathrm{X}_\bullet^\infty$

$$\mathrm{X}_\bullet \,:\, \mathrm{X}_\lambda\,(x,y) := \begin{bmatrix} x^2+\lambda \\ y-\left(x^2+\lambda\right)a_\lambda\,(x) \end{bmatrix}\,, \tag{22.5}$$

where $(\lambda, x) \longmapsto a_\lambda\,(x)$ is a given analytic function near the origin of $\mathbb{C}^2$. Although for such elementary families all computations can be performed explicitly (variation of constant), some natural questions and non-trivial answers arise already in this case-study. Generalizing the constructions and objects introduced in that simple situation to arbitrary bifurcation-preserving *analytic* perturbations of the model family $\mathrm{X}_\bullet^\infty$ is the main concern of the rest of the text.

Reducing the setting to analytic parametric families may seem rather restrictive. Yet the geometric approach we present here could be inherited by less regular situations, or could give insights as to where sources of peculiar behavior may lie. On the other end of the argument, the obvious added benefit stemming from this restriction is the rigidity of holomorphic functions and diffeomorphisms of complex (compact) manifolds. Also the analytic class comprises polynomial vector fields, of special interest for planar vector fields, e.g., regarding Hilbert's 16th problem on the number/position of limit cycles, or Poincaré's problem on the existence of rational first integrals.

22.2 Scope of the Study

In the sequel we investigate the links between local dynamics on the one hand, local classification (i.e., up to local changes of analytic coordinates and parameters) on the other hand, while at the same time hinting at how they can help measuring divergence of some class of "summable" power series. We particularly explain the role of complex geometry and analysis in understanding saddle-node bifurcations. We wish to underline that the two objects Z_0 and $(Z_\lambda)_{\lambda\neq 0}$ are intertwined, as dynamical properties for one can be deduced from studying objects attached to the other and vice versa.

Rousseau has pioneered the classification of some non-linear (discrete or continuous) dynamical systems having a saddle-node bifurcation [19, 24–28]. She has also contributed to the study of families of vector fields corresponding to linear differential systems in finite-dimensional complex linear spaces, with Fuchsian singularities merging to an irregular singularity [12, 13, 15–17]. The Stokes matrix of the irregular system is recovered as the limit of well-chosen monodromy matrices of the Fuchsian systems. The linear/non-linear and discrete/continuous settings all share the same basic idea, which can be summarized as the following recipe:

- find a finite decomposition of both parameter space and dependent-variable space in "sectors" over which the family is conjugate to some known, simple normal form;
- form the classification invariants as transition maps between overlapping normalization sectors.

We do not wish to emphasize too much the link between local orbital classification of $Z_\bullet$ and local classification of its strong holonomy $\mathfrak{h}_\bullet$, the family of holomorphic first-return map of Z_λ on a fixed horizontal disc which crosses $\{x^2 + \lambda = 0\}$. The connection is very clearly explained by Rousseau, for instance, in [24] for saddle-node bifurcations, or again in [26] for deformations of a resonant saddle stationary point. Although both objects encode somehow the same dynamics, and are classified by the same invariants under local analytic equivalence and change of parameters, we take advantage of the extra dimension the complex plane $\mathbb{C}^2$ offers to deploy more geometrical constructions à la Martinet–Ramis [21, 31]. Instead of simply deducing the classification of vector fields from that of holonomies, which would frankly spoil all the fun, the present text is focused on building objects specifically from the continuous nature of the dynamics of Z_λ. Although both moduli spaces end up with the same presentation, some formulations for vector fields yield different characterizations of, e.g., the "compatibility condition" as compared to holonomies [27].

A by-product of that approach is an explicit family of normal forms for bifurcations $Z_\bullet$ having persistent heteroclinic connections, generalizing [30] to the case $\lambda \neq 0$ as done in [29]. There is as yet no such known explicit universal family for holonomies $\mathfrak{h}_\bullet$ (not even for $\mathfrak{h}_0$).

22.3 Contents Description

The text begins with two preliminary sections devoted to covering basic examples, objects and tools, as well as fixing notations.

- The example of affine families (22.5) is presented in Sect. 22.4. The thread of the exposition is the link existing between the (lack of) analytic center manifold of X_0 and the (lack of) persistence of heteroclinic connections in the family $X_\bullet$. In the process of revealing this bond through the use of elementary complex analysis, we perform the analytical classification of all affine families and present a collection of normal forms.
- The other sections are framed in a geometric setting, with its own standard terminology. We present in Sect. 22.5 basic objects attached to singular holomorphic vector fields: directional derivative, flow, change of coordinates, and most importantly singular foliations, first integrals, leaves space, and normal forms. We recall related basic results of differential geometry. Readers familiar with these concepts should skim briefly through this section mainly to fix notations.

The next six sections form a survey on formal and analytical classification of saddle-node vector field bifurcations. The choice has been made to focus mainly on precise constructions, while providing sketches of proof whenever doing so helps the exposition. The missing technical details are to be found mostly in [28, 29].

- Section 22.6 is devoted to an introductory text, giving a brief historical overview of the emergence of moduli à la Martinet–Ramis for saddle-node vector fields and their deformations. The dynamical nature of these invariants is explained

from their very construction as transition maps of a *rigid* analytic atlas of the corresponding leaves space.

- We present a more detailed account of the formal and local classifications in Sect. 22.7, where the main theorems are stated and the structure of the construction is presented. Each one of subsequent sections develops a particular aspect.
  - The formal normalization is performed in Sect. 22.8 by reducing the problem to solving a couple of well-chosen cohomological equations.
  - Sections 22.9 and 22.10 contain the precise construction of the normalization sectors in parameter-space and dependent-variable-space, respectively. Ascertaining the rigidity of the corresponding leaves space atlas is literally what shapes the normalization sectors.
  - Section 22.11 contains the material needed to perform the sectorial normalization, by solving on such sectors the cohomological equations. The classification theorem is finally proven: to each family $Z_\bullet = (Z_\lambda)_\lambda$ corresponds a functional invariant $\mathfrak{m}(Z_\bullet)$, and the fiber of $\mathfrak{m}$ over $\mathfrak{m}(Z_\bullet)$ consists precisely of the conjugacy class of Z. The general inverse problem (to determine the range of $\mathfrak{m}$) is still open.

The rest of the chapter is mostly concerned with the inverse problem and its dynamical ramifications. The last two sections involve more recent (and perhaps more difficult) material, coming with full proofs.

- Section 22.12 is concerned with the dynamical interpretation of the invariant of classification $\mathfrak{m}(Z_\bullet)$, from which is formulated the "orbital compatibility condition". This condition is expected to solve the orbital inverse problem. The fact that (the orbital part of) $\mathfrak{m}$ is not onto the natural candidate is established. The proof is based on the characterization of those $Z_\bullet$ for which $\mathfrak{m}(Z_\bullet)$ is analytic in the parameter, which turns out to seldom happen.
- Section 22.13 is concerned by partial answers to the inverse problem. First we formulate the "temporal compatibility condition" and prove it completely characterizes the range of $\mathfrak{m}$ for given orbital part, hence describing the moduli space of saddle-node bifurcations inducing the same given bifurcation of foliations. At last we provide analytical normal forms in the case of persistent heteroclinic connections. The combination of both temporal and orbital compatibility conditions is proved to solve the inverse problem completely in that (non-generic) case.

22.4 Affine Saddle-Node Bifurcations

The study of affine saddle-node vector fields was initiated in the second half of the nineteenth century by Bouquet and Briot [3], starting with a collection of examples of invariant manifolds existing at a formal level but not at an analytic one, generalizing the famous behavior displayed by Euler's differential equation

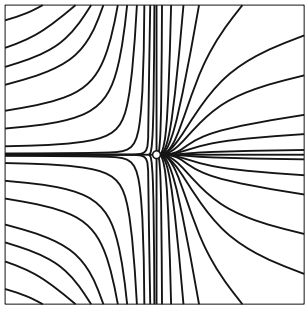


Fig. 22.2 Solutions of $x^2y' = y$

$x^2y' = y - x$. Their first significant result is the existence of a formal **weak separatrix** for X_0, that is an invariant formal curve $\{y = \hat{s}(x)\}$ with $\hat{s} \in \mathbb{C}[[x]]$. They obtained an explicit criterion for convergence of $\hat{s}$ in terms of the Taylor coefficients of a_0, which we recover in Proposition 3 after a study aimed at understanding how the trajectories of X_λ, $\lambda \neq 0$, evolve into those of X_0 (Fig. 22.2).

We begin our study of the affine collection Affine (1), whose members are given by (22.5), by presenting the easiest instance $a_\bullet := 0$. Standard results describing the regularity of parametric solutions state that solutions y_λ in (22.3) and (22.4) converge to corresponding solutions for $\lambda = 0$

$$x^2y'(x) = y(x) \tag{22.6}$$

$$y_0 : x \longmapsto c\exp\left(-\frac{1}{x}\right) \quad , \ c \in \mathbb{C}$$

uniformly on compact subsets of $\mathbb{R}_{\neq 0}$ as $\lambda \to 0$ (it suffices to wait until the stationary points have left the compact set). Now, can we say something about the convergence near 0? Obviously the question only makes sense for families of solutions bounded near 0 as $\lambda \to 0$. For $\lambda = 0$ the limiting objects are center manifolds of the saddle-node stationary point $(0,0)$ of X_0^∞. As a real vector field X_0^∞ has infinitely many center manifolds passing through $(0,0)$, each one given by the graph $\{y = s(x)\}$ of the smooth (meaning C^∞) function

$$\begin{aligned} s &: \mathbb{R} \longrightarrow \mathbb{R} \\ x \leq 0 &\longmapsto 0 \\ x > 0 &\longmapsto c\exp\left(-\frac{1}{x}\right) \end{aligned}$$

for arbitrary $c \in \mathbb{R}$. Those are the only bounded solutions of (22.6) at 0. Only one of them is analytic there, namely $\mathcal{S}_0 := \{y = 0\}$, all others being non-zero flat functions. This property identifies uniquely a distinguished center manifold, called the **weak separatrix** of X_0^∞, with the most regular dynamics. The weak separatrix is the limiting curve of the family $(\mathcal{S}_\lambda)_\lambda$ collecting the only smooth integral curve

connecting both stationary points $\left(\pm\sqrt{-\lambda},0\right)$ for $\lambda < 0$. In this simple situation the only such **heteroclinic** integral curve is $\mathcal{S}_\lambda = \{y = 0\}$, since y_λ in (22.3) is not of class C^{r+1} at the node, $r := \left\lceil \frac{1}{2\sqrt{-\lambda}} \right\rceil$, save for $c = 0$.

Consider next a quadratic perturbation of $\mathrm{X}_\bullet^\infty$, the Euler family

$$\mathrm{E}_\lambda\,(x,y) := \begin{bmatrix} x^2+\lambda \\ y-\left(x^2+\lambda\right) \end{bmatrix} \tag{22.7}$$

whose stationary points are again located at $(\pm s, 0)$ where, for the sake of simplicity, we set:

$$s := \sqrt{-\lambda}\,.$$

This is a special member of Affine (1) obtained by setting $a_\bullet := 1$, yet we are to prove that together with $\mathrm{X}_\bullet^\infty$ they somehow span all possible behaviors for members of the whole collection.

For $\lambda = 0$ infinitely many smooth center manifolds persist through $(0,0)$, given by the graphs of

$$\begin{aligned} \mathfrak{s}_0 \;:\; \mathbb{R} &\longrightarrow \mathbb{R} \\ x<0 &\longmapsto \exp\left(-\frac{1}{x}\right)\int_x^0 \exp\left(\frac{1}{u}\right)\mathrm{d}u \\ 0 &\longmapsto 0 \\ x>0 &\longmapsto \exp\left(-\frac{1}{x}\right)\left(c+\int_x^1 \exp\left(\frac{1}{u}\right)\mathrm{d}u\right)\,. \end{aligned} \tag{22.8}$$

A standard calculus exercise consists in checking for the smoothness of $\mathfrak{s}_0$. Yet none of these functions can be analytic, as if one were it would possess a convergent Taylor series $\hat{\mathfrak{s}}$ at 0 solving

$$x^2\hat{\mathfrak{s}}'\,(x) = \hat{\mathfrak{s}}\,(x) - x^2\,. \tag{22.9}$$

A straightforward computation yields the unique formal power series

$$\hat{\mathfrak{s}}\,(x) = x\sum_{n\geq 0} n!x^{n+1}\,, \tag{22.10}$$

which has null radius of convergence. We say in that case that we encounter a **divergent weak separatrix**. It is worth mentioning that the Taylor expansion of each $\mathfrak{s}_0$ at 0 is $\hat{\mathfrak{s}}$.

Here we cannot distinguish a preferred center manifold in the class of analytic objects at $(0,0)$. Although the divergence of the weak separatrix can be explained

computationally for the Euler family, the generic perturbation $X_\bullet \in \mathrm{Affine}\,(1)$ is impossible to deal with this way since no reasonable closed-form formulas for the coefficients of $\hat{\mathfrak{s}}$ exist in general. Even so the basic formal approach, computing coefficients of $\hat{\mathfrak{s}}$ one after the other, cannot prove nor disprove the power series convergence in finite time. We propose a dynamical approach instead to trace back the source of the divergence (Theorem 1), which leads to the semi-decidability of the convergence of $\hat{\mathfrak{s}}$: there exists an algorithm taking a "computable" $a_\bullet$ as input and stopping in finite time if and only if $\hat{\mathfrak{s}}$ diverges. The key is to check whether the complex contour integral

$$\varphi_0^{\mathrm{n}} := \frac{1}{2\mathrm{i}\pi} \oint_{r\mathbb{S}^1} a_0\,(z) \left(\frac{z+s}{z-s}\right)^{\frac{1}{2s}} \mathrm{d}z \in \mathbb{C}$$

vanishes (meaning convergence), for $r > 0$ small enough. This viewpoint also allows us to find a complete collection of normal forms (Theorem 2).

When $\lambda < 0$ write $\mathcal{S}_\lambda^-$ the (analytic) stable manifold of the saddle-point located at $(-s, 0)$ and, when it exists, $\mathcal{S}_\lambda^+$ the (analytic) unstable manifold of the node-point at $(s, 0)$. What happens in the Euler family is that no heteroclinic connection between stationary points takes place: $\mathcal{S}_\lambda^-$ does not coincide with $\mathcal{S}_\lambda^+$. We aim at establishing this property has a predominant bearing on the convergence of the weak separatrix.

Theorem 1. *Consider a family* $\mathrm{X}_\bullet \in \mathrm{Affine}\,(1)$ *as in* (22.5)*. The implications (1)⇒(2)⇒(3) hold, and if moreover* $\frac{\partial a_\lambda}{\partial \lambda} = 0$ *then (3)⇒(1).*

1. *The vector field* X_λ *has a heteroclinic connection for all* $\lambda < 0$ *sufficiently close to* 0*.*
2. *The vector field* X_λ *has a heteroclinic connection for values of* $\lambda < 0$ *accumulating on* 0*.*
3. *The vector field* X_0 *admits a convergent weak separatrix (that is, an analytic center manifold).*

Remark 1.

1. In each item of the theorem the corresponding property is equivalent to the existence of an open interval $I \ni 0$ such that the differential equation

$$\left(x^2 + \lambda\right) y'\,(x) = y\,(x) - \left(x^2 + \lambda\right) a_\lambda\,(x) \tag{22.11}$$

   admits a solution analytic on I, for every corresponding values of λ. The solution is necessarily unique.
2. The practical usefulness of the theorem is by contraposition: if we happen to know that X_0 has a divergent weak separatrix, then *any* unfolding $\mathrm{X}_\bullet \in \mathrm{Affine}\,(1)$ of X_0 eventually sheds all heteroclinic connections as λ goes to 0.

We prove this theorem for the Euler family $\mathrm{E}_\bullet$ in the next Sect. 22.4.1 for (2)⇒(3) and Sect. 22.4.2 for (3)⇒(1). After that step there are two ways to process the general case. On the one hand, the proof performed in Euler's case could be

adapted straightforwardly to fit the more general setting. On the other hand, we can provide a collection of normal forms $X_\bullet^\kappa$ for Affine (1) on which the validity of the equivalences is easily read. This approach brings also the benefit of characterizing completely situations for which (3)⇒(1) holds.

Theorem 2. *Consider a family* $X_\bullet \in$ Affine (1) *as in* (22.5).

1. *There exists a unique*

$$\kappa \in \overline{\mathbb{N}} := \mathbb{Z}_{\geq 0} \cup \{\infty\}$$

such that $X_\bullet$ *is conjugate to one of the models* $X_\bullet^\kappa$

$$X_\lambda^\kappa(x,y) := \begin{bmatrix} x^2+\lambda \\ y - \lambda^\kappa\left(x^2+\lambda\right) \end{bmatrix}$$

where we conventionally identify λ^∞ *with* 0. *This conjugacy can be chosen fibered in the variables* x *and* λ. *Moreover families* $X_\bullet^\kappa$ *are mutually orbitally non-equivalent for differing values of* κ.
2. *The implication (3)⇒(1) in Theorem 1 holds if and only if* $\kappa \in \{0, \infty\}$. *Notice that the condition* $\frac{\partial a_\lambda}{\partial \lambda} = 0$ *implies* $\kappa \in \{0, \infty\}$.

This theorem, proved in Sect. 22.4.3 below, discriminates all three possible qualitative dynamical behaviors occurring in Affine (1).

$\kappa = 0$ **Pure divergence**. For every $\lambda \neq 0$ sufficiently close to 0 the vector field X_λ has no heteroclinic connection while X_0 has a divergent weak separatrix.

$\kappa \in \mathbb{N}_{>0}$ **Sly convergence**. For every $\lambda \neq 0$ sufficiently close to 0 the vector field X_λ has no heteroclinic connection although X_0 has a convergent weak separatrix.

$\kappa = \infty$ **Pure convergence**. For every $\lambda \neq 0$ sufficiently close to 0 the vector field X_λ has a heteroclinic connection so that X_0 has a convergent weak separatrix.

Here the modulus space $\overline{\mathbb{N}}$ for analytical orbital classification is discrete. The property no longer persists for families unfolding a more degenerate saddle-node, i.e. the coalescence of $k+1$ stationary points with $k > 1$. We refer to [28] for this more involved situation.

22.4.1 From Heteroclinic Connections to Convergence

Forget for now that Euler's series (22.10) is divergent. We want to recover its divergence at $\lambda = 0$ by dynamical properties arising in the family when $\lambda \in (\mathbb{C}, 0) \setminus \{0\}$. This story is told during the next two Sects. 22.4.1 and 22.4.2.

First, we must exclude values of the parameter λ for which there are no analytic unstable manifold through the node of the Euler vector field E_λ. Although this phenomenon is not generic, it still turns up for an infinite discrete set of parameters accumulating on 0. The stable manifold is always unique, and the variation of parameters yields that it is given by the graph of

$$\begin{aligned} \mathfrak{s}_\lambda^- &:]-s, s[\longrightarrow \mathbb{R} \\ &x \longmapsto \left(\frac{s-x}{s+x}\right)^{1/2s} \int_x^{-s} \left(\frac{s+u}{s-u}\right)^{1/2s} \mathrm{d}u \,. \end{aligned} \tag{22.12}$$

Proposition 1.

1. E_λ *admits a (unique) analytic unstable manifold* $\mathcal{S}_\lambda^+$ *if and only if*

$$\lambda \in \hat{\Lambda} := \mathbb{R}_{<0} \backslash \frac{-1}{4\mathbb{N}^2} \,.$$

2. *There exists a unique function* $\mathfrak{c} : \sqrt{-\hat{\Lambda}} \to \mathbb{R}$ *such that for all* $\lambda \in \hat{\Lambda}$ *the manifold* $\mathcal{S}_\lambda^+$ *coincides with the graph of*

$$\begin{aligned} \mathfrak{s}_\lambda^+ &:]-s, s[\longrightarrow \mathbb{R} \\ &x \longmapsto \left(\frac{s-x}{s+x}\right)^{1/2s} \left(\mathfrak{c}(s) + \int_x^0 \left(\frac{s+u}{s-u}\right)^{1/2s} \mathrm{d}u\right) . \end{aligned}$$

3. $\mathfrak{c}$ *is analytic.*

Proof. First notice that whatever the value of $\mathfrak{c}(s)$ may be, the graph of $\mathfrak{s}_\lambda^+$ is an integral curve of E_λ even when $\lambda \notin \hat{\Lambda}$. Swapping the order of summation and integration operations in the expansion

$$(u+s)^{\frac{1}{2s}} =: \sum_{n=0}^{\infty} \alpha_n^+(s)\,(u-s)^n \,,$$

which converges uniformly on compact subsets of $]-s, 3s[$, we isolate the candidate singular term of $\mathfrak{s}_\lambda^+$ at s:

$$\mathfrak{s}_\lambda^+(x) = \left(\frac{s-x}{s+x}\right)^{\frac{1}{2s}} \left(\mathfrak{c}(s) - \sum_{n+1 \neq \frac{1}{2s}} \alpha_n^+(s)\,(-1)^n \frac{s^{n+1-\frac{1}{2s}}}{n+1-\frac{1}{2s}} - \alpha_*(s) \ln \frac{s-x}{s}\right)$$

$$+ \text{(analytic at } s\text{)}$$

where

$$\alpha_*(s) := \begin{cases} 0 & \text{if } \frac{1}{2s} \notin \mathbb{N} \\ 1 & \text{otherwise} \end{cases} .$$

Notice that $(s-x)^{\frac{1}{2s}}$ cancels out the non-integral exponent in the power series of the right-hand side. If $\frac{1}{2s} \in \mathbb{N}$, no choice of $\mathfrak{c}(s) \in \mathbb{C}$ may yield an analytic $\mathfrak{s}_\lambda^+$. On the contrary for $\lambda \in \hat{\Lambda}$ we can only have

$$\mathfrak{c}(s) = \sum_{n=0}^{\infty} \alpha_n^+(s)(-1)^n \frac{s^{n+1-\frac{1}{2s}}}{n+1-\frac{1}{2s}} ,$$

which is an analytic function of $s \in \sqrt{-\hat{\Lambda}}$. □

A consequence of the proposition is the following: if $\frac{1}{2s} \in \mathbb{N}$, there is no heteroclinic connection, while a heteroclinic connection for $\lambda \in \hat{\Lambda}$ occurs exactly if

$$\mathfrak{s}_\lambda^-(0) = \mathfrak{s}_\lambda^+(0) ,$$

that is, if

$$\varphi(s) := \mathfrak{c}(s) + \int_{-s}^{0} \left(\frac{s+u}{s-u}\right)^{1/2s} \mathrm{d}u$$

vanishes.

Corollary 1. *If φ vanishes on a set accumulating on 0, then $\hat{\mathfrak{s}}$ converges.*

The proof requires our switching to complex analysis in order to use compactness of normal families of holomorphic functions. The main ingredient is therefore to show that $\left(\mathfrak{s}_\lambda^-\right)_{-1<\lambda<0}$ extends to a uniformly bounded family of analytic functions on the slit unit disc

$$\{z \in \mathbb{C} \backslash [s, \infty[: \ |z| < 1\} .$$

We need to slit the disc because the complex (multivalued) extension of $\mathfrak{s}_\lambda^-$ is given by taking path integrals in the variation of constant method

$$\mathfrak{s}_\lambda^-(z) = \left(\frac{s-z}{s+z}\right)^{1/2s} \int_{\gamma(z)} \left(\frac{s+u}{s-u}\right)^{1/2s} \mathrm{d}u \tag{22.13}$$

where γ is a piecewise smooth path linking $z \neq s$ to $-s$. We choose the determination of the logarithm in such a way that $\mathfrak{s}_\lambda^-(z)$ coincides with (22.12) on $]-s, s[$.

Remark 2. We will discuss the relevance of the multivaluedness of $\mathfrak{s}_\lambda^-$ regarding the question of convergence of $\hat{\mathfrak{s}}$ in the next section, when proving the converse of the corollary.

Lemma 1. *There exists $C > 0$ such that for every $(\lambda, z) \in]-1, 0[\times\left(\mathbb{S}^1 \setminus \{1\}\right)$ we have*

$$|\mathfrak{s}_\lambda^-(z)| \leq C .$$

Proof. Let us build an adequate integration path $\gamma(z)$ for which bounds are easily obtained.

- When $\Im(z) < 0$ we first follow the shortest anticlockwise arc $\gamma^-(z)$ of $\mathbb{S}^1$ joining z to -1, then the interval

$$I_\lambda := [-1, -s] .$$

- Otherwise we follow the shortest clockwise arc $\gamma^+(z)$ of $\mathbb{S}^1$ joining z to -1 before I_λ.

For $u \in I_\lambda$ we have $0 < \frac{u+s}{u-s} < 1$ so that

$$\int_{-1}^{-s} \left(\frac{u+s}{u-s}\right)^{1/2s} \mathrm{d}u \leq (1-s) .$$

Moreover there exists $C_1 \geq 0$ for which

$$|\mathfrak{s}_\lambda^-(z)| \leq \left|\frac{s-z}{s+z}\right|^{1/2s} \left((1-s) + \int_{\gamma^\pm(z)} \left|\frac{s+u}{s-u}\right|^{1/2s} \mathrm{d}u\right) \leq C_1 (1-s) + \pi$$

because, on the one hand, $\left|\frac{s+u}{s-u}\right| \leq \left|\frac{s+z}{s-z}\right|$ when $z \in \mathbb{S}^1$ and $u \in \gamma^\pm(z)$, while, on the other hand, $\left|\frac{s+z}{s-z}\right| \leq \frac{1+s}{1-s}$ and $\lim_{s\to 0} \left(\frac{1+s}{1-s}\right)^{\frac{1}{2s}} = \mathrm{e}$. □

We get on now with proving Corollary 1.

Proof. Each function $\mathfrak{s}_\lambda^-$, holomorphic on the slit unit disc, can be analytically extended to the whole $\mathbb{D}$ precisely when it is analytic near s. This happens precisely when the stable manifold of the node is analytic, in other words when $\varphi(s) = 0$. Let $\Omega \subset \varphi^{-1}(0)$ be a set accumulating on 0. Because of Lemma 1 and of the maximum principle we know that $\mathfrak{s}_\lambda^-$ is bounded on $\mathbb{D}$ uniformly in $-1 < \lambda < 0$, i.e. the family $\left(\mathfrak{s}_\lambda^-\right)_{\lambda\in\Omega}$ is normal. Thus by Montel's theorem we can consider an adherence value (for uniform convergence on compacts sets of $\mathbb{D}$) which must be a solution of Euler's equation (22.9) with analytic Taylor expansion at 0. But there is only one such formal power series solving Euler's equation, namely $\hat{\mathfrak{s}}$. □

Remark 3. We can give a series representation for φ using the expansions

$$(s \pm u)^{\pm\frac{1}{2s}} =: \sum_{n=0}^{\infty} \alpha_n^\pm(s) (u \mp s)^n$$

where the determination of the logarithm on the left-hand side is chosen in such a way that the function is real on $]-s, s[$. In particular

$$\alpha_n^+(-s) = (-1)^{\frac{1}{2s}} \alpha_n^-(s) = \alpha_n^-(s) \exp\frac{\mathrm{i}\pi}{2s} .$$

In that setting

$$\mathfrak{c}(s) := \sum_{n=0}^{\infty} \frac{\alpha_n^+(s)}{n+1-\frac{1}{2s}} (-1)^n s^{n+1-\frac{1}{2s}}$$

and it is easy to compute

$$\int_{-s}^{0} \left(\frac{s+u}{s-u}\right)^{\frac{1}{2s}} \mathrm{d}u = \sum_{n=0}^{\infty} \frac{\alpha_n^-(s)}{n+1+\frac{1}{2s}} (-1)^n s^{n+1+\frac{1}{2s}} = -\mathfrak{c}(-s) \quad ,$$

therefore

$$\varphi(s) = \mathfrak{c}(s) - \mathfrak{c}(-s) \ . \tag{22.14}$$

22.4.2 From Convergence to Heteroclinic Connections

We just observed that if $\mathfrak{s}_\lambda^-$ is uniform (that is, not multivalued) then $\hat{\mathfrak{s}}$ converges. We want to establish the converse statement in the following way. When the formal solution of (22.9) converges it defines a real, entire holomorphic function, in particular for given $x_* > s$

$$(x_*, y_*) := (x_*, \hat{\mathfrak{s}}(x_*))$$

is a well-defined point in $\mathbb{R}^2$. Consider the solution y_λ of (22.11) with $a = 1$ and initial value (x_*, y_*), and perform its local analytic (i.e. multivalued) continuation over

$$\overline{\mathbb{D}}_\lambda := \overline{\mathbb{D}} \setminus \{\pm s\} \quad .$$

We are more particularly interested in the analytic continuation of y_λ along the unit circle, which can be performed in the universal cover of $\overline{\mathbb{D}} \setminus [-s, s]$ as we explain below. We identify the action of the deck transform of this covering with the symbolic multiplication of z by $\exp 2\mathrm{i}\pi$, so that the analytic continuation of y_λ along $\mathbb{S}^1$ can be conveniently written $y_\lambda(x_* \exp 2\mathrm{i}\pi)$. Because $(y_\lambda)_\lambda$ converges uniformly on compact subsets of the preimage of $\mathbb{S}^1$ in the universal cover as $\lambda \underset{<}{\to} 0$, if $\hat{\mathfrak{s}}$ converges then its sum y_0 is uniform and we must have

$$\lim_{\lambda \underset{<}{\to} 0} y_\lambda(x_* \exp 2\mathrm{i}\pi) = y_* \ .$$

Let us see how this observation relates to the persistence of heteroclinic connections.

Proposition 2. *For every* $\lambda \in \hat{\Lambda}$ *and* $(x_*, y) \in \mathbb{R}_{>s} \times \mathbb{R}$ *write* $z \mapsto y_\lambda(z, y)$ *the solution of* (22.11) *with initial value* (x_*, y). *The local analytic continuation of* $y_\lambda(\bullet, y)$ *over* $\mathbb{S}^1$ *follows the rule*

$$y_\lambda(x_* \exp 2i\pi, y) = y - 2i\left(\frac{x_* - s}{x_* + s}\right)^{\frac{1}{2s}} \varphi(s) \sin\frac{\pi}{2s},$$

where φ *is defined by* (22.14). *In particular*

$$\lim_{\substack{\lambda \to 0 \\ <}} y_\lambda(x_* \exp 2i\pi) = y_*$$

if and only if

$$\lim_{s \to 0} \varphi(s) \sin\frac{\pi}{2s} = 0.$$

Remark 4. Notice that when $y \in \mathbb{R}$ and $\lambda \in \hat{\Lambda}$ the continued value $y_\lambda(x_* \exp 2i\pi, y)$ is never real since $\left(\frac{x_*-s}{x_*+s}\right)^{\frac{1}{2s}} \varphi(s) \sin\frac{\pi}{2s} \in \mathbb{R}$.

In order to establish the proposition we need to understand the monodromy of

$$\hat{\mathfrak{g}}_s : z \neq \pm s \longmapsto \left(\frac{s+z}{s-z}\right)^{\frac{1}{2s}}.$$

We fix a determination $\hat{\mathfrak{g}}_s^*$ of $\hat{\mathfrak{g}}_s$ on $\left(\overline{\mathbb{D}}\backslash\mathbb{R}\right) \cup]-s, s[$ in such a way that $\hat{\mathfrak{g}}_s^*|_{]-s,s[}$ coincides with the canonical real determination used previously. For any path γ, starting from 0 with image included in $\overline{\mathbb{D}}_\lambda$, we define

$$\hat{\mathfrak{g}}_s(\gamma) := \hat{\mathfrak{g}}_s^*(0) \exp \int_\gamma \frac{du}{u^2 + \lambda}.$$

Fix a system $\gamma^\pm$ of generators of $\pi_1\left(\overline{\mathbb{D}}_\lambda, 0\right)$ whose index around $\pm s$ is 1 and 0 around the other point as in Fig. 22.3. The monodromy of $\hat{\mathfrak{g}}_s$ is multiplicative and given by

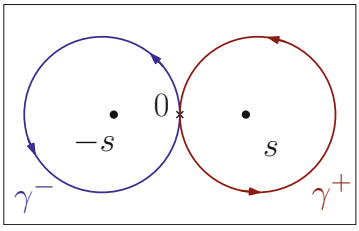


Fig. 22.3 Generators of $\pi_1\left(\overline{\mathbb{D}}_\lambda, 0\right)$

$$\hat{\mathfrak{g}}_s\left(\gamma^{\pm}\gamma\right) = \hat{\mathfrak{g}}_s\left(\gamma\right)\exp\oint_{\gamma^{\pm}}\frac{\mathrm{d}u}{u^2+\lambda} = \hat{\mathfrak{g}}_s\left(\gamma\right)\exp\frac{\pm\mathrm{i}\pi}{s}\ ,$$

according to the residue formula and the identity

$$\frac{-1}{z^2+\lambda} = \frac{1}{2s}\left(\frac{1}{s+z}+\frac{1}{s-z}\right)\ .$$

In particular

$$\hat{\mathfrak{g}}_s\left(\gamma^{-}\gamma^{+}\gamma\right) = \mathfrak{g}_s\left(\gamma\right)$$

so that $\hat{\mathfrak{g}}_s^*$ is actually holomorphically extendable to $\overline{\mathbb{D}}\backslash[-s,s]$, as claimed. We prove now the proposition.

Proof. The method of variation of parameters yields the following expression for the monodromy

$$y_\lambda\left(x_*\exp 2\mathrm{i}\pi, y\right) - y = -\frac{1}{\hat{\mathfrak{g}}_s^*\left(x_*\right)}\oint_{\mathbb{S}^1}\hat{\mathfrak{g}}_s^*\left(u\right)\mathrm{d}u\ .$$

It can be computed by deforming $\mathbb{S}^1$ into the concatenation $\gamma^-\gamma^+$ of generators of $\pi_1\left(\overline{\mathbb{D}}_\lambda, 0\right)$ given by

$$\begin{aligned}\gamma^- &: t\in[0,1]\longmapsto -s+s\exp\left(2\mathrm{i}\pi t\right)\\ \gamma^+ &: t\in[0,1]\longmapsto s-s\exp\left(2\mathrm{i}\pi\left(t-1\right)\right)\end{aligned}$$

using the relation

$$\oint_{\mathbb{S}^1}\hat{\mathfrak{g}}_s^*\left(u\right)\mathrm{d}u = \int_{\gamma^-}\hat{\mathfrak{g}}_s\left(u\right)\mathrm{d}u + \int_{\gamma^+}\hat{\mathfrak{g}}_s\left(u\right)\mathrm{d}u$$

(notice that we do not use the symbol $\oint$ for the paths $\gamma^{\pm}$ because the integration is actually performed in the universal cover of $\overline{\mathbb{D}}_\lambda$ and the lift of $\gamma^{\pm}$ is not a loop). Using the notations and formulas presented in Remark 3 we compute

$$\begin{aligned}\int_{\gamma^-}\hat{\mathfrak{g}}_s\left(u\right)\mathrm{d}u &= \sum_{n=0}^{\infty}\frac{\alpha_n^-\left(s\right)}{n+1+\frac{1}{2s}}\left[z^{n+1+\frac{1}{2s}}\right]_s^{s\exp 2\mathrm{i}\pi}\\ &= \left(1-\exp\frac{\mathrm{i}\pi}{s}\right)\mathfrak{c}\left(-s\right)\end{aligned}$$

with $z := u + s$, then

$$\int_{\gamma^+} \hat{\mathfrak{g}}_s(u)\,\mathrm{d}u = \sum_{n=0}^{\infty} \frac{\alpha_n^+(s)\,(-1)^{n+1}}{n+1-\frac{1}{2s}} \left[z^{n+1-\frac{1}{2s}} \right]_{s\exp(-2\mathrm{i}\pi)}^{s}$$
$$= \left(\exp\frac{\mathrm{i}\pi}{s} - 1 \right) \mathfrak{c}(s)$$

with $z := s - u$. The conclusion follows from

$$\frac{1}{\hat{\mathfrak{g}}_s^*(x_*)} = \left(\frac{x_* - s}{x_* + s} \right)^{\frac{1}{2s}} \exp\frac{-\mathrm{i}\pi}{2s}$$

and from (22.14). □

We end the story by an explicit computation which settles the question of the divergence of $\hat{\mathfrak{s}}$.

Lemma 2. *For every $\lambda \in \hat{\Lambda}$ we have*

$$\varphi(s) \sin\frac{\pi}{2s} = \pi\ .$$

Proof. We just proved

$$\varphi(s) \sin\frac{\pi}{2s} = \pi \times \exp\frac{-\mathrm{i}\pi}{2s} \times \frac{1}{2\mathrm{i}\pi} \oint_{\mathbb{S}^1} \hat{\mathfrak{g}}_s^*(u)\,\mathrm{d}u$$
$$= \pi \times \frac{1}{2\mathrm{i}\pi} \oint_{\mathbb{S}^1} \left(\frac{u+s}{u-s} \right)^{\frac{1}{2s}} \mathrm{d}u\ .$$

The latter integral can be evaluated using the residue formula at ∞ since $z \mapsto \left(\frac{z+s}{z-s}\right)^{\frac{1}{2s}}$ is holomorphic at this point. Setting $w := \frac{1}{u}$ we compute

$$\left(\frac{z+s}{z-s} \right)^{\frac{1}{2s}} = \left(\frac{1+ws}{1-ws} \right)^{\frac{1}{2s}} = 1 + w + \mathrm{o}(w)$$

and

$$\frac{1}{2\mathrm{i}\pi} \oint_{\mathbb{S}^1} \left(\frac{u+s}{u-s} \right)^{\frac{1}{2s}} \mathrm{d}u = \frac{1}{2\mathrm{i}\pi} \oint_{\mathbb{S}^1} \exp(w + \mathrm{o}(w)) \frac{\mathrm{d}w}{w^2} = 1\ .$$

22.4.3 Normal Forms

We just established the equivalence in the Euler family between

- divergence of $\hat{\mathfrak{s}}$,
- absence of heteroclinic connections (non-vanishing of φ),
- non-vanishing of the integral

$$\varphi_s^{\mathrm{n}} := \frac{1}{2\mathrm{i}\pi} \oint_{\mathbb{S}^1} \left(\frac{u+s}{u-s} \right)^{\frac{1}{2s}} \mathrm{d}u = \frac{1}{\pi} \varphi(s) \sin \frac{\pi}{2s} = 1 \, .$$

Remark 5. Be careful that the exponent "n" refers to "node", and is not meant to be thought of as a variable. The terminology choice will be explained in the next part of the chapter.

In order to establish the classification Theorem 2 we need to find an (almost) invariant quantity under changes of coordinates. This invariant turns out to be $\varphi_\bullet^{\mathrm{n}}$. One can argue that it suffices to consider φ instead, which is somehow nicer because of its dynamical flavor. Yet φ is afflicted of serious drawbacks:

- φ presents an accumulation of poles as $s \underset{>}{\longrightarrow} 0$, and there is no hope of extending it analytically at 0,
- φ is not even: there is no hope of extending it holomorphically on any open annulus surrounding 0 *as a function of* λ.

None of these shortcomings hinder $\varphi_\bullet^{\mathrm{n}}$, even in the more general setting of affine unfoldings.

Proposition 3. *For* $\mathrm{X}_\bullet \in \mathrm{Affine}(1)$ *as in* (22.5) *we may find* $\rho > 0$ *such that* $(\lambda, x) \mapsto a_\lambda(x)$ *is holomorphic on (a neighborhood of)* $\rho^2\mathbb{D} \times \rho\overline{\mathbb{D}}$. *For* $s \in \rho\mathbb{D}\backslash\{0\}$ *define*

$$\varphi_s^{\mathrm{n}} := \frac{1}{2\mathrm{i}\pi} \oint_{\rho\mathbb{S}^1} a_{-s^2}(u) \left(\frac{u+s}{u-s} \right)^{\frac{1}{2s}} \mathrm{d}u \, . \tag{22.15}$$

1. *The holomorphic mapping* $s \mapsto \varphi_s^{\mathrm{n}}$ *can be continued to an even germ of a holomorphic function at* 0 *satisfying*

$$\varphi_0^{\mathrm{n}} = \frac{1}{2\mathrm{i}\pi} \oint_{\mathbb{S}^1} a_0(u) \exp \frac{1}{u} \mathrm{d}u \, .$$

2. *Write* $a_\lambda(x) = \sum_{n=0}^{\infty} \phi_n(\lambda) x^n$. *Then for* $s \in \rho\mathbb{D}\backslash\{0\}$

$$\varphi_s^{\mathrm{n}} = \sum_{n=0}^{\infty} \frac{\phi_n(-s^2)}{(n+1)!} \times \frac{1}{2^n} \sum_{p+q=n} \binom{n}{p} \prod_{j=1}^{p} (1+2sj) \prod_{j=1}^{q} (1-2sj) \, ,$$

with limit

$$\varphi_0^{\mathrm{n}} = \sum_{n=0}^{\infty} \frac{\phi_n(0)}{(n+1)!} .$$

3. *The formal solution* $\hat{\mathfrak{s}}$ *with* $\hat{\mathfrak{s}}(0) = 0$ *of*

$$x^2 y'(x) = y(x) - x^2 a_0(x)$$

converges if and only if $\varphi_0^{\mathrm{n}} = 0$.

The third statement of the proposition is actually Briot–Bouquet's result [3].

Remark 6. We deduce the determination of $\mathfrak{g}_s := \left(\frac{\bullet+s}{\bullet-s}\right)^{\frac{1}{2s}}$ from that of the function $\hat{\mathfrak{g}}_s$ built in Sect. 22.4.2 by setting

$$\mathfrak{g}_s := \hat{\mathfrak{g}}_s \exp \frac{-\mathrm{i}\pi}{2s} .$$

The multiplicative monodromy of $\mathfrak{g}_s$ is the same as that of $\hat{\mathfrak{g}}_s$.

Proof.

1. Although it is a consequence of 2, we can prove directly the property. First notice that $\varphi_{-s}^{\mathrm{n}} = \varphi_s^{\mathrm{n}}$. Also $z \mapsto a_{-s^2}(z)\, \mathfrak{g}_s(z)$ converges uniformly to $z \mapsto a_0(z) \exp \frac{1}{z}$ on $\rho\mathbb{S}^1$ as $s \to 0$, so that $s \mapsto \varphi_s^{\mathrm{n}}$ is bounded on a pointed neighborhood of 0. Riemann's removable singularity theorem yields the conclusion. This is a trick used extensively in this text.
2. For $n \in \mathbb{Z}_{\geq 0}$ let us evaluate

$$t_s(n) := \frac{1}{2\mathrm{i}\pi} \oint_{\mathbb{S}^1} u^n \mathfrak{g}_s(u)\, \mathrm{d}u .$$

The residue formula used in Lemma 2 to compute $t_s(0)$ sure works here, yet one would have to formally derive a closed-form for the Taylor coefficients of $z \mapsto \mathfrak{g}_s(z)$ at ∞, which is no trivial task. We relate instead the computation at hands to the Beta function, more precisely its integral representation along a Pochhammer contour around 0 and 1. Introduce first the contour around $-s$ and s

$$\mathcal{P} := \gamma^+ \gamma^- \left(\gamma^+\right)^{-1} \left(\gamma^-\right)^{-1} \tag{22.16}$$

where $\gamma^{\pm}$ are generators of $\pi_1\left(\overline{\mathbb{D}} \setminus \{\pm s\}, 0\right)$ as described in Fig. 22.3. The identity

$$\oint_{\mathcal{P}} u^n \mathfrak{g}_s(u)\, \mathrm{d}u = \left(\exp \frac{-\mathrm{i}\pi}{s} - 1\right) \oint_{\mathbb{S}^1} u^n \mathfrak{g}_s(u)\, \mathrm{d}u$$

holds because the value of $\mathfrak{g}_s$ above $\gamma^+\gamma^-$ is multiplied by $\exp\frac{-\mathrm{i}\pi}{s}$ as compared to that above $\left(\gamma^+\right)^{-1}(\gamma^-)^{-1}$.

We invoke now the standard formula

$$(1-\exp 2\mathrm{i}a\pi)(1-\exp 2\mathrm{i}b\pi)\,\mathrm{B}(a,b)=\oint_{\hat{\mathcal{P}}} z^{a-1}(1-z)^{b-1}\,\mathrm{d}z\,, \qquad (22.17)$$

where $\hat{\mathcal{P}}$ is a Pochhammer contour around 0 and 1. We can take for $\hat{\mathcal{P}}$ the image of $\mathcal{P}$ under the change of variable

$$z:=\frac{1}{2s}(s-u)$$

which transforms $u-s$ into $-2sz$ (maps s on 0) and $u+s$ into $2s(1-z)$ (maps $-s$ on 1). It is therefore relevant to work with the expansion

$$u^n=\frac{1}{2^n}\sum_{p+q=n}\binom{n}{p}(u+s)^p(u-s)^q\ .$$

From (22.17) we compute, for $p+q=n$ non-negative integers,

$$\begin{aligned}
t_{p,q}&:=\oint_{\mathbb{S}^1}(u+s)^{p+\frac{1}{2s}}(u-s)^{q-\frac{1}{2s}}\,\mathrm{d}u\\
&=\frac{1}{\exp\frac{-\mathrm{i}\pi}{s}-1}\oint_{\mathcal{P}}(u+s)^{p+\frac{1}{2s}}(u-s)^{q-\frac{1}{2s}}\,\mathrm{d}u\\
&=\frac{(2s)^{n+1}}{1-\exp\frac{-\mathrm{i}\pi}{s}}(-1)^q\exp\frac{-\mathrm{i}\pi}{2s}\oint_{\hat{\mathcal{P}}}(1-z)^{p+\frac{1}{2s}}z^{q-\frac{1}{2s}}\mathrm{d}z\\
&=(-1)^q\frac{(2s)^{n+1}}{1-\exp\frac{-\mathrm{i}\pi}{s}}\exp\frac{-\mathrm{i}\pi}{2s}\left(1-\exp\frac{-\mathrm{i}\pi}{s}\right)\left(1-\exp\frac{\mathrm{i}\pi}{s}\right)\\
&\quad\times\mathrm{B}\left(1+q-\frac{1}{2s},1+p+\frac{1}{2s}\right)\\
&=(-1)^q\,2\mathrm{i}\,(2s)^{n+1}\sin\frac{\pi}{2s}\mathrm{B}\left(1+q-\frac{1}{2s},1+p+\frac{1}{2s}\right)\\
&=(-1)^q\frac{2\mathrm{i}}{(n+1)!}(2s)^{n+1}\sin\frac{\pi}{2s}\Gamma\left(1+q-\frac{1}{2s}\right)\Gamma\left(1+p+\frac{1}{2s}\right)\ .
\end{aligned}$$

Since $\Gamma(z+1)=z\Gamma(z)$ and $\Gamma(1-z)\,\Gamma(z)=\frac{\pi}{\sin\pi z}$ we deduce finally

$$t_{p,q} = \frac{2\mathrm{i}\pi}{(n+1)!} \prod_{j=1}^{p} (1+2sj) \prod_{j=1}^{q} (1-2sj)$$

and

$$t_s(n) = \frac{1}{2^n (n+1)!} \sum_{p+q=n} \binom{n}{p} \prod_{j=1}^{p} (1+2sj) \prod_{j=1}^{q} (1-2sj) \ .$$

Because φ_s^{n} is obtained by integrating a holomorphic 1-form on a compact loop we can swap the order of summation operators:

$$\oint_{\mathbb{S}^1} a_{-s^2}(u)\, \mathfrak{g}_s(u)\, \mathrm{d}u = \sum_{n=0}^{\infty} \phi_n\left(-s^2\right) t_s(n) \ .$$

3. After applying a convenient linear scaling of the x-coordinate we can assume that a_0 is holomorphic on $\overline{\mathbb{D}}$. For $z \in \overline{\mathbb{D}} \setminus [0,1]$ consider a path $\gamma(z)$ joining 0 directly to -1, then reaching z within the domain. The function

$$\mathfrak{s}_0^- \ : \ z \in \overline{\mathbb{D}} \setminus [0,1] \longmapsto \exp \frac{-1}{z} \int_{\gamma(z)} a_0(u) \exp \frac{1}{u} \mathrm{d}u$$

is well defined and holomorphic on $\overline{\mathbb{D}} \setminus [0,1]$. It is the only solution of the equation which tends to 0 at 0 over $\mathbb{R}_{<0}$. It must therefore coincide with $\hat{\mathfrak{s}}$ when one of the two objects represents a holomorphic function on $\overline{\mathbb{D}}$. The conclusion follows from the fact that φ_0^{n} embodies the monodromy of the multivalued continuation of $\mathfrak{s}_0^-$ on $\overline{\mathbb{D}} \setminus \{0\}$.

□

Let us present now the classification theorem.

Theorem 3. *Take two families* $\mathrm{X}_\bullet$ *and* $\widetilde{X}_\bullet$ *of* Affine (1). *The following properties are equivalent.*

1. *There exists a germ of a holomorphic function* $\lambda \mapsto c(\lambda)$ *with* $c(0) \neq 0$ *such that for all* s *sufficiently close to* 0

$$\tilde{\varphi}_s^{\mathrm{n}} = c\left(-s^2\right) \varphi_s^{\mathrm{n}} \ .$$

2. $\mathrm{X}_\bullet$ *and* $\widetilde{X}_\bullet$ *are conjugate.*

Any conjugacy between the two families must fix λ*, and in that case a change of coordinates* $\Psi_\bullet$ *such that* $\Psi_\bullet^* \mathrm{X}_\bullet = \widetilde{X}_\bullet$ *exists in the form*

$$(\lambda, x, y) \longmapsto (\lambda \ , \ x \ , \ yc(\lambda) + \phi_\lambda(x)) \ .$$

Proof.

1. $\Rightarrow$ 2. We find a germ of a holomorphic function $(s,x) \mapsto \psi_s(x)$ such that

$$\Psi_{-s^2}(x,y) := \left(x, yc\left(-s^2\right) + \psi_s(x)\right)$$

satisfies $\Psi^*_{-s^2} \mathrm{X}_{-s^2} = \widetilde{X}_{-s^2}$. We prove next that $s \mapsto \psi_s$ is even, so that there exists a holomorphic function $(\lambda, x) \mapsto \phi_\lambda(x)$ with $\phi_{-s^2} = \psi_s$. By definition we need to solve the equation

$$\mathrm{D}\Psi_\lambda\left(\mathrm{X}_\lambda\right) = \widetilde{X}_\lambda \circ \Psi_\lambda$$

where $\lambda := -s^2$, that is

$$z^2 \psi_s'(z) = \psi_s(z) - \delta_\lambda(z)\left(z^2 + \lambda\right)$$

where

$$\delta_\lambda(z) := \tilde{a}_\lambda(z) - c(\lambda)\, a_\lambda(z) \ .$$

Without loss of generality we can assume that c is holomorphic on $\mathbb{D}$. Suppose first that $0 < s < 1$. The method of variation of the constant yields

$$\psi_s(z) = \frac{1}{\mathfrak{g}_s(z)} \int_z^{-s} \delta_\lambda(u)\, \mathfrak{g}_s(u)\, \mathrm{d}u \ ,$$

which is holomorphic on $\overline{\mathbb{D}} \backslash [s,1]$. Because $\tilde{\varphi}^{\mathrm{n}}_s = c\left(-s^2\right) \varphi^{\mathrm{n}}_s$ the function ψ_s extends to a uniform (holomorphic) function on $\overline{\mathbb{D}} \backslash \{s\}$. As in Lemma 1 it is easy to prove that ψ_s is bounded on $\mathbb{S}^1$ (uniformly in s). Using the maximum modulus principle and Riemann's removable singularity theorem we deduce that ψ_s extends holomorphically to $\overline{\mathbb{D}}$. Montel's theorem ensures that $(\psi_s)_s$ converges uniformly on $\overline{\mathbb{D}}$ to some function ψ_0 for which $\Psi_0^* \mathrm{X}_0 = \widetilde{X}_\lambda$.

The above construction can be holomorphically continued for all $s \in \mathbb{D} \backslash \mathbb{R}$, in that case the graph of ψ_s coincides with the invariant manifold of the collection $\Delta_\bullet \in \mathrm{Affine}\,(1)$,

$$\Delta_\lambda(x,y) := \begin{bmatrix} x^2 + \lambda \\ y - \delta_\lambda(x)\left(x^2 + \lambda\right) \end{bmatrix} ,$$

passing through the hyperbolic point $(-s, 0)$ and transverse to the line $\{z = -s\}$. This manifold is unique, as other non-vertical trajectories of Δ_λ are multivalued. This property guarantees that a heteroclinic connection occurs in Δ_λ, otherwise ψ_s would not be uniform near $(s, 0)$. Therefore the local graph of ψ_s near $(s, 0) = (-(-s), 0)$ coincides with that of ψ_{-s}. From the analytic continuation principle

we derive $\psi_{-s} = \psi_s$ for $s \in \mathbb{D}\backslash\mathbb{R}$, which allows to extend holomorphically $(s,x) \mapsto \psi_s(x)$ to $\mathbb{D}\times\mathbb{D}$ to an even function of s, as expected.

2. $\Rightarrow$1. Take an orbital equivalence

$$\Psi : (\lambda, x, y) \longmapsto (\phi(\lambda), \Psi_\lambda(x,y)) \in \mathrm{Diff}\left(\mathbb{C}^3, 0\right)$$

between $\mathrm{X}_\bullet$ and $\widetilde{X}_\bullet$. Assuming that Ψ is holomorphic on $\mathbb{D}\times\overline{\mathbb{D}}\times\mathbb{D}$ does not lessen the generality of our argument. We prove that $\phi = \mathrm{Id}$. The key ingredient is the following classical fact.

Lemma 3. *Take $p \in \mathbb{C}^2$ a stationary point of a holomorphic vector field X, and consider the linear part of X at p, i.e. the linear mapping $\mathrm{D}X(p)$. Let $L_p(X)$ denote the equivalence class of its spectrum under the equivalence*

$$\{\lambda_1,\lambda_2\} \leftrightsquigarrow \left\{\tilde{\lambda}_1,\tilde{\lambda}_2\right\} \iff \left(\exists c \in \mathbb{C}_{\neq 0}\right) : \{\lambda_1,\lambda_2\} = c\left\{\tilde{\lambda}_1,\tilde{\lambda}_2\right\}.$$

Then $L_p(X)$ is invariant under orbital equivalence.

In our situation for given λ the diffeomorphism Ψ_λ maps $p_\pm := \left(\pm\sqrt{-\lambda}, 0\right)$ to $\tilde{p}_\pm = \left(\pm(-1)^\ell\sqrt{-\phi(\lambda)}, 0\right)$ for some integer ℓ. Because the spectrum of the linearization of X_λ at $p_\pm$ is $\left\{\pm 2\sqrt{-\lambda}, 1\right\}$ we must have

$$\begin{cases} \sqrt{-\lambda} & = \sqrt{-\phi(\lambda)}\,(-1)^\ell \\ \text{or} & \\ 1 & = \sqrt{-\lambda}\sqrt{-\phi(\lambda)}\,(-1)^\ell \end{cases}.$$

The former identity yields $\lambda = \phi(\lambda)$ while the latter $\lambda\phi(\lambda) = 1$ cannot hold on a neighborhood of 0. Also Ψ_λ must fix each stationary point $(\pm s, 0)$. We will not prove that Ψ_λ can be taken to act identically on the x-variable, although it is the case (see, e.g., [28, 29]). The other claims can be recovered by a formal computation.

□

Corollary 2. *For $\mathrm{X}_\bullet \in \mathrm{Affine}(1)$ there exists a unique*

$$\kappa \in \overline{\mathbb{N}} := \mathbb{Z}_{\geq 0} \cup \{\infty\}$$

such that $\mathrm{X}_\bullet$ is conjugate to one of the models $\mathrm{X}^\kappa_\bullet$

$$\mathrm{X}^\kappa_\lambda(x,y) := \begin{bmatrix} x^2+\lambda \\ y - \lambda^\kappa\left(x^2+\lambda\right) \end{bmatrix}$$

where we conventionally identify λ^∞ with 0. Moreover families $\mathrm{X}^\kappa_\bullet$ are mutually orbitally non-equivalent for differing values of κ.

Proof. There exists a unique κ such that

$$\varphi_s^{\mathrm{n}} = \frac{s^{2\kappa}}{c\left(-s^2\right)}$$

for a germ c of a holomorphic function at 0 satisfying $c(0) \neq 0$. Observe that the invariant $\tilde{\varphi}_s^{\mathrm{n}}$ associated with $\mathrm{X}_\lambda^\kappa$ equals $s^{2\kappa}$. Using Theorem 3 we obtain the first claim. The theorem also implies that if $\mathrm{X}_\bullet^\kappa$ is orbitally equivalent to $\mathrm{X}_\bullet^{\tilde{\kappa}}$ then $\lambda^{\tilde{\kappa}} = \lambda^\kappa c(\lambda)$ for some holomorphic function c with $c(0) \neq 0$ and every λ close enough to 0. Therefore $\kappa = \tilde{\kappa}$. □

22.5 Basic Objects and Notations

22.5.1 Standard Notations

In this paragraph n is a positive integer. All rings are commutative and unital.

- We conventionally use $\mathbb{N} := \{1, 2, \ldots\}$. By putting expressions as index of $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, or $\mathbb{C}$ we build subsets of the space satisfying said expressions, e.g. $\mathbb{R}_{<-1} =]\infty, -1[$ or $\mathbb{Z}_{\geq 0} = \{0\} \cup \mathbb{N}$.
- The open unit disc of $\mathbb{C}$ is written

  $$\mathbb{D} := \{z \in \mathbb{C} \,:\, |z| < 1\}$$

  and we denote by $\overline{\mathbb{D}} := \mathrm{adh}(\mathbb{D})$ the closed unit disc. Also

  $$\mathbb{S}^1 := \overline{\mathbb{D}} \backslash \mathbb{D} = \partial\mathbb{D} = \{z \in \mathbb{C} \,:\, |z| = 1\}$$

  stands for the unit circle of the complex line.
- A complex number $z \in \mathbb{C}$ has real part $\Re(z)$ and imaginary part $\Im(z)$.
- The standard Riemann sphere is written $\overline{\mathbb{C}} := \mathbb{C} \cup \{\infty\}$.
- The multiplicative group of invertible elements of a ring $\mathcal{R}$ is written $\mathcal{R}^\times$.
- The ring of polynomials in the multi-variable $\left(z_j\right)_{1 \leq j \leq n}$ over a ring $\mathcal{R}$ is written

  $$\mathcal{R}[z_1, \ldots, z_n],$$

  while for $\star \in \{<, \leq, =, \geq, >\}$ and $d \in \mathbb{N}$ the notation $\mathcal{R}[z_1, \ldots, z_n]_{\star d}$ stands for the set of such polynomials of homogeneous degree δ satisfying $\delta \star d$.
- The ring of all formal power series in the variables $\left(z_j\right)_{1 \leq j \leq n}$ over $\mathcal{R}$ is written

  $$\mathcal{R}[[z_1, \ldots, z_n]].$$

- For $p \in \mathbb{C}^n$ the notation

  $$(\mathbb{C}^n, p)$$

should stand for the set of domains of $\mathbb{C}^n$ containing p, but by a standard and convenient abuse of notations we actually write $(\mathbb{C}^n, p)$ to mean some small enough such domain, much like the usage for Landau's o $(\bullet)$ and O $(\bullet)$ notations.

- The algebra of holomorphic functions on an open set $\mathcal{U} \subset \mathbb{C}^n$ is written Holo $(\mathcal{U})$. We say that a function is holomorphic on $A \subset \mathbb{C}^n$ if it belongs to some Holo $(\mathcal{U})$ for $A \subset \mathcal{U}$. The algebra of all such germs of a function is denoted by Holo (A).
- In the special case $A = \{p\} \subset \mathbb{C}^n$ we more conventionally refer to Holo $(\{p\})$ as

$$\text{Holo}\,(\mathbb{C}^n, p)\ ,$$

the algebra of germs at p of a holomorphic functions. The group Holo $(\mathbb{C}^n, p)^\times$ consists of all germs $U \in \text{Holo}\,(\mathbb{C}^n, p)$ such that $U(p) \neq 0$.
- If moreover $p = 0$, we identify Holo $(\mathbb{C}^n, 0)$ with the sub-algebra

$$\mathbb{C}\{z_1, \ldots, z_n\}$$

of $\mathbb{C}[[z_1, \ldots, z_n]]$ consisting in formal power series which are absolutely convergent on a neighborhood of 0.
- For a domain $\mathcal{D} \subset \mathbb{C}^{n+1}$ of the form $\bigcup_{s \in \Sigma} \{s\} \times D_s$ with $\Sigma \subset \mathbb{C}$ and $D_s \subset \mathbb{C}^n$ we define the functional space

$$\text{Holo}_{\text{c}}\,(\mathcal{D}) := \left\{ f_\bullet \in C^0\,(\text{adh}\,(\mathcal{D})) \ : \ f_\bullet \in \text{Holo}\,(\mathcal{D})\ ,\ (\forall s \in \text{adh}\,(\Sigma))\ f_s \in \text{Holo}\,(D_s) \right\}\ .$$

- The Holo $(\mathcal{U})$-module of all holomorphic vector fields on $\mathcal{U}$ is written $\mathfrak{X}\,(\mathcal{U})$. The Holo $(\mathbb{C}^n, p)$-module of all germs at p of a holomorphic vector field is written

$$\mathfrak{X}\,(\mathbb{C}^n, p)\ .$$

- The set of biholomorphic mappings $\mathcal{U} \to \tilde{\mathcal{U}}$ from an open set $\mathcal{U} \subset \mathbb{C}^n$ onto another one $\tilde{\mathcal{U}}$ is written $\text{Diff}\left(\mathcal{U} \to \tilde{\mathcal{U}}\right)$. As before this construction can be germified near any $A,\ \tilde{A} \subset \mathbb{C}^n$, yielding the set $\text{Diff}\left(A \to \tilde{A}\right)$ whose elements Ψ belong to some $\text{Diff}\left(\mathcal{U} \to \tilde{\mathcal{U}}\right)$ with $A \subset \mathcal{U}$, $\tilde{A} \subset \tilde{\mathcal{U}}$, and $\Psi(A) = \tilde{A}$.
- In the special case $A = \{p\}$ and $\tilde{A} = \{\tilde{p}\}$ we conventionally write $\text{Diff}\,((\mathbb{C}^n, p) \to (\mathbb{C}^n, \tilde{p}))$ instead. If moreover $p = \tilde{p}$, we name

$$\text{Diff}\,(\mathbb{C}^n, p)$$

the (pseudo)group of germs of a diffeomorphism fixing p.
- A tuple of power series $\Psi = \left(\Psi_j\right)_{1 \leq j \leq n} \in \mathbb{C}[[z_1, \ldots, z_n]]^n$ is a **formal diffeomorphism** when $\Psi(0) = 0$ and Ψ is invertible for the composition of formal power series (that is, $\text{D}\Psi(0) \in \text{GL}_n(\mathbb{C})$). The group of all such formal diffeomorphisms is written

$$\widehat{\text{Diff}}\,(\mathbb{C}^n, 0)\ .$$

22.5.2 Lie Derivative

Till the end of Sect. 22.5 we are given a vector field $Z \neq 0$ holomorphic on a domain $\mathcal{U} \subset \mathbb{C}^2$, which we understand as a section

$$\begin{aligned} \mathcal{U} &\longrightarrow \mathrm{T}\mathcal{U} = \mathcal{U} \times \mathbb{C}^2 \\ p &\longmapsto (p\ ,\ Z(p)) \end{aligned}$$

of the tangent bundle of $\mathcal{U}$. We write vector fields $Z = \begin{bmatrix} A \\ B \end{bmatrix}$ as derivations expressed in the canonical basis $\left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)$, say

$$Z = A\frac{\partial}{\partial x} + B\frac{\partial}{\partial y}$$

for two functions $A,\ B \in \mathrm{Holo}(\mathcal{U})$ not both identically zero.

Example 1. The simplest saddle-node encountered in (22.1) can be written

$$\mathrm{X}_0^\infty(x, y) = x^2\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\ .$$

The associated **Lie (directional) derivative** on functions f (or formal power series) is defined by

$$Z \cdot f := A\frac{\partial f}{\partial x} + B\frac{\partial f}{\partial y}\ .$$

Considering $Z(p)$ as an element of the tangent space of $\mathcal{U}$ at p we have

$$(Z \cdot f)(p) = \mathrm{D}_p f(Z(p))\ .$$

The **Lie bracket** of two vector fields X and Y is the vector field whose action by derivation is

$$[X, Y] \cdot f := X \cdot Y \cdot f - Y \cdot X \cdot f\ .$$

We write for short $[X, Y] = X \cdot Y - Y \cdot X$, which makes sense component-wise and endows the space of vector fields with a Lie algebra structure. When $[X, Y] = 0$ we say that X and Y **commute**.

We define inductively for $m \in \mathbb{Z}_{\geq 0}$

$$\begin{aligned} Z^{\cdot 0} f &:= f \\ Z^{\cdot m+1} &:= Z \cdot (Z^{\cdot m} f)\ . \end{aligned}$$

The action is extended component-wise to vectors or matrices of functions.

Any holomorphic function $H \in \mathrm{Holo}\,(\mathcal{U})$ such that

$$Z \cdot H = 0$$

is called a **first integral** of Z.

Example 2. The function $H\ :\ (x,y) \mapsto y \exp \frac{1}{x}$ is a first integral of the saddle-node X_0^∞ on $\mathbb{C}^\times \times \mathbb{C}$.

22.5.3 Flow, Integral Curves and Singularities

The **local flow** of Z at $p \in \mathcal{U}$ is the germ of a mapping

$$\begin{aligned} \Phi_Z^\bullet\ :\ \left(\mathbb{C}^2, p\right) \times (\mathbb{C}, 0) &\longrightarrow \mathbb{C}^2 \\ (x,y,t) &\longmapsto \Phi_Z^t\,(x,y) \end{aligned}$$

defined as the unique local solution of the flow-system of Z

$$\begin{aligned} \frac{\mathrm{d}\Phi_Z^t\,(x,y)}{\mathrm{d}t} &= Z \circ \Phi_Z^t\,(x,y) \\ \Phi_Z^0\,(x,y) &= (x,y)\ . \end{aligned}$$

The Lie formula gives a series expansion, normally convergent near $p \times \{0\}$, in the form

$$\Phi_Z^t = \sum_{m=0}^{\infty} \frac{t^m}{m!} Z \cdot^m \mathrm{Id}\,, \tag{22.18}$$

where $\mathrm{Id}\ :\ (x,y) \mapsto (x,y)$ is the identity of the complex plane. More generally for any $G \in \mathrm{Holo}\,(\mathcal{U})$ the Lie identity holds (locally for all $t \in (\mathbb{C}, 0)$)

$$G \circ \Phi_Z^t = \sum_{m=0}^{\infty} \frac{t^m}{m!} Z \cdot^m G\,. \tag{22.19}$$

In particular G is a first integral of Z if and only if G is constant along every integral curves of Z. If $0 \in \mathcal{U}$, the formula also holds for any formal power series $G \in \mathbb{C}\,[[x,y]]$, the right-hand side belonging to $\mathbb{C}\,[[x,y,t]]$.

Example 3. We compute easily

$$\Phi_{\mathrm{X}_0^\infty}^t\,(x,y) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \left(n! x^{n+1}\ ,\ y\right) = \left(\frac{x}{1-tx}\ ,\ y \exp t\right)$$

For fixed p we perform the maximal analytic continuation of $t \mapsto \Phi_Z^t(p)$ by patching in the appropriate fashion well-chosen solutions to nearby flow systems. The result is a curve parameterization $\Phi_Z^\bullet(p) : S_p \to \mathcal{U}$ from a connected Riemann surface S_p onto the **integral curve** of Z passing through p. One encounters also the terminology «orbit of p under (the flow of) Z», which is not used as such here but helps explaining some terminology we employ below for changes of coordinates. The parameterization itself may be referred to as the **trajectory** of Z passing through p. It is the natural parameterization of the integral curve by the time of Z. Two vector fields Z and X on $\mathcal{U}$ have same integral curves if and only if

$$Z = UX$$

for some $U \in \mathrm{Holo}(\mathcal{U})^\times$.

Notice that according to (22.19) the following identity holds (locally for all $t \in (\mathbb{C}, 0)$)

$$Z \cdot \Phi_Z^t = Z \circ \Phi_Z^t .$$

A **singularity** (or stationary point) of Z is a point $p \in \mathcal{U}$ such that $Z(p) = 0$. The set of singular points of Z is written $\mathrm{Sing}(Z)$. Outside $\mathrm{Sing}(Z)$ we say that Z is **regular**. Singularities of Z are the only constant trajectories.

Example 4. The only singularity of X_0^∞ is located at 0. It is therefore isolated. Any other integral curve, distinct from $\{x = 0\ ,\ y \neq 0\}$, coincides with a level $\{H = \mathrm{cst}\}$ of the first integral $H : (x, y) \in \mathbb{C}^\times \times \mathbb{C} \mapsto y \exp \frac{1}{x}$. For arbitrary $p = (x, y) \in \mathbb{C}^2$, the trajectory of the integral curve passing through p is defined for times belonging to $S_p = \mathbb{C} \setminus \left\{\frac{1}{x}\right\}$. See also Example 3.

22.5.4 Holomorphic Foliations

We wish to describe the holomorphic **singular foliation** $\mathcal{F} = \mathcal{F}_Z$ associated with Z on $\mathcal{U}$. Roughly speaking it is the partition of $\mathcal{U}$ into singular points and leaves, the latter corresponding to non-constant integral curves (without referring to a particular parameterization). There is a small catch, though, when Z is singular at p but the singularity is not isolated. In that case we can factor out a greatest common divisor in the components of Z, yielding a (local) decomposition $Z = UX$, where $U \in \mathrm{Holo}\left(\mathbb{C}^2, p\right)$ vanishes at p and $X \in \mathfrak{X}\left(\mathbb{C}^2, p\right)$ either is regular or has an isolated singularity at p. All such eventually isolated singularities $p \in \mathrm{Sing}(X)$ form the **singular set** $\mathrm{Sing}(\mathcal{F})$ of $\mathcal{F}$. By each point $p \notin \mathrm{Sing}(\mathcal{F})$ passes a unique **leaf** $\mathcal{L}_p$ of the foliation, which is the maximal connected smooth complex curve tangent to Z and containing p. It is obtained by gluing integral curves of corresponding local vector fields X.

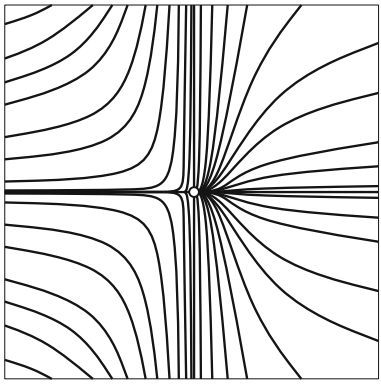


Fig. 22.4 Some leaves of the foliation induced by X_0^∞ on $\mathbb{R}^2$. Mixed saddle (*left*) and node (*right*) behaviors are apparent

Two foliations $\mathcal{F}_Z$ and $\mathcal{F}_X$ are identical if and only if there exist $V,\ W \in \text{Holo}(\mathcal{U}) \setminus \{0\}$ such that $VZ = WX$. If Z has only isolated singularities in $\mathcal{U}$ then the conditions boils down to $Z = UX$ for some $U \in \text{Holo}(\mathcal{U})^\times$.

The **restriction** of $\mathcal{F}$ to a subdomain $\mathcal{V} \subset \mathcal{U}$ is the foliation

$$\mathcal{F}|_{\mathcal{V}}$$

of $\mathcal{V}$, with singularities located at points of $\mathcal{V} \cap \text{Sing}(\mathcal{F})$ and whose leaves are the connected components of $\mathcal{V} \cap \mathcal{L}_p$ for each $p \in \mathcal{V}$.

Example 5. Take $\mathcal{U} := \mathbb{C}^2$ and $Z : (x, y) \mapsto yX_0^\infty(x, y)$. The vector field Z has the line $\{y = 0\}$ for singular set. Yet $\mathcal{F}_Z = \mathcal{F}_{X_0^\infty}$ has only one singularity at 0, all other leaves are either of the form $\left\{y = h \exp \frac{-1}{x}\ ,\ x \neq 0\right\}$ for some $h \in \mathbb{C}$ or coincide with $\{x = 0\ ,\ y \neq 0\}$. See Fig. 22.4.

Some leaves play a special role for the foliation. A **separatrix** of $\mathcal{F}$ at the singularity $p \in \mathcal{U}$ is a leaf whose adherence in $\mathcal{U}$ is perhaps a singular analytic curve containing p.

Example 6. The foliation induced by X_0^∞ on $\left(\mathbb{C}^2, 0\right)$ has exactly two separatrices, which are the connected components of $\{xy = 0\} \setminus \{0\}$.

22.5.5 Changes of Coordinates

We define the action of $\text{Diff}\left(\tilde{\mathcal{U}} \to \mathcal{U}\right)$ by change of coordinates on vector fields. On the source space $\mathcal{U}$ of the vector field, $\Psi \in \text{Diff}\left(\tilde{\mathcal{U}} \to \mathcal{U}\right)$ acts as a usual mapping by composition. The action on the range space $\text{T}\tilde{\mathcal{U}}$ is induced by the direct product $\Psi \oplus \text{D}\Psi$, sending $(p, \mathbf{v}) \in \tilde{\mathcal{U}} \times \mathbb{C}^2$ to $(\Psi(p), \text{D}\Psi(p)(\mathbf{v}))$. We write $\Psi^* Z$ the element of $\mathfrak{X}(\mathcal{U})$ defined in such a way that the following diagram commutes

$$\begin{array}{ccc} \tilde{\mathcal{U}} & \xrightarrow{\Psi} & \mathcal{U} \\ \downarrow{\scriptstyle \Psi^* Z} & & \downarrow{\scriptstyle Z} \\ \mathrm{T}\tilde{\mathcal{U}} & \xrightarrow[\Psi \oplus \mathrm{D}\Psi]{} & \mathrm{T}\mathcal{U} \end{array}$$

that is

$$\Psi^* Z = (\mathrm{D}\Psi)^{-1} (Z \circ \Psi) \ . \tag{22.20}$$

The vector field $\Psi^* Z$ is called the **pullback** of Z by Ψ. In that situation trajectories of $\Psi^* Z$ are mapped to trajectories of Z, leaving the natural time unchanged (locally for all $t \in (\mathbb{C}, 0)$):

$$\Psi \circ \Phi^t_{\Psi^* Z} = \Phi^t_Z \circ \Psi \ . \tag{22.21}$$

We say that $Z \in \mathfrak{X}(\mathcal{U})$ and $\tilde{Z} \in \mathfrak{X}(\tilde{\mathcal{U}})$ are **analytically conjugate** if there exists $\Psi \in \mathrm{Diff}(\tilde{\mathcal{U}} \to \mathcal{U})$ such that $\tilde{Z} = \Psi^* Z$. This is equivalent to the **conjugacy equation**

$$\tilde{Z} \cdot \Psi = Z \circ \Psi \tag{22.22}$$

being satisfied.

We say that Z and $\tilde{Z}$ are **analytically orbitally equivalent** when there exists $\tilde{U} \in \mathrm{Holo}(\tilde{\mathcal{U}})^\times$ such that Z is conjugate to $\tilde{U}\tilde{Z}$. This means that Z is conjugate to a vector field with same integral curves as $\tilde{Z}$, in other words that integral curves of $\tilde{Z}$ are mapped under Ψ onto integral curves of Z, yet the natural time changes in general.

Naturally all these notions can be germified. We then speak of **local conjugacy** and **local orbital equivalence**. If (22.22) holds at a formal level for some $\Psi \in \widehat{\mathrm{Diff}}(\mathbb{C}^2, 0)$, then Z is **formally conjugate** to $\tilde{Z}$. If Z is formally conjugate to some $\tilde{U}\tilde{Z}$ with $\tilde{U} \in \mathbb{C}[[x, y]]^\times$, then Z is **formally orbitally equivalent** to $\tilde{Z}$.

In case Z is (analytically, locally) orbitally equivalent to $\tilde{Z}$, a bijection Ψ realizing the equivalence maps $\mathrm{Sing}(\mathcal{F}_{\tilde{Z}})$ onto $\mathrm{Sing}(\mathcal{F}_Z)$ and sends each leaf of $\mathcal{F}_{\tilde{Z}}$ onto a leaf of $\mathcal{F}_Z$. We say the foliations $\mathcal{F}_Z$ and $\mathcal{F}_{\tilde{Z}}$ are **(analytically, locally) conjugate** and define

$$\Psi^* \mathcal{F}_Z := \mathcal{F}_{\Psi^* Z} \ .$$

We extend the terminology in the obvious way for formal diffeomorphisms, speaking of **formal conjugacy** between foliations.

Example 7. If one lets $\widehat{\Psi}$ be $(x, y) \mapsto (x, y - \hat{\mathfrak{s}}(x))$, where $\hat{\mathfrak{s}}$ is the formal solution of (22.9), then $\mathrm{E}_0 = \widehat{\Psi}^* \mathrm{X}_0^\infty$ since

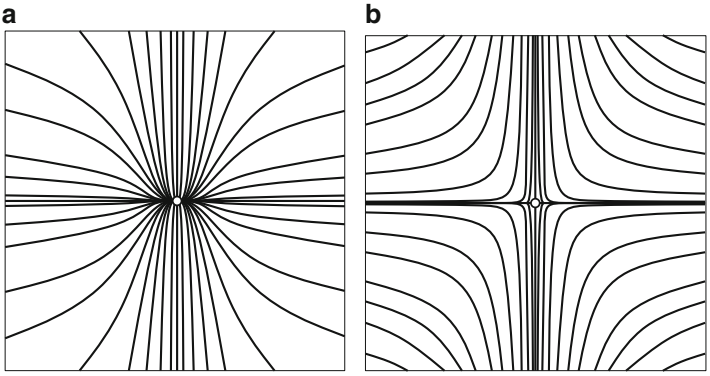


Fig. 22.5 Real foliation of the vector field Z when $l \in \mathbb{R}^{\times}$. (**a**) A node ($l > 0$); (**b**) a saddle ($l < 0$)

$$\mathrm{X}_0^{\infty} \circ \widehat{\Psi}(x,y) = x^2 \frac{\partial}{\partial x} + (y - \hat{\mathfrak{s}}(x)) \frac{\partial}{\partial y}$$

$$\mathrm{E}_0 \cdot \widehat{\Psi}(x,y) = x^2 \frac{\partial}{\partial x} + \left(y - x^2 \hat{\mathfrak{s}}'(x) - x^2\right) \frac{\partial}{\partial y} .$$

22.5.6 *Leaves Spaces*

The **leaves space** of $\mathcal{F}$ on $\mathcal{U}$ is, as a set, the quotient

$$\Omega_{\mathcal{F}} := {}^{(\mathcal{U} \setminus \mathrm{Sing}(\mathcal{F}))}/_{\mathcal{F}}$$

where two points of $\mathcal{U} \setminus \mathrm{Sing}(\mathcal{F})$ are equivalent when they belong to the same leaf of $\mathcal{F}$. It is endowed with the quotient topology, which (generally) is non-Hausdorff (Fig. 22.5).

Example 8. Take $\mathcal{U} := \mathbb{C}^2$ and the linear vector field $Z : (x,y) \mapsto lx\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}$ with $l \in \mathbb{C}^{\times}$. Both components of $\{xy = 0\} \setminus \{0\}$ is a separatrix of $\mathcal{F}_Z$; set $\mathcal{L}_x := \{x = 0\ ,\ y \neq 0\}$ and $\mathcal{L}_y := \{x \neq 0\ ,\ y = 0\}$. The foliation $\mathcal{F}_Z$ has a singularity at 0, whose dynamical type and leaves space structure depend on the rationality of l.

$l = \frac{p}{q} \in \mathbb{Q}_{>0}$: **resonant node**. Each leaf corresponds to a level $\{H = \mathrm{cst}\}$ of the rational first integral $H : (x,y) \in \mathcal{U} \setminus \{0\} \mapsto x^q y^{-p}$ (which is a connected complex curve). The leaves space Ω is homeomorphic to the Riemann sphere $H(\mathcal{U}) = \overline{\mathbb{C}}$, where $\mathcal{L}_x$ corresponds to 0 and $\mathcal{L}_y$ to ∞. Taking $\mathcal{U} = \left(\mathbb{C}^2, 0\right)$ instead of the whole plane does not change Ω.

$l \in \mathbb{R}_{>0} \setminus \mathbb{Q}$: **quasi-resonant node**. The vector field Z has no meromorphic first integral, although the multivalued function $H : (x,y) \mapsto xy^{-l}$

satisfies $Z \cdot H = 0$ algebraically. Apart from $\mathcal{L}_x$ and $\mathcal{L}_y$, each leaf corresponds to a "level" $\{H(x,y) = h\}$ of H, $h \in \mathbb{C}^\times$, which cannot be closed in $\mathcal{U} \setminus \{(0,0)\}$. The leaves space Ω is homeomorphic to the quotient of $\overline{\mathbb{C}}$ by the action of the irrational rotation $h \mapsto h \exp 2i\pi l$, whose orbits are dense in circles $\{|h| = \text{cst}\}$. The quotient cannot be Hausdorff, and actually no two leaves in the same circle $\{|h| = \text{cst}\}$ can have separating neighborhoods in Ω. Taking $\mathcal{U} = \left(\mathbb{C}^2, 0\right)$ instead of the whole plane does not change Ω.

$l = -\frac{p}{q} \in \mathbb{Q}_{<0}$: **resonant saddle**. Each leaf corresponds to a level of the polynomial first integral $H : (x,y) \in \mathcal{U} \mapsto x^q y^p$ (which is a connected complex curve) save for $H^{-1}(0) = \{0\} \cup \mathcal{L}_x \cup \mathcal{L}_y$. The leaves space Ω is isomorphic as a set to the punctured line $\mathbb{C}^\times$ joined to two abstract points $\left\{0_x, 0_y\right\}$. As a topological space it is homeomorphic to $\mathbb{C}^\times \cup \left\{0_x, 0_y\right\}$ equipped with the following topology: a non-empty subset $U \subset \mathbb{C}^\times \cup \left\{0_x, 0_y\right\}$ is open if and only if $U \subset \mathbb{C}^\times$ or $U \cap \mathbb{C}^\times$ is an open, punctured neighborhood of $0 \in \mathbb{C}$. This space is not Hausdorff as 0_x and 0_y have no separating neighborhoods. Taking $\mathcal{U} = \left(\mathbb{C}^2, 0\right)$ results in a smaller leaves space Ω where the role played by $\mathbb{C}^\times$ is replaced by $(\mathbb{C}, 0) \setminus \{0\}$.

$l \in \mathbb{R}_{<0} \setminus \mathbb{Q}$: **quasi-resonant saddle**. It is a composite situation that can be obtained from resonant saddles by quotienting out the action of the irrational rotation, as for quasi-resonant nodes. Details are left to the reader.

$l \notin \mathbb{R}$: **hyperbolic singularity**. Apart from $\mathcal{L}_x$ and $\mathcal{L}_y$, each leaf corresponds to a "level" $\left\{xy^{-l} = h\right\}$, $h \in \mathbb{C}^\times$, whose adherence in $\mathcal{U} \setminus \{0\}$ contains $\mathcal{L}_x \cup \mathcal{L}_y$. The punctured leaves space $\Omega \setminus \left\{\mathcal{L}_x\ ,\ \mathcal{L}_y\right\}$ is homeomorphic to the quotient of $\mathbb{C}^\times$ by the action of the linear map $h \mapsto h \exp 2i\pi l$, which is a torus. Yet Ω is not Hausdorff as $\mathcal{L}_x$ and $\mathcal{L}_y$ cannot be separated from any other leaf. Taking $\mathcal{U} = \left(\mathbb{C}^2, 0\right)$ instead of the whole plane does not change Ω.

It is not always possible to endow the leaves space with an analytic atlas, although Ω is locally homeomorphic to an open set of $\mathbb{C}$. Indeed around a regular point $p \notin \text{Sing}(\mathcal{F})$ we can apply the rectification theorem to some regular $X \in \mathfrak{X}\left(\mathbb{C}^2, p\right)$ defining the foliation: there exists a local diffeomorphism $\Psi : \left(\mathbb{C}^2, 0\right) \to \left(\mathbb{C}^2, p\right)$ such that $\Psi^* X = \frac{\partial}{\partial x}$. Hence the leaves of $\mathcal{F}|_{\left(\mathbb{C}^2, p\right)}$ are images of small "horizontal" discs included in $\{y = \text{cst}\}$. A pair $(\mathcal{D}, \Psi)$ of a domain $\mathcal{D} = \left(\mathbb{C}^2, 0\right)$ and a map $\Psi \in \text{Diff}(\mathcal{D} \to (\mathbb{C}^n, p))$ sending X to $\frac{\partial}{\partial x}$ is called a **rectifying chart** (or **flow-box**) for $\mathcal{F}$. Since level sets of

$$\begin{aligned} H\ :\ \Psi(\mathcal{D}) &\longrightarrow \mathbb{C} \\ \Psi(x,y) &\longmapsto y \end{aligned}$$

coincide with leaves of $\mathcal{F}|_{\Psi(\mathcal{D})}$, the local leaves space has a holomorphic parameterization

$$\Omega_{\mathcal{F}|_{\Psi(\mathcal{D})}} \simeq H\left(\Psi\left(\mathcal{D}\right)\right) = (\mathbb{C}, 0) \ .$$

Yet transition maps between two rectifying charts may fail to be either bijective or holomorphic.

Instead of trying to force a superfluous analytic structure on $\Omega_{\mathcal{F}}$, it will be sufficient for our purposes to use the sheaf $\mathrm{FirstIntegral}\left(\bullet\right)$ of first integrals of $\mathcal{F}$. Any holomorphic function $H \in \mathrm{Holo}\left(\mathcal{V}\right)$ on a subdomain $\mathcal{V} \subset \mathcal{U}$, which is a first integral of any vector field locally defining $\mathcal{F}$, is called a **first integral** of $\mathcal{F}$ on $\mathcal{V}$. Level sets of H are saturated by $\mathcal{F}|_{\mathcal{V}}$. When a connected level set of a non-constant first integral H does not contain a singularity of $\mathcal{F}$ then it coincides with a single leaf of $\mathcal{F}|_{\mathcal{V}}$. We say that H has **connected fibers** when it is non-constant and every level set is connected. (Notice this property is fulfilled and used in the previous Example 8.) First integrals with connected fibers will play a central role in the sequel, as the algebra $\mathrm{FirstIntegral}\left(\mathcal{V}\right)$ is functionally generated by some (and in fact any) first integral H with connected fibers. The mapping

$$\begin{array}{ccc} \mathrm{Holo}\left(H\left(\mathcal{V}\right)\right) & \longrightarrow & \mathrm{FirstIntegral}\left(\mathcal{V}\right) \\ f & \longmapsto & f \circ H \end{array}$$

is indeed bijective: any first integral factors uniquely and holomorphically through H.

Example 9. The function $H\left(x, y\right) := xy^2$ is a first integral of the resonant linear saddle $X\left(x, y\right) := 2x\frac{\partial}{\partial x} - y\frac{\partial}{\partial y}$. Its fibers are the connected Riemann surfaces $\left\{xy^2 = c\right\}$, $c \in \mathbb{C}$. Notice that the two branches of $\{xy = 0\}$ are disconnected when the singularity 0 is removed from them.

The equation $X \cdot F = 0$ has formal solutions $F\left(x, y\right) = \sum_{n,m \geq 0} f_{n,m} x^n y^m$ satisfying $f_{n,m} = 0$ if $2n \neq m$, while each $f_{n,2n}$ is free to chose in $\mathbb{C}$. Therefore $F\left(x, y\right) = \sum_{n \geq 0} f_{n,2n} \left(xy^2\right)^n = f\left(xy^2\right)$ where $f\left(t\right) := \sum_{n \geq 0} f_{n,2n} t^n$.

22.5.7 Moduli Spaces, Normal Forms

The local rectification theorem asserts the existence of a single equivalence class for local conjugacy near a regular point. One important goal in the theory of vector fields is therefore to understand qualitative behaviors near *singular* points up to diverse conjugacy notions (and their orbital counterparts for foliations). This means to describe the quotients, called **moduli spaces**, of $\mathfrak{X}\left(\mathbb{C}^2, 0\right)$ under the action by conjugacy of $\mathrm{Diff}\left(\mathbb{C}^2, 0\right)$ or $\widehat{\mathrm{Diff}}\left(\mathbb{C}^2, 0\right)$, respectively, i.e. to perform the (local, formal) classification by identifying a complete set of objects invariant under conjugacy. We call such objects (**local, formal**) **invariants**.

An important invariant is the following. Take $p \in \mathrm{Sing}\,(Z)$ and consider the linear part of Z at p, i.e. the linear mapping $\mathrm{D}Z\,(p)$. Then its spectrum, written $\mathrm{Spec}\,(Z,p)$ for the sake of simplicity, is invariant under formal conjugacy:

$$\forall \Psi \in \mathrm{Diff}\,((\mathbb{C}^n,\tilde{p}) \to (\mathbb{C}^n,p)) \qquad \mathrm{Spec}\left(\Psi^* Z,\tilde{p}\right) = \mathrm{Spec}\,(Z,p)\ .$$

Besides, let $\Lambda\,(Z,p)$ be the equivalence class of $\mathrm{Spec}\,(Z,p)$ under

$$\{l_1,l_2\} \leftrightsquigarrow \left\{\tilde{l}_1,\tilde{l}_2\right\} \quad \Longleftrightarrow \quad \left(\exists c \in \mathbb{C}^\times\right)\ :\ \{l_1,l_2\} = c\left\{\tilde{l}_1,\tilde{l}_2\right\}\ .$$

Then $\Lambda\,(Z,p)$ is invariant under formal orbital equivalences.

Remark 7. The quotient space is naturally isomorphic to the double cover of $\mathbb{P}_1\,(\mathbb{C})$ ramified over $[1:1]$ and $[-1:1]$, obtained under the obvious $\frac{\mathbb{Z}}{2\mathbb{Z}}$-action $[l_1:l_2] \mapsto [l_2:l_1]$ in homogeneous coordinates. The quotient is a smooth, compact Riemann surface of genus 0 parameterized by $[l_1:l_2] \mapsto \left[l_1^2+l_2^2 : l_1 l_2\right]$, therefore itself a conformal projective line $\mathbb{P}_1\,(\mathbb{C})$.

Generically this is the only invariant, as if $l_1 l_2 \neq 0$ and $\frac{l_1}{l_2} \notin \mathbb{R}$ then the vector field is **hyperbolic** and Poincaré's theorem guarantees that Z is locally **linearizable**: there exists $\Psi \in \mathrm{Diff}\left(\mathbb{C}^2,0\right)$ such that $\Psi^* Z = l_1 x\frac{\partial}{\partial x} + l_2 y\frac{\partial}{\partial y}$. This particularly means that the local leaves space near a hyperbolic singularity is a conformal torus with two points in the adherence of every others, corresponding to the two local separatrices passing through p (see Example 8).

To be altogether correct, we need to mention that the group $\widehat{\mathrm{Diff}}\left(\mathbb{C}^2,0\right)$ does not really act on $\mathfrak{X}\left(\mathbb{C}^2,0\right)$. If $\Psi \in \widehat{\mathrm{Diff}}\left(\mathbb{C}^2,0\right)$ and $Z \in \mathfrak{X}\left(\mathbb{C}^2,0\right)$, there is no reason why $\Psi^* Z$ should be a holomorphic vector field, even though (22.20) defines a perfectly valid vector field with formal power series components. Yet being formally conjugate defines an equivalence relation, and we write resulting quotients as if they were quotients of a group action, for convenience sake.

The complete invariants we seek should differ in nature from simply stating "the equivalence class in the quotient". We particularly wish to build non-trivial bijective mappings between the various flavors of moduli spaces and some functional spaces. Classifying vector fields is out of reach in such a general form, although it can be carried out for smaller classes of vector fields. We take $\mathfrak{F} \subset \mathfrak{X}\left(\mathbb{C}^2,0\right)$ and write respectively

$$\mathrm{Mod}_{\mathrm{loc}}\,(\mathfrak{F}) := {}^{\mathfrak{F}}\!/_{\mathrm{Diff}\left(\mathbb{C}^2,0\right)}$$

$$\mathrm{Mod}_{\mathrm{for}}\,(\mathfrak{F}) := {}^{\mathfrak{F}}\!/_{\widehat{\mathrm{Diff}}\left(\mathbb{C}^2,0\right)}$$

the corresponding moduli spaces. Since formal conjugacy is weaker than local conjugacy there is a canonical onto mapping $\mathrm{Mod}_{\mathrm{loc}}\,(\mathfrak{F}) \twoheadrightarrow \mathrm{Mod}_{\mathrm{for}}\,(\mathfrak{F})$, and this is why in practice we fix a formal equivalence class within which the local classification is conducted. The notation

$$[Z]_\star \in \mathrm{Mod}_\star(\mathfrak{F}) \qquad , \star \in \{\mathrm{for}, \mathrm{loc}\}$$

stands for the equivalence class of $Z \in \mathfrak{F}$ with respect to corresponding class of conjugacy.

Let $\mathrm{Mod}_\star(\mathfrak{F}) = {}^{\mathfrak{F}}/_{G}$ stand for one of the above quotients and let Ω be a set. We call

- an injective mapping $C : \mathrm{Mod}_\star(\mathfrak{F}) \to \Omega$ a **classification** of $\mathrm{Mod}_\star(\mathfrak{F})$ (it is **complete** when surjective),
- a surjective mapping $R : \Omega \to \mathrm{Mod}_\star(\mathfrak{F})$ a **realization** of $\mathrm{Mod}_\star(\mathfrak{F})$.

The best way to realize a moduli space $\mathrm{Mod}_\star(\mathfrak{F})$ in a concrete form is to find a collection of (**local, formal**) **normal forms** $\mathrm{NF}_\star(\mathfrak{F}) \subset \mathfrak{F}$ satisfying the first two following properties:

Versality	The natural map ${}^{\mathrm{NF}_\star(\mathfrak{F})}/_{G} \to {}^{\mathfrak{F}}/_{G}$ is bijective
Uniqueness	There exist $\nu \in \mathbb{N}$ and a smooth $\mathbb{C}^\nu$-action on $\mathrm{NF}_\star(\mathfrak{F})$ such that for any $Z \in \mathrm{NF}_\star(\mathfrak{F})$ the whole equivalence class $[Z] \in \mathrm{Mod}_\star(\mathfrak{F})$ is included in a single orbit.
Simplicity	Although the notion of "simple" expression is primarily opinion-based, it is generally expected that elements of $\mathrm{NF}_\star(\mathfrak{F})$ have "simple" expressions in some "natural" basis of the tangent bundle.

Remark 8. The clause of uniqueness states that a normal form $Z \in \mathrm{NF}_\star(\mathfrak{F})$ is unique "up to a finite-dimensional space". A notion of smoothness on spaces of germs (endowed with a convenient locally convex topology) adapted to this context can be found, for instance, in [34]. Once these normal forms are given it is in general straightforward to refine the study and pinpoint unique representatives for a given equivalence class. This work can be messy, though in practice seldom reaching further than linear algebra. In this text we stick to finite-dimensional uniqueness.

Example 10. Let $\mathfrak{H} := \left\{Z \in \mathfrak{X}\left(\mathbb{C}^2, 0\right) : Z \text{ be hyperbolic at } 0\right\}$. Then

$$\begin{aligned} C_{\mathrm{loc}} : \mathrm{Mod}_{\mathrm{loc}}(\mathfrak{H}) &\longrightarrow {(\mathbb{C}^\times)^2}/_{\frac{\mathbb{Z}}{2\mathbb{Z}}} \\ [Z]_{\mathrm{loc}} &\longmapsto \mathrm{Spec}(Z, 0) \end{aligned}$$

is a classification for local conjugacy, which is not complete. Injective realizations are given by

$$\begin{aligned} R_{\mathrm{loc}} : (\mathbb{C}\backslash\mathbb{R}) \times \mathbb{C}^\times &\longrightarrow \mathrm{Mod}_{\mathrm{loc}}(\mathfrak{H}) \\ (\rho, l_2) &\longmapsto \left[l_2 \rho x \frac{\partial}{\partial x} + l_2 y \frac{\partial}{\partial y}\right]_{\mathrm{loc}} . \end{aligned}$$

Normal forms are given by

$$\mathrm{NF}_{\mathrm{loc}}\left(\mathfrak{H}\right) := \left\{ l_1 x\frac{\partial}{\partial x} + l_2 y\frac{\partial}{\partial y} \ : \ (l_1, l_2) \in \left(\mathbb{C}^{\times}\right)^2 \ , \ \tfrac{l_1}{l_2} \notin \mathbb{R} \right\} \ .$$

Example 11. Theorem 2 asserts

$$\mathrm{Mod}_{\mathrm{loc}}\left(\mathrm{Affine}\,(1)\right) \simeq \overline{\mathbb{N}}$$

with normal forms

$$\mathrm{NF}_{\mathrm{loc}}\left(\mathrm{Affine}\,(1)\right) := \left\{ \mathrm{X}^{\kappa}_{\bullet} \ : \ \kappa \in \overline{\mathbb{N}} \right\} \ .$$

22.6 General Saddle-Node Bifurcations

From now on we deal with the general case of a holomorphic germ of a planar saddle-node bifurcation. For the bifurcation value of the parameter $\lambda \in \Lambda$, which we conveniently locate at the origin of a complex affine space of which Λ is a (sufficiently small) domain, the vector field Z_0 is of **saddle-node** type near, say, the origin of $\mathbb{C}^2$, that is:

- 0 is an isolated singularity of Z_0,
- the differential at 0 of the vector field has exactly one non-zero eigenvalue or, with notations introduced earlier, $\mathrm{Spec}\,(Z, 0) = \{0, l_2\}$ for some $l_2 \in \mathbb{C}^{\times}$ (the singularity is elementary degenerate).

To stick to general terminology, a (holomorphic germ of a) parametric family of (germs at $0 \in \mathbb{C}^2$ of) vector fields $Z_{\bullet} = (Z_{\lambda})_{\lambda \in \Lambda}$ is called a holomorphic germ of an **unfolding** of Z_0. We study in detail only "generic" unfoldings, those which possess the "right number" of parameters to encode the bifurcation structure. Let us be more specific.

Definition 1. Let $k \in \mathbb{N}$ be given. A **generic unfolding of codimension** k is a germ of an unfolding $Z_{\bullet}$ for which the following conditions hold.

- There exist $\Lambda = \left(\mathbb{C}^k, 0\right)$ and $\mathcal{U} = \left(\mathbb{C}^2, 0\right)$ such that $(\lambda, x, y) \mapsto Z_{\lambda}\,(x, y)$ is holomorphic on $\Lambda \times \mathcal{U}$, and the vector field Z_0 has only one singularity in $\mathcal{U}$.
- For a dense open set $\widehat{\Lambda} \subset \Lambda$ and all $\lambda \in \widehat{\Lambda}$ the vector field Z_{λ} has exactly $k + 1$ (distinct) singularities in $\mathcal{U}$, which are all hyperbolic and merge to 0 as $\lambda \to 0$.
- $\lambda \in \widehat{\Lambda} \mapsto \mathrm{Sing}\,(Z_{\lambda})$ is injective.

The limiting saddle-node Z_0 has codimension k.

Families with $\tilde{k} > k$ parameters can be dealt with by changing the parameters (say, using the implicit function theorem, after desingularization if need be) in such a way

that the first k components govern the location of the singularities, while the rest (seen as extra parameters) do not move them around. To be more specific, we mean that each fiber of $\lambda \in \widehat{\Lambda} \mapsto \mathrm{Sing}\,(Z_\lambda)$ is included in a single fiber of the natural projection on the first k components. All results presented here hold also for these extra-parametric generic unfoldings, as it will appear clearly that the constructions depend holomorphically on extraneous parameters.

Families $\left(\tilde{Z}_{\tilde{\lambda}}\right)_{\tilde{\lambda}\in\tilde{\Lambda}}$ having singularities either generically elementary degenerate (e.g., coalescing saddle-nodes) or non-degenerate but reached multiple times can be studied through (extra-parametric) generic unfoldings $(Z_\lambda)_{\lambda\in\Lambda}$ by specializing values $\lambda = \phi\left(\tilde{\lambda}\right)$.

We postpone a formal definition of conjugacy/orbital equivalence between unfoldings till the end of this section. Just keep in mind that we do not allow parameter changes involving the spatial coordinates (x, y).

Affine unfoldings, as detailed in Sect. 22.4, suggest that dynamical questions regarding saddle-node bifurcations can be understood from the local classification of generic unfoldings. The classification of unfoldings contains the classification of Z_0 by specialization, which is why we deliberately elude presenting this degenerate situation in detail. Yet as the strategies adopted to address saddle-node singularities will serve us well, we present them briefly as a steppingstone to the unfolded case. We refer the reader to the works cited below for a comprehensive study of the subject.

The analytic unstable manifold of Z_0, tangent at 0 to the eigenspace associated with l_2, is called the **strong separatrix**. The other eigenspace corresponds to a "formal separatrix" called the **weak separatrix** (generically divergent [23], always summable in the sense of Borel [11]). We say that a saddle-node is convergent or divergent according to the nature of its weak separatrix.

The formal orbital classification was performed by Poincaré and Dulac [6, 7], yielding polynomial normal forms. It was known from the very beginning that the formal conjugacy cannot always converge, and as a matter of fact divergence is the rule. After some inspiring works by Birkhoff [1] on local classification of resonant diffeomorphisms, a complete local orbital classification was achieved in the early 1980s by Martinet and Ramis [21]. At about the same time Bruno [4] presented formal normal forms for saddle-node vector fields. These works were complemented with a complete local classification in the early 2000s simultaneously by Meshcheryakova and Voronin [36] ($k = 1$) and by Teyssier [31] for the general case. These studies reveal that classifying vector fields can be dissociated into two independent process:

1. classify the orbital part (the foliation),
2. classify the "time" (the vector field for fixed foliation).

Item (2) is a linear problem, simpler to deal with. Hence in the current introduction we only present how orbital classification is achieved. Also, for the sake of keeping notations light, only the case $k = 1$ is presented below.

The foundational viewpoint introduced in [21] bridges the gaps between classification on the one hand, dynamical and analytical properties on the other hand.

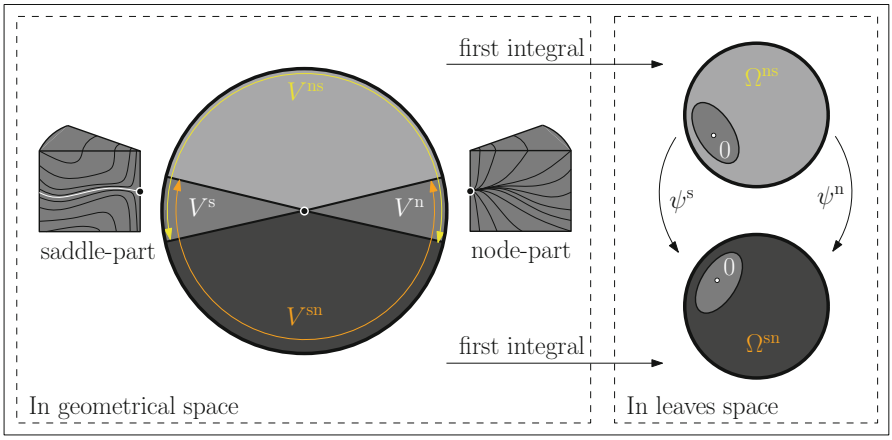


Fig. 22.6 The sectorial decomposition of a saddle-node of codimension 1. The *white* invariant curve in the saddle-part (sectorial separatrix) is mapped to 0 in leaves space

It consists in interpreting the orbital invariants as transition maps defining a holomorphic atlas of some analytic space Ω_0 closely related to the space of leaves of $\mathcal{F}_{Z_0}$. The conformal class of Ω_0 is reasonably a complete orbital invariant of Z_0. As illustrated in Fig. 22.6, this space is obtained by a conformal gluing between two Riemann spheres Ω_0^{sn} and Ω_0^{ns}. The spheres correspond to a decomposition of $(\mathbb{C}^2, 0)$ in two overlapping fibered sectors. Each $\Omega_0^\sharp$ is the union of $\mathbb{C}$, the range of a holomorphic sectorial first integral with connected fibers, and ∞, standing for the strong separatrix. It can be arranged that the first integral maps the sectorial weak separatrix of Z_0 to 0. The transition map near 0 (*resp.* ∞) is a germ of diffeomorphism ψ_0^s (*resp.* a translation ψ^n) induced by the inclusions of the sectors in $(\mathbb{C}^2, 0)$. The resulting complex 1-manifold is Martinet–Ramis's *chapelet de sphères*, which we rather refer to as the **orbital necklace** of Z_0.

Remark 9. The necklace Ω_0 is not exactly the space of leaves $\widehat{\Omega}_0$ of $\mathcal{F}_{Z_0}$. For one thing, most leaves accumulate on the strong separatrix so that $\widehat{\Omega}_0$ is far from being Hausdorff: ∞ is not topologically separable from any other point of $\widehat{\Omega}_0$. Secondly $\widehat{\Omega}_0$ is obtained from Ω_0 by modding out the action of the global (analytic) monodromy of the necklace $\psi_0^n \circ \psi_0^s$. This fact is explained in detail later on for unfoldings (Sect. 22.12).

That the pair of transition maps (ψ_0^s, ψ_0^n) classifies, up to a finite dimensional space, the foliation $\mathcal{F}_{Z_0}$ is a consequence of the conformal rigidity of orbital necklaces: their group of diffeomorphisms is small (the automorphism group of $\overline{\mathbb{C}}$ is $\mathbb{P}\mathrm{GL}_2(\mathbb{C})$ acting by homography). Since any conjugacy between foliations induces a diffeomorphism between the respective necklaces, corresponding transition maps must be $\mathbb{C}^\times$-conjugate. Indeed, with the choice of leaves for 0 and ∞, and yet

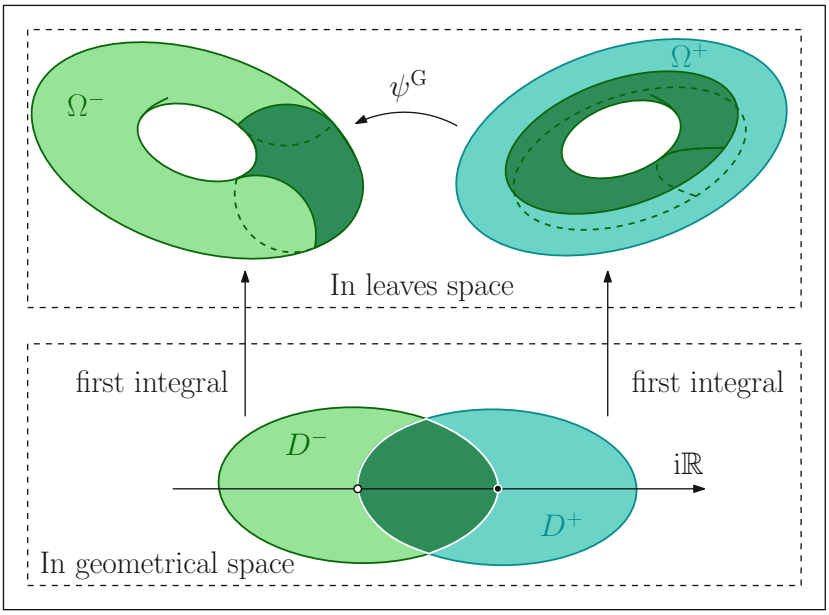


Fig. 22.7 The Glutsyuk connection

another canonical normalization, the only remaining degree of freedom for the choice of $\left(\psi_0^{\mathrm{s}}, \psi_0^{\mathrm{n}}\right)$ is the linear action of $\mathbb{C}^\times$, simultaneously on all spheres.

The first technique based on a deformation of a saddle-node vector field in order to recover Martinet–Ramis invariants was presented in the early 2000s, after an analogous work by Martinet [9, 20] for unfoldings of parabolic diffeomorphisms. Glutsyuk [10] embedded Z_0 in a generic unfolding of codimension 1. Restricting λ to $\widehat{\Lambda}$, so that both singularities are hyperbolic, he let the singularities merge as $\lambda \to 0$. He proved that the domains $D_\lambda^\pm$ of linearization of Z_λ (near the respective singular point $\pm\sqrt{-\lambda}$) overlap and their union D_λ contains a domain $\left(\mathbb{C}^2, 0\right)$ independent on λ. The "orbital link" of $\mathcal{F}_{Z_\lambda}|_{D_\lambda}$ is built by gluing the two spaces of leaves $\Omega_\lambda^\pm$ of $\mathcal{F}_{Z_\lambda}|_{D_\lambda^\pm}$, which are (rigid) conformal tori $\mathbb{C}^\times/\mathbb{Z}$ (see Example 8), through the Glutsyuk connection $\psi_\lambda^{\mathrm{G}}$ coming from the inclusions $D_\lambda^\pm \hookrightarrow D_\lambda$, as summarized in Fig. 22.7.

When $\lambda \to 0$ each torus gets pinched more and more sharply along a meridian, converging towards a sphere $\overline{\mathbb{C}}$ with the points 0 and ∞ identified. This continuous process lifts to the family of connections $\left(\psi_\lambda^{\mathrm{G}}\right)_{\lambda\in\widehat{\Lambda}}$: Martinet–Ramis transition maps $\left(\psi_0^{\mathrm{s}}, \psi_0^{\mathrm{n}}\right)$ can be recovered from the limiting Glutsyuk connection ψ_0^{G}. We refer to Fig. 22.8.

The natural continuation of Glutsyuk's use of the saddle-node bifurcation would be to classify all generic unfoldings. Yet this approach is doomed to fail, because $\widehat{\Lambda}$ is not connected. Glutsyuk's construction critically depends on the local behavior

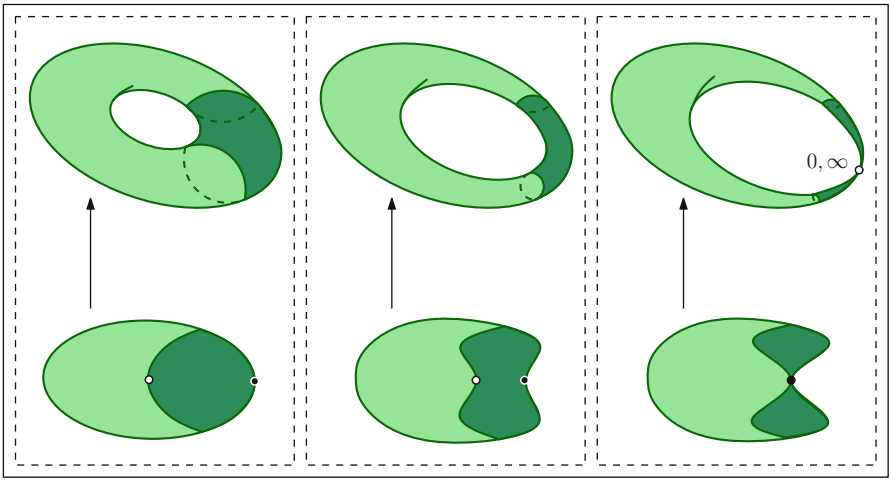


Fig. 22.8 Transition from Glutsyuk's torus to Martinet–Ramis's sphere at the limit $\lambda \to 0$

of hyperbolic singularities, and as such cannot be extended to $\Lambda \backslash \widehat{\Lambda}$. Indeed most (resonant) saddle points are not linearizable and, even if they were, their local leaves space would be a conformal disc, which is not rigid: the automorphism group of germs at 0 of conformal discs coincides with the infinite dimensional group Diff $(\mathbb{C}, 0)$. Roughly speaking, one cannot turn a saddle or node singularity into a hyperbolic singularity. But one can the converse, as done by C. Rousseau and L. Teyssier [28]. By cutting "sectors" attached to hyperbolic points with a special, asymptotically spiraling shape, the restricted foliation can be forced to behave locally very much like a node or a saddle. The process yields an orbital necklace Ω_λ obtained exactly in the same way as for $\lambda = 0$, with an additional linear identification. This approach allows to cover the whole parameter space. The conformal structure of the necklace Ω_λ depends locally analytically on λ and is continuous as $\lambda \to 0$ (in particular sectors for Z_λ converge toward standard sectors for Z_0 in the Hausdorff distance). Therefore unfoldings of foliations are locally classified by families of gluing mappings unfolding the local orbital invariants of Z_0.

The problem of giving a *complete* local classification of generic unfoldings (identifying the total image of the classification) is still open, except for the case $k = 1$ if one compiles the results of [24, 27]. The realization problem is double:

1. realize, for fixed λ, a given necklace as the leaves space of some Z_λ,
2. glue all Z_λ for $\lambda \in \Lambda$ to form an unfolding of Z_0.

Item (1) poses no specific problem and can be dealt with in the usual manner using tools borrowed from complex geometry, by building an abstract almost-complex realization, then invoking Newlander–Nirenberg theorem to incarnate it as a germ of a foliated analytic manifold. On the contrary Item (2) is linked with the combinatorial structure of the covering of the parameter space by (contractible) open cells on which the invariants $\lambda \mapsto \mathfrak{m}_\lambda$ of the unfoldings are holomorphic.

This decomposition is not trivial: it is indeed impossible to perform the previous sectorial decomposition uniformly for all values of λ. The reason is the following: the transfiguration of a saddle point into a sectorial saddle-like hyperbolic point cannot be pursued after a certain point, corresponding to parameter values for which a saddle-like singularity tends to a genuine node. A node cannot be tricked into believing it behaves like a saddle. One must therefore deal with finite families $\left((\mathfrak{m}_\lambda)_{\lambda\in\Lambda_\ell}\right)_\ell$ of unfoldings of invariants of Z_0, where $\Lambda = \bigcup_\ell \mathrm{adh}\left(\Lambda^\ell\right)$. On neighboring intersections $\Lambda^\ell \cap \Lambda^{\tilde{\ell}}$ all singularities are hyperbolic and the configuration is that of a Glutsyuk deformation. The invariants $\mathfrak{m}_\bullet^\ell$ and $\mathfrak{m}_\bullet^{\tilde{\ell}}$ must relate to Glutsyuk tori decomposition, since all three objects encode the same leaves space and the same underlying dynamics. Expressing this identity yields necessary **compatibility conditions** (Sect. 22.12.3) that $\mathfrak{m}_\bullet^\ell$ and $\mathfrak{m}_\bullet^{\tilde{\ell}}$ must obey, as explained clearly in [26]. In the case $k = 1$ the compatibility conditions guarantee that Newlander–Nirenberg theorem applies in parameter space. Corresponding conditions for $k > 1$ have not been written down and proved sufficient, although there is little doubt that they are. We do not present further details of this approach.

Another way of achieving a complete classification would be to describe a collection of normal forms for generic unfoldings. Normal forms have been devised recently by Schäfke and Teyssier [30] for convergent saddle-nodes. These families can be unfolded to families of normal forms [29], in the case of **pure convergence**: every member Z_λ of the unfolding has a heteroclinic connection or, equivalently, sectorial weak separatrices patch continuously for all $\lambda \in (\mathbb{C}, 0)$. This method leaves open the generic case where there are no homoclinic connection in the unfolding, although general normal forms are expected to be worked out soon.

To conclude this introduction we give a precise definition of what changes of variables and parameters we allow between unfoldings. Section 22.5 recalled the diverse notions of (formal, local) conjugacy and orbital equivalence between vector fields in $\mathfrak{X}\left(\mathbb{C}^2, 0\right)$. We need to precise the corresponding notions for unfoldings in order to perform their classification.

Definition 2. We say that two unfoldings $(Z_\lambda)_{\lambda\in\Lambda}$ and $\left(\widetilde{Z}_{\tilde{\lambda}}\right)_{\tilde{\lambda}\in\tilde{\Lambda}}$ of codimension k are (formally, locally) **conjugate** (*resp.* **orbitally equivalent**) if there exists an association

$$\Psi \ : \ (\lambda, x, y) \longmapsto (\phi\,(\lambda)\ ,\ \Psi_\lambda\,(x, y))$$

in the corresponding regularity class, such that:

1. $\lambda \in \left(\mathbb{C}^k, 0\right) \mapsto \tilde{\lambda} = \phi\,(\lambda)$ has invertible derivative at 0,
2. the identity $\widetilde{Z}_{\phi(\lambda)} \cdot \Psi_\lambda = Z_\lambda \circ \Psi_\lambda$ is satisfied at a formal level. When Ψ is analytic this is equivalent to the property that, for each $\lambda \in \left(\mathbb{C}^k, 0\right)$, the component Ψ_λ is a conjugacy (*resp.* orbital equivalence) between Z_λ and $\widetilde{Z}_{\phi(\lambda)}$.

If the above conditions are fulfilled, we write

$$\Psi^*\,(Z_\lambda)_\lambda = \left(\widetilde{Z}_{\tilde{\lambda}}\right)_{\tilde{\lambda}}\ .$$

We wish to describe the (formal, local) classification in the set

$$\mathrm{SNU}\,(1) := \{(Z_\lambda)_\lambda \ : \ (Z_\lambda)_\lambda \ \text{generic unfolding of multiplicity } 1\}\ .$$

Definition 3. We use the notations $\mathrm{Mod}_{\mathrm{for}}\,(1)$ and $\mathrm{Mod}_{\mathrm{loc}}\,(1)$ to stand for the **moduli spaces** of $\mathrm{SNU}\,(1)$ under corresponding conjugacy. The **orbital moduli spaces** under orbital equivalence is written $\mathrm{Mod}^{\mathrm{orb}}_{\mathrm{for}}\,(1)$ and $\mathrm{Mod}^{\mathrm{orb}}_{\mathrm{loc}}\,(1)$, respectively. The same notational convention is used for the class of an unfolding:

$$[(Z_\lambda)_\lambda]^\sharp_\star \in \mathrm{Mod}^\sharp_\star\,(1)\ .$$

We need to slacken a little the clause of uniqueness for **normal forms**: we require that there exists $\nu \in \mathbb{N}$ such that the equivalence class of $[Z_\lambda]^\sharp_\star$ is contained in the orbit of a $\mathbb{C}\,\{\lambda\}^\nu$-action.

22.7 Every Step of the Way

For the sake of clarity we present only the case of codimension $k = 1$. Unlike saddle-node singularities, where a general complete classification is not harder to obtain than for $k = 1$, unfoldings for $k > 1$ are more difficult to deal with this way, due to the need of splitting the parameter space into many cells. The specific problems and corresponding results are detailed in [28].

Let us summarize the different steps leading to the classification of generic saddle-node unfoldings $Z_\bullet$ of codimension 1. The section ends by the statement of the main theorems. Most important items in the list below are developed later on in the course of the chapter. There we will outline the precise setting of and problems occurring in the upcoming constructions, while referring to [28] for all details concerning actual proofs.

22.7.1 Preparation of the Family (Sect. 22.8.1)

We can choose local conformal coordinates (λ, x, y) in which $Z_\bullet$ can be put under **prepared form**

$$Z_\lambda = U_\lambda \left(\mathrm{X}^\infty_\lambda + (\cdots)\,\frac{\partial}{\partial y} \right)$$

where $\mathrm{X}^\infty_\bullet$, called the **orbital model**, is formally orbitally equivalent to $Z_\bullet$, the notation $(\cdots)\,\frac{\partial}{\partial y}$ denotes a transverse holomorphic perturbation and $U_\bullet \in \mathbb{C}\,\{\lambda, x, y\}^\times$. Moreover the singularities of Z_λ are located at points in $\{y = 0\}$ corresponding to the roots of the polynomial

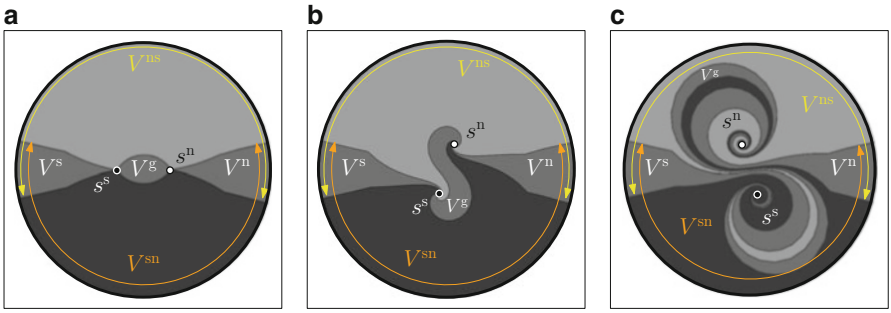


Fig. 22.9 A sectorial decomposition of the x-variable when $k = 1$ for three values of the parameter. (**a**) $\lambda = -1$; (**b**) $\lambda = \exp \frac{-i\pi}{8}$; (**c**) $\lambda = \exp \frac{i\pi}{8}$

$$P_\lambda(x) := x^2 + \lambda \,. \tag{22.23}$$

A formal conjugacy between prepared unfoldings must fix the parameter λ: it is a formal invariant. In the sequel we only consider prepared unfoldings, which allows us to work for fixed λ. To lighten notations we omit to mention the subscript "λ" in all the following items.

22.7.2 Sectorial Decomposition (Sect. 22.10)

The local invariants of $Z_\bullet$ are built by comparing transition maps between two neighboring normalizing charts. By this we mean to cut $\left(\mathbb{C}^2, 0\right) \setminus P^{-1}(0)$ up into two overlapping open **canonical sectors** $\mathcal{V}^{ns}$ and $\mathcal{V}^{sn}$ on every one of which Z is orbitally equivalent to the model X^∞. The canonical sectors are fibered over **squid sectors** $V^\sharp$ in the x-variable, displayed in Fig. 22.9. The intersection $V^{ns} \cap V^{sn}$ has three components:

- a **saddle-part** V^s having only s^s in its adherence,
- a **node-part** V^n having only s^n in its adherence,
- a **gate-part** V^g having both points in its adherence.

We write

$$\mathcal{V}^\sharp := V^\sharp \times (\mathbb{C}, 0) \,, \qquad \sharp \in \{\mathrm{n}\,,\ \mathrm{ns}\,,\ \mathrm{s}\,,\ \mathrm{sn}\,,\ \mathrm{g}\}$$

the corresponding fibered sectors. There are two transitions to consider for orbital equivalence (happening in the saddle- and node-part), and one more to account for conjugacy (over the saddle-part).

We denote by $\mathcal{F}$ the (germ of a) singular foliation induced by Z. Each canonical sector is attached to both singular points s^n and s^s. The boundary of a sector is carved in such a way that the leaves of $\mathcal{F}|_{\mathcal{V}^\sharp}$ near s^n behave like those of a node: every

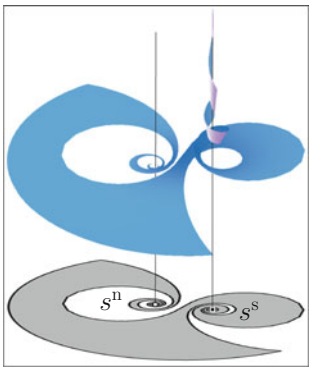


Fig. 22.10 Modulus of the y-component of a leaf over a sector $V^\sharp$ in the x-variable represented as a height map in real 3-space

leaf accumulates on the singular point. Near s^{s} the foliation is similar to a saddle: every leaf but one (the local invariant manifold) stays far away from the singular point (Fig. 22.10). This topological configuration allows us to mimic constructions performed in Sect. 22.4 for affine unfoldings. In particular the sectorial space of leaves is a conformal line

$$\Omega^\sharp = H^\sharp\left(\mathcal{V}^\sharp\right) = \mathbb{C}\,,$$

where $H^\sharp \in \mathrm{Holo}\left(\mathcal{V}^\sharp\right)$ is the **canonical first integral** of $\mathcal{F}|_{\mathcal{V}^\sharp}$, having connected fibers.

22.7.3 Straightening of the Weak Separatrices

If a singularity p of X is not a node, there exists only one integral curve with smooth analytic closure passing through p (a separatrix, or invariant manifold) and transverse to the vertical lines $\{x = \mathrm{cst}\}$. It is given by the graph of a holomorphic function $x \mapsto \mathfrak{s}(x)$, with holomorphic continuation over every canonical sector. We call such an analytic continuation a **sectorial weak separatrix**. It corresponds to the level 0 of the canonical first integral $H^\sharp$. In particular both sectorial weak separatrices coincide in saddle- and gate-parts.

Applying the change of coordinates $\psi : (x, y) \mapsto (x, y + \mathfrak{s}(x))$ to Z straightens the sectorial weak separatrix into $\{y = 0\}$. Notice that $\mathfrak{s}$ cannot (in general) be analytically continued on a whole neighborhood of the other singularity, as a given sectorial weak separatrix may not coincide with the other one when continued (the typical $\mathfrak{s}$ is multivalued). When $\mathfrak{s}$ does extend holomorphically on a neighborhood of all singular points we say that a **heteroclinic connection** occurs between s^{s} and s^{n}.

22.7.4 Normalization Strategy (Sect. 22.8.2)

After straightening, we can write:

$$Z = UX$$

$$X = \mathrm{X}^{\infty} + yR\frac{\partial}{\partial y}$$

for some (sectorial) holomorphic functions U and R. It turns out the conjugacy equation $\mathcal{O}^{*}\mathrm{X}^{\infty} = X$ is equivalent to the following **orbital cohomological equation**

$$X \cdot O = -R \tag{22.24}$$

if one seeks a conjugacy in the form

$$\mathcal{O}(x, y) = (x, y \exp O(x, y)) \ .$$

Also the conjugacy equation $\mathcal{T}^{*}X = UX$ takes the form of a **temporal cohomological equation**

$$X \cdot T = \frac{1}{U} - 1 \tag{22.25}$$

for changes of coordinates

$$\mathcal{T}(x, y) = \Phi_{X}^{T(x,y)}(x, y)$$

obtained by taking a dependent time in the flow $\Phi_{X}^{t}(x, y)$ of X. Therefore the normalization process has been reduced to solving two cohomological equations.

22.7.5 Formal Classification (Sect. 22.8.3)

The straightening step can be realized at a formal level. The orbital cohomological equation (22.24) can be solved formally, more or less by construction of the model X^{∞} and of the prepared form. The temporal cohomological equation (22.25) needs some adjustment since $\frac{1}{U} - 1$ may not belong to the image of the Lie derivative associated with X. We can find an affine function $\mathfrak{u} \in \mathbb{C}[x]_{\leq 1}$, relatively prime with P, such that $\frac{1}{U} - \frac{1}{\mathfrak{u}}$ belongs to that image. Hence $\mathfrak{u}\mathrm{X}^{\infty}$ is formally conjugate to Z, finally yielding a polynomial **formal model**

$$Z^{\mathfrak{u}} := \mathfrak{u}\mathrm{X}^{\infty} \ .$$

This complete formal classification of generic unfoldings ranges in the space of germs $\mathbb{C}\{\lambda\}^{3}$.

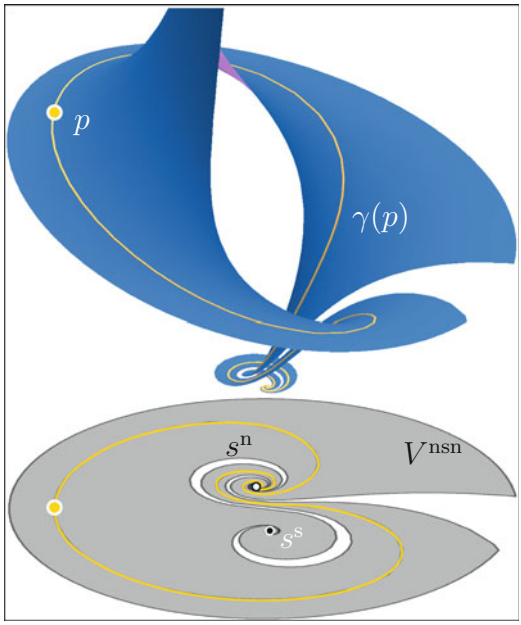


Fig. 22.11 An asymptotic cycle and its projection on the x-variable (Modulus of the y-coordinate of the leaf as a height-map)

22.7.6 Local Classification (Sect. 22.11)

The normalizing cohomological equations (22.24) and (22.25) admit a bounded solution on each canonical sector. They are obtained by integrating the right-hand side against $\frac{\mathrm{d}x}{P}$ along **asymptotic paths** tangent to $\mathcal{F}$, ending at the point (x, y) and accumulating on s^{n} in backward time. The node-like nature of s^{n} guarantees that every point in the sector can be reached by asymptotic paths, and that both solutions coincide in the node- and gate-part.

Therefore Z is conjugate to the formal model Z^{u} on each canonical sectors. The local class of $Z_\bullet$ is thus completely determined by the following data.

- Its orbital class $\varphi := \left(\varphi^{\mathrm{n}}, \varphi^{\mathrm{s}}\right)$, obtained by comparing:
  - the sectorial weak separatrices in the node-part, measuring how far the vector field is from having a heteroclinic connection and encoded in a translation $h \mapsto h + \varphi^{\mathrm{n}}$,
  - the sectorial solutions to (22.24) in the saddle-part, measuring how far the continued sectorial solution of the orbital cohomological equation is from uniformity (that is, continuity).
- Its temporal class f, encoded by comparing the sectorial solutions to (22.25) in the saddle-part.

In both cases where cohomological equations are involved, comparing sectorial solutions means to consider the **period**

$$g\,(p) := \frac{1}{2\mathrm{i}\pi}\int_{\gamma(p)} G\frac{\mathrm{d}x}{P}$$

of the right-hand side G along an **asymptotic cycle** $\gamma\,(p)$ circling around s^{s} and tending asymptotically to s^{n} both in forward and backward time, while passing through $p \in \mathcal{V}^{\mathrm{s}}$, as displayed in Fig. 22.11. This gives an integral representation of the invariants over the saddle-part (Sect. 22.11.5.2). The value of the integral does not depend on the choice of p in a fixed leaf of the restriction of $\mathcal{F}$ to the sector

$$\mathcal{V}^{\mathrm{nsn}} := \left(\mathbb{C}^2, 0\right) \setminus \left(\mathcal{V}^{\mathrm{n}} \cup \mathcal{V}^{\mathrm{g}}\right)$$

because the leaf is simply connected. Hence g is a first integral of X and for that matter factors as a (germ of a) holomorphic function of the canonical first integral H^{ns}, holomorphic on $\mathcal{V}^{\mathrm{ns}}$:

$$g = \mathfrak{T}_{\bullet}^{X}\,(G) \circ H^{\mathrm{ns}} \quad , \quad \mathfrak{T}_{\bullet}^{X}\,(G) \in h\mathbb{C}\,\{h\}\ .$$

Remark 10. For the period $\mathfrak{T}_{\bullet}^{X}\,(G)$ to be well defined, we need to mod it out by the degree of freedom in the choice of the sectorial first integrals H^{ns}. This freedom results from a faithful action of $\mathbb{C}^{\times}$ by linear changes of variables $h \mapsto ch$. We leave such essentially irrelevant technicalities out till subsequent sections.

For a fixed formal model we finally obtain a local classification ranging in a functional space

$$[(Z_\lambda)_\lambda]_{\mathrm{loc}} \longmapsto (\lambda \mapsto \mathfrak{m}_\lambda) \quad , \quad \mathfrak{m}_\lambda = \left(\varphi_\lambda^{\mathrm{n}}, \varphi_\lambda^{\mathrm{s}}, f_\lambda\right) \in \mathbb{C} \times h\mathbb{C}\,\{h\} \times h\mathbb{C}\,\{h\}\ .$$

Notice how cowardly we shied away from discussing the dependence in the parameter λ until now. The result to come, proved in Sects. 22.11 and 22.13.1, gives a precise description of that dependence.

Theorem 4. *There exists a cover of* $(\mathbb{C}, 0)$ *by the adherence of two germs of sectors* Λ^{+}, Λ^{-} *attached to* 0 *(called cells), for which we can find a complete local classification of generic unfoldings of codimension* 1 *with fixed formal class, ranging in the space of those collections* $\left(\mathfrak{m}_{\bullet}^{-}, \mathfrak{m}_{\bullet}^{+}\right)$ *satisfying:*

- $(\lambda, h) \in \Lambda^{\pm} \times (\mathbb{C}, 0) \mapsto m_\lambda^{\pm}\,(h)$ *is holomorphic with continuous extension to* $\mathrm{adh}\left(\Lambda^{\pm}\right) \times (\mathbb{C}, 0)$*, and* $\mathfrak{m}_\lambda^{\pm}$ *is holomorphic for every* $\lambda \in \partial\Lambda^{\pm}$*,*
- $\mathfrak{m}_\lambda^{\pm} \in \mathbb{C} \times h\mathbb{C}\,\{h\} \times h\mathbb{C}\,\{h\}$*,*
- *the* ***compatibility condition****, given explicitly in Definitions 16 and 17.*

By construction the model $\mathfrak{u}_{\bullet}\mathrm{X}_{\bullet}^{\infty}$ has local class $\mathfrak{m}_{\bullet}^{\pm} = 0$.

22.7.7 Normal Forms for Pure Convergence (Sect. 22.13.2)

When $\varphi_\bullet^{\mathrm{n}} = 0$ (we say the unfolding is **purely convergent**) we provide normal forms for $Z_\bullet$. Remark both sectorial separatrices glue to form a holomorphic weak separatrix (heteroclinic connection) and therefore purely convergent unfoldings are locally conjugate to prepared unfoldings for which $\{y = 0\}$ is a leaf, and vice versa.

Theorem 5. *Let* Convergent(1) *be the space of all purely convergent, generic unfoldings of codimension* 1. *Define*

$$\tau := \begin{cases} 0 & \text{if } \mu_0 \notin \mathbb{R}_{\leq 0} \\ 1 + \lfloor -\mu_0 \rfloor & \text{otherwise} \end{cases}$$

$$\mathrm{Section}\,(1,\tau) := x^{\tau+1} y \mathbb{C}\{\lambda, x^\tau y\} \ .$$

Then the collection

$$\left\{ \frac{\mathrm{u}_\bullet}{1 + \mathrm{u}_\bullet Q_\bullet} \left(\mathrm{X}_\bullet^\infty + y R_\bullet \frac{\partial}{\partial y} \right) \ : \ Q_\bullet, \ R_\bullet \in \mathrm{Section}\,(1,\tau) \right\}$$

is a family of normal forms for Convergent(1).

More precisely, two unfoldings in that form for $(Q_\bullet, R_\bullet)$ *and* $\left(\widetilde{Q}_\bullet, \widetilde{R}_\bullet\right)$ *are locally analytically conjugate if there exists* $c_\bullet \in \mathbb{C}\{\lambda\}^\times$ *such that*

$$\begin{cases} R_\lambda(x,y) & = \widetilde{R}_\lambda(x, c_\lambda y) \\ Q_\lambda(x,y) & = \widetilde{Q}_\lambda(x, c_\lambda y) \end{cases},$$

amounting to a linear $\mathbb{C}\{\lambda\}^\times$*-action.*

22.8 Preparation and Formal Classification

22.8.1 Preparation

Theorem 6. *There exists local conformal coordinates* (λ, x, y) *in which* $Z_\bullet$ *has the following* ***prepared form***

$$\begin{aligned} Z_\lambda &= U_\lambda X_\lambda \\ X_\lambda(x,y) &= P_\lambda(x) \frac{\partial}{\partial x} + \left(y(1 + \mu_\lambda x) + P_\lambda(x) R_\lambda(x,y) \right) \frac{\partial}{\partial y} \end{aligned} \qquad (22.26)$$

where

$$\begin{cases} \lambda & \in (\mathbb{C}, 0) \\ \mu_\bullet & \in \mathbb{C}\{\lambda\} \\ R_\bullet & \in \mathbb{C}\{\lambda, x, y\} \\ U_\bullet & \in \mathbb{C}\{\lambda, x, y\}^\times \end{cases}$$

are arbitrary.

Notice that every prepared form is a generic unfolding of codimension 1, for $\mathrm{Sing}(X_\lambda) = \{y = 0\} \cap P_\lambda^{-1}(0)$. It turns out that formal changes of coordinates fixing the general form of the family (22.26) must leave the parameter λ invariant. This parameter therefore plays a special role for the unfolding, and we call it the **canonical parameter**. It thus suffices to work with fixed λ in order to perform the local classification.

22.8.2 From Normalization to Cohomological Equations

The whole procedure relies on writing the conjugacy equation $\Psi^* Z = \widetilde{Z}$ as **cohomological equations**

$$Z \cdot F = G \tag{22.27}$$

for well-chosen right-hand sides. The key computation is the following proposition.

Proposition 4. *Let X and Y be two germs of a holomorphic vector field on a domain $\mathcal{U}$ such that $[X, Y] = 0$. If f is holomorphic on $\mathcal{U}$ (*resp. *a formal power series at some point $p \in \mathcal{U}$), then*

$$\Psi(x, y) := \Phi_Y^{f(x,y)}(x, y)$$

has same regularity as f, and satisfies

$$\Psi^* X = X - \frac{X \cdot f}{1 + Y \cdot f} Y .$$

In particular, the following properties hold.

1. (**Temporal conjugacy**) *UZ is conjugate to VZ by Φ_{UZ}^T if and only if*

$$Z \cdot T = \frac{1}{U} - \frac{1}{V} .$$

2. (**Orbital conjugacy**) *Assume $X \pitchfork Y$. Then X is conjugate to $X + RY$ by Φ_Y^O if and only if*

$$X \cdot O = R \,.$$

Proof. It is sufficient to perform computations at a formal level. We use Lie formula (22.19) so that, because $X \cdot Y = Y \cdot X$,

$$\begin{aligned} X \circ \Psi &= \sum_{n \geq 0} \frac{f^n}{n!} Y \cdot^n (X \cdot \mathrm{Id}) \\ &= \sum_{n \geq 0} \frac{f^n}{n!} X \cdot (Y \cdot^n \mathrm{Id}) \,. \end{aligned}$$

Besides

$$\begin{aligned} \mathrm{D}\Psi \,(X - RY) &= (X - RY) \cdot \Psi \\ &= \sum_{n \geq 0} (X - RY) \cdot \left(\frac{f^n}{n!} Y \cdot^n \mathrm{Id} \right) \\ &= (X - RY) \cdot f \times \sum_{n \geq 0} \frac{f^n}{n!} Y \cdot^{n+1} \mathrm{Id} + \sum_{n \geq 0} \frac{f^n}{n!} (X - RY) \cdot Y \cdot^n \mathrm{Id} \\ &= X \circ \Psi + (X \cdot f - R\,(1 + Y \cdot f)) \times Y \circ \Psi \,. \end{aligned}$$

Therefore

$$R = \frac{X \cdot f}{1 + Y \cdot f} \,.$$

To prove 1. it suffices to take $X := UZ$ and $Y := UZ$. □

By taking $(Z_s)_s$ in prepared form (22.26) we can write

$$\begin{aligned} Z_\lambda &= U_\lambda X_\lambda \\ X_\lambda &= \left(\mathrm{X}_\lambda^\infty + P_\lambda R_\lambda \frac{\partial}{\partial y} \right) \\ \mathrm{X}_\lambda^\infty &:= P_\lambda \frac{\partial}{\partial x} + y\,(1 + \mu_\lambda x) \frac{\partial}{\partial y} \,. \end{aligned}$$

Notice that

$$\left[\mathrm{X}_\lambda^\infty, y \frac{\partial}{\partial y} \right] = 0$$

so we can apply Proposition 4 Item (2), as soon as R_s can be factored by y, that is once the weak separatrix is straightened into $\{y = 0\}$.

22.8.3 Formal Classification

The strategy is the following.

- There exists a unique weak separatrix family, that is a formal family of curves $\{y - \mathfrak{s}_\lambda(x) = 0\}$ such that

$$P_\lambda(x)\,\hat{\mathfrak{s}}'_\lambda(x) = \hat{\mathfrak{s}}_\lambda(x)\,(1+\mu_\lambda x) + P_\lambda(x)\,R_\lambda(x,\hat{\mathfrak{s}}_\lambda(x)) \qquad , \ \hat{\mathfrak{s}}_\bullet \in \mathbb{C}[[\lambda,x,y]] \ .$$

- After applying the change of coordinates $(x,y) \mapsto (x, y - \hat{\mathfrak{s}}_\lambda(x))$ to X_λ we obtain the formal vector field

$$\widehat{X}_\lambda = \mathrm{X}^\infty_\lambda + P_\lambda \widehat{R}_\lambda y \frac{\partial}{\partial y}$$

where

$$\widehat{R}_\lambda(x,y) := \frac{R_\lambda(x, y+\hat{\mathfrak{s}}_\lambda(x)) - R_\lambda(x,\hat{\mathfrak{s}}_\lambda(x))}{y} \ . \tag{22.28}$$

- The cohomological equation

$$\widehat{X}_\lambda \cdot \widehat{O}_\lambda = -P_\lambda \widehat{R}_\lambda$$

admits a unique formal family of solutions $\widehat{O}_\bullet \in \mathbb{C}[[\lambda,x,y]]$ such that $\widehat{O}_\lambda(0) = 0$. Therefore $\mathrm{X}^\infty_\bullet$ is orbitally formally conjugate to $\widehat{X}_\bullet$ by

$$\widehat{\mathcal{O}}_\bullet := \Phi^{\widehat{O}_\bullet}_{y\frac{\partial}{\partial y}} \ ,$$

thus formally conjugate to $X_\bullet$ by a (λ,x)-fibered formal conjugacy. Remark that the previous cohomological equation is equivalent to one involving X_λ in the original coordinates:

$$X_\lambda \cdot O_\lambda = -P_\lambda \frac{R_\lambda(x,y) - R_\lambda(x,\hat{\mathfrak{s}}_\lambda(x))}{y - \hat{\mathfrak{s}}_\lambda(x)} \ . \tag{22.29}$$

- The cohomological equation with $\mathfrak{u}_\bullet \in \mathbb{C}[[\lambda,x]]^\times$

$$X_\lambda \cdot \widehat{T}_\lambda = \frac{1}{U_\lambda} - \frac{1}{\mathfrak{u}_\lambda} \tag{22.30}$$

admits a formal solution if and only if

$$U_\lambda(x,y) = \mathfrak{u}_\lambda(x) + \mathrm{O}\left(x^2\right) + \mathrm{O}(y)$$

where $\mathfrak{u}_\lambda$ is relatively prime with P_λ in the factorial ring $\mathbb{C}[[\lambda, x, y]]$. The holomorphic germ $\mathfrak{u}_\bullet$ can therefore be taken as the remainder of the Euclidean division of $U_\lambda(x,0)$ by $P_\lambda(x)$. In particular for each value of the parameter $\mathfrak{u}_\lambda$ is affine. Because $\widehat{\mathcal{O}}_\lambda$ is x-fibered, the vector field $\mathfrak{u}_\lambda \mathrm{X}_\lambda^\infty$ is formally conjugate to $\mathfrak{u}_\lambda X_\lambda$, thus to Z_λ.

The last two claims derive from the following easy computational lemma.

Lemma 4. *[29] Let $\widehat{G}_\bullet \in \mathbb{C}[[\lambda, x, y]]$ be given. The cohomological equation*

$$X_\lambda \cdot \widehat{F}_\lambda = \widehat{G}_\lambda$$

admits a formal solution $\widehat{F}_\bullet \in \mathbb{C}[[\lambda, x, y]]$ if and only if $\widehat{G}_\lambda = \mathrm{O}(P_\lambda(x)) + \mathrm{O}(y)$. This family of solution is unique up to the addition with an arbitrary formal power series belonging to $\mathbb{C}[[\lambda]]$, corresponding to the choice of the value $\widehat{F}_\lambda(0)$.

Discounting additional straightforward computations, we just established a formal classification with normal forms.

Theorem 7. *[29] We have complete classifications*

$$\mathrm{Mod}_{\mathrm{for}}(1) \simeq \mathbb{C}\{\boldsymbol{\lambda}\}[x]^\times_{\leq 1} \times \mathrm{Mod}^{\mathrm{orb}}_{\mathrm{for}}(1)$$

$$\mathrm{Mod}^{\mathrm{orb}}_{\mathrm{for}}(1) \simeq \mathbb{C}\{\boldsymbol{\lambda}\}$$

with normal forms

$$\mathrm{NF}_{\mathrm{for}}(1) := \mathbb{C}\{\lambda\}[x]^\times_{\leq 1}\, \mathrm{NF}^{\mathrm{orb}}_{\mathrm{for}}(1)$$

$$\mathrm{NF}^{\mathrm{orb}}_{\mathrm{for}}(1) := \left\{ P_\bullet(x)\frac{\partial}{\partial x} + y(1+\mu_\bullet x)\frac{\partial}{\partial y} \;:\; \mu_\bullet \in \mathbb{C}\{\lambda\} \right\}.$$

Definition 4. An unfolding $\widehat{Z}_\bullet$ in $\mathrm{NF}^{\mathrm{orb}}_{\mathrm{for}}(1)$ is now fixed, corresponding to the choice of the holomorphic germs $\mu_\bullet : \lambda \mapsto \mu_\lambda$ and $\mathfrak{u}_\bullet : (\lambda, x) \mapsto \mathfrak{u}_\lambda(x)$. It is referred to as the **model**.

The local classification is obtained by repeating the construction on sectors over which $\widehat{O}_\lambda$ and $\widehat{T}_\lambda$ have holomorphic "sums". The local invariant measures how far from converge these power series are.

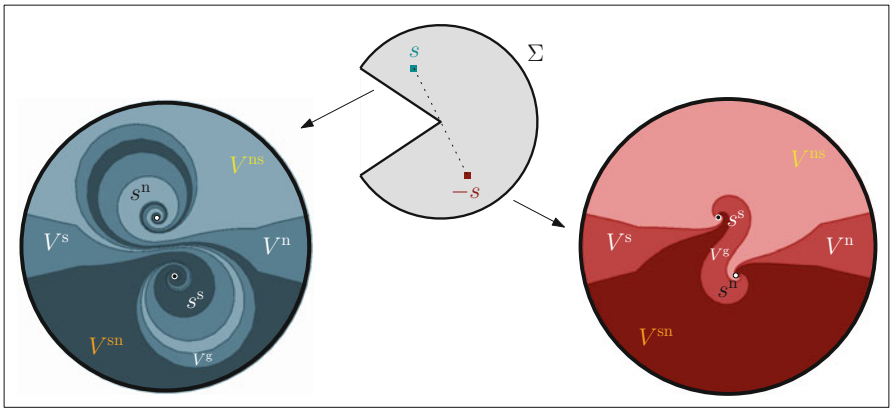


Fig. 22.12 Two non-equivalent decompositions by squid sectors with the same value of the parameter $\lambda = -s^2$. Notice that the nature of the singular points (node- or saddle-like) swaps from one configuration to the other

22.9 Parameter Space

To carry out the construction of the sectorial decomposition we need to follow singularities as λ varies. Although the *set* Sing (Z_λ) depends continuously on λ, it is not possible to *mark* and follow continuously singularities, as these get exchanged by turning around the bifurcation value of λ (corresponding to $P_0(x) = x^2$). This phenomenon has a prominent bearing on the construction of sectors, since for the same value of λ one obtains dynamically non-equivalent coverings, as illustrated in Fig. 22.12. We resolve the ambiguity in the labeling of the singularities by using the two-fold branched covering

$$\widehat{P}_\bullet \; : \; s \in \mathbb{C} \longmapsto (x-s)\,(x+s)$$

and take s, which determines completely the position of the roots $\pm s$, as new parameter for the constructions.

Definition 5. Let us call Param (1) the complex line $\mathbb{C}$ viewed as the s-space. The branched covering

$$\begin{aligned} \lambda \; : \; \mathrm{Param}\,(1) &\longrightarrow \mathbb{C} \\ s &\longmapsto \lambda(s) := -s^2 \,, \end{aligned}$$

satisfying the identity

$$P_{\lambda(s)} = \widehat{P}_s \,,$$

is called the **canonical re-parameterization**. We will actually use only strict subsectors of Param (1), given for $\rho > 0$ by

$$\Sigma := \left\{ s \ : \ 0 < |s| < \rho \ , \ |\arg s| < \frac{2\pi}{3} \right\}$$

(where the principal determination of the argument on $\mathbb{C}\backslash\mathbb{R}_{\leq 0}$ is used). For each $s \in$ Param (1) let us define

$$s^{\mathrm{n}} := (s, 0)$$
$$s^{\mathrm{s}} := (-s, 0) \ ,$$

which are the singularities of X_λ.

Remark 11. Notice that $\lambda(\Sigma \cup \{0\}) = (\mathbb{C}, 0)$ so every original parameter is covered this way. The explanation as to why we cannot take $\Sigma \cup \{0\} = (\mathbb{C}, 0)$ will be given in the course of the upcoming sections (especially Sect. 22.10.3).

The next properties will be used without explicitly referencing the trivial lemma beneath.

Lemma 5.

1. *The automorphism group of the covering λ is isomorphic to $\frac{\mathbb{Z}}{2\mathbb{Z}}$.*
2. *The critical set Δ of λ is the origin.*
3. *A (germ of a) holomorphic function $\hat{f}$: Param (1) $\to \mathbb{C}$ factors as $f \circ \lambda$ with f holomorphic if and only if $\hat{f}$ is even (i.e., $\frac{\mathbb{Z}}{2\mathbb{Z}}$-invariant).*

Definition 6. In all the remaining text, we make the following notational conventions:

- when an object Ω is subscripted with "s" we imply $s \mapsto \Omega_s$ depends on s in a (holomorphic, continuous) way on Σ,
- when an object $\tilde{\Omega}$ is subscripted with "λ" we imply $\lambda \mapsto \tilde{\Omega}_\lambda$ depends (formally, holomorphically, continuously) on $\lambda \in (\mathbb{C}, 0)$.

22.10 Canonical Sectors

22.10.1 Splitting Vector Fields

The boundary of a squid sector will be defined by real trajectories of vector fields

$$\Xi_s(x) := \vartheta P_\lambda(x) \frac{\partial}{\partial x}$$

for some suitable choice of a direction $\vartheta = \vartheta(s) \in \mathbb{S}^1$. In order to ensure the upcoming construction matches our needs, we must ensure that the vector field is

sufficiently generic. Let us explain what we mean by describing some dynamical data attached to the planar real-analytic foliation $\mathcal{F}$ induced by Ξ_s on $\overline{\mathbb{C}}$. The most complete reference on the subject is [2].

- $\mathrm{Sing}(\mathcal{F}) = P_\lambda^{-1}(0)$, in particular $\mathcal{F}$ is regular near ∞. We call **local separatrix** of Ξ_s any one of the two leaves of $\mathcal{F}|_{(\overline{\mathbb{C}},\infty)\setminus\{\infty\}}$ accumulating on ∞. We call separatrix the corresponding leaf in $\mathcal{F}|_{\mathbb{C}}$.
- Because Ξ_s is holomorphic, $\mathcal{F}$ is free from limit (poly)cycles. Therefore the fate of a non-singular trajectory Γ of Ξ_s can only be one of the following (Bendixon–Poincaré theorem).
  - Γ is a separatrix and its adherence connects ∞ to either a singular point $\pm s$ of Ξ_s, or ∞. In the former case we say that Γ **lands** at $\pm s$. In the latter case we say Γ is a **homoclinic connection** (happening exactly when the continuations of both local separatrices meet *en route*).
  - Γ connects (asymptotically) both singular points s and $-s$.
  - Γ is a non-isolated simple loop. In that case both $\pm s$ are center points.

Definition 7. Take $s \in \mathrm{Param}(1)$. We say that Ξ_s is **splitting** if it admits no homoclinic connection.

Lemma 6. *The following conditions are equivalent.*

1. *Ξ_s is splitting.*
2. *There exists a leaf connecting s and $-s$.*
3. *$\vartheta P'_\lambda(\pm s) \notin \mathrm{i}\mathbb{R}$.*

Let us briefly explain why this lemma holds. When $s \neq 0$ the vector field Ξ_s is locally linearizable around each root $\pm s$ of P_λ, and its linear part is given by $\vartheta P'_\lambda(\pm s)(x \mp s)\frac{\partial}{\partial x}$. Therefore $\vartheta P'_\lambda(\pm s)$ is purely imaginary (non-zero) if and only if $\pm s$ is a center point of Ξ_s.

- When $\pm s$ is a center, it lies within an open basin $B^\pm$ of periodic trajectories. In particular no leaf can connect s and $-s$. Moreover $\partial B^\pm$ must be a homoclinic connection, and Ξ_s is not splitting.
- When $\pm s$ is not a center, it is either a focus or a sink. In any case it lies within an open basin of attraction $B^\pm$ (respectively in forward or backward time) and any integral curve of Ξ_s crossing the basin must accumulate on $\pm s$ one way or the other. There are no trajectory accumulating on s (or on $-s$) in both directions.

22.10.2 Transvestite Hyperbolic Points

For $l \in \mathbb{C}^\times$ we consider the linear vector field (Fig. 22.13)

$$W(x,y) := lx\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\,.$$

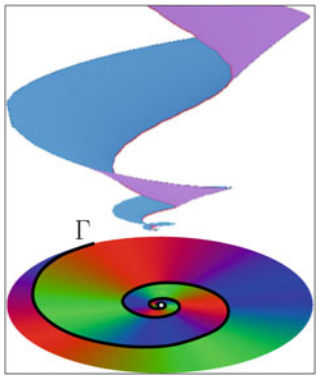


Fig. 22.13 A leaf of a node-like foliation over a slit disc, cut out from a hyperbolic singularity (modulus represented as a height-map, argument as *colors* in the base) (Color figure online)

Save for the separatrix $\{x = 0\}$, the leaves of $\mathcal{F}_W$ are included in level sets of the (in general) multivalued first integral

$$H(x, y) := yx^{-\frac{1}{l}} .$$

As explained already in Example 8, the space of leaves of $\mathcal{F}_W$ is the quotient

$$\Omega_W := {}^{\mathbb{C}^\times}\!/_{\mathbb{Z}} \cup \{0_x, 0_y\} ,$$

where 0_x and 0_y represent the branches of $\{xy = 0\} \setminus \{0\}$. The quotient corresponds to the multiplicative action of $\mathbb{Z}$ on the space of initial values $\mathbb{C}^\times$

$$y \longmapsto y \exp \tfrac{-2i\pi n}{l} \quad , \; n \in \mathbb{Z} ,$$

encoding the monodromy of H.

Choose $\vartheta \in \mathbb{S}^1$ and pick a real-time trajectory of $\vartheta lx\frac{\partial}{\partial x}$, given by

$$t \in \mathbb{R} \longmapsto x(t) := x_* \exp(\vartheta l t) , \qquad x_* \in \mathbb{C}^\times .$$

We can lift this path into $\mathcal{F}_W$ through the projection

$$\Pi \; : \; (x, y) \longmapsto x ,$$

starting from some $(x_*, y_*) \in \mathbb{C}^\times \times \mathbb{C}$. We obtain the path tangent to W

$$t \in \mathbb{R} \longmapsto (x(t), y(t)) \quad , \quad y(t) = y_* \exp(\vartheta t)$$

satisfying the identity $H(x(\bullet), y(\bullet)) = \text{cst}$.

Notice that

$$\lim_{t\to\pm\infty} x(t) = 0 \iff \pm\Re(\vartheta l) < 0$$
$$\lim_{t\to\pm\infty} y(t) = 0 \iff \pm\Re(\vartheta) < 0 \text{ or } y_* = 0 .$$

Definition 8. For given $l \in \mathbb{C}^\times$ we say that $\vartheta \in \mathbb{S}^1$ is a **saddle-direction** (*resp.* **node-direction**) for l if:

- $\Re(\vartheta) > 0$,
- $\Re(\vartheta l) < 0$ (*resp.* $\Re(\vartheta l) > 0$).

Remark 12. Just pointing out the obvious.

1. l admits a saddle-direction if and only if $l \notin \mathbb{R}_{>0}$ (i.e., W is not a node).
2. l admits a node-direction if and only if $l \notin \mathbb{R}_{<0}$ (i.e., W is not a saddle).

For such a choice of ϑ, the curve $t \mapsto x(t)$ is a spiral (specializing to a straight line when $\vartheta l \in \mathbb{R}$). Consider the domain

$$V := \mathbb{C}\setminus \exp\left(\vartheta l\overline{\mathbb{R}}\right)$$

obtained by slitting the complex line along the adherence Γ of a real integral curve of $\vartheta l x\frac{\partial}{\partial x}$, and build the fibered domain of the complex plane

$$\mathcal{V} := V \times \mathbb{C} .$$

Then $\mathcal{F}_W|_{\mathcal{V}}$ is **saddle-like** (*resp.* **node-like**) in the sense that only one (*resp.* every) leaf accumulates on 0. Notice that a saddle-like (*resp.* node-like) singularity is reached in positive (*resp.* negative) time. Also, (any determination of) the first integral H on $\mathcal{V}$ is holomorphic. The following properties are immediate to establish.

Lemma 7.

1. *$\mathcal{F}_W|_{\mathcal{V}}$ is saddle-like if and only if for every $\mathcal{U} = \left(\mathbb{C}^2, 0\right)$ we have $H(\mathcal{U}\cap\mathcal{V}) = (\mathbb{C}, 0)$ and its diameter goes to 0 as that of $\mathcal{U}$ does.*
2. *$\mathcal{F}_W|_{\mathcal{V}}$ is node-like if and only if for every $\mathcal{U} = \left(\mathbb{C}^2, 0\right)$ we have $H(\mathcal{U}\cap\mathcal{V}) = \mathbb{C}$.*

22.10.3 Sectorial Decomposition

We work here within a fixed formal class $\mu_\bullet$. We find a covering of $(\mathbb{C}, 0)\setminus\{\pm s\}$ by two squid sectors V_s^{ns} and V_s^{sn} attached to $\pm s$, such that near $s^{\mathrm{n}} = (s, 0)$ (*resp.* $s^{\mathrm{s}} = (-s, 0)$) the linear part of X_λ defines a node-like (*resp.* saddle-like) foliation over both sectors, except for forbidden values of s we shall describe afterward.

The linear part of X_λ at the singularity $(\pm s, 0)$ is

$$\pm 2s\,(x \mp s)\,\frac{\partial}{\partial x} + y\,(1 \pm \mu_\lambda s)\,\frac{\partial}{\partial y}\,,$$

while its local analytic invariant is given by

$$l_s^{\pm} := \frac{\pm 2s}{1 \pm \mu_\lambda s}\,. \tag{22.31}$$

Both functions $s \mapsto l_s^{\pm}$ are holomorphic on $(\mathbb{C}, 0)$. Notice that

$$l_s^{+} = l_{-s}^{-}$$

and, when $s \neq 0$,

$$\frac{1}{l_s^{+}} + \frac{1}{l_s^{-}} = \mu_\lambda\,.$$

Definition 9. The curves

$$\left\{s \neq 0\ :\ \pm l_s^{\pm} < 0\right\}$$

are called the **forbidden curves** associated with $\mu_\bullet$. Values of the parameter for "+" (*resp.* "−") correspond to configurations where X_λ is a saddle at s^{n} (*resp.* a node at s^{s}).

Till the end of the section we consider the principal determination of the argument on $\mathbb{C}\backslash\mathbb{R}_{\leq 0}$. For each $s \in \Sigma$ set

$$\vartheta\,(s) := \exp\frac{-\mathrm{i}\arg s}{2}\,. \tag{22.32}$$

Since

$$l_s^{\pm} \sim_0 \pm 2s$$

the following lemma holds (Fig. 22.14).

Lemma 8. *There exists $\rho > 0$ so that for all $s \in \Sigma$*

$$\begin{aligned} \left|\arg\left(\vartheta\,(s)\,l_s^{+}\right)\right| &< \frac{3\pi}{8} \\ \left|\arg\left(\vartheta\,(s)\,l_s^{-}\right)\right| &> \frac{5\pi}{8}\,. \end{aligned} \tag{22.33}$$

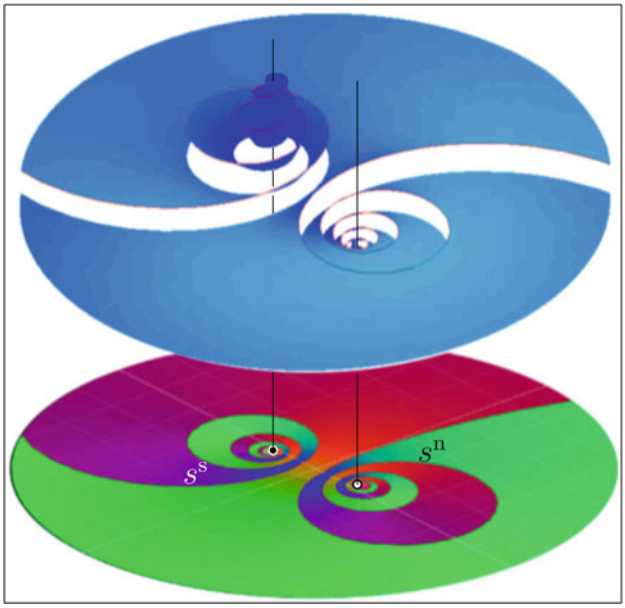


Fig. 22.14 The typical leaf above a slit disc (modulus represented as a height-map, argument as *colors* in the base) (Color figure online)

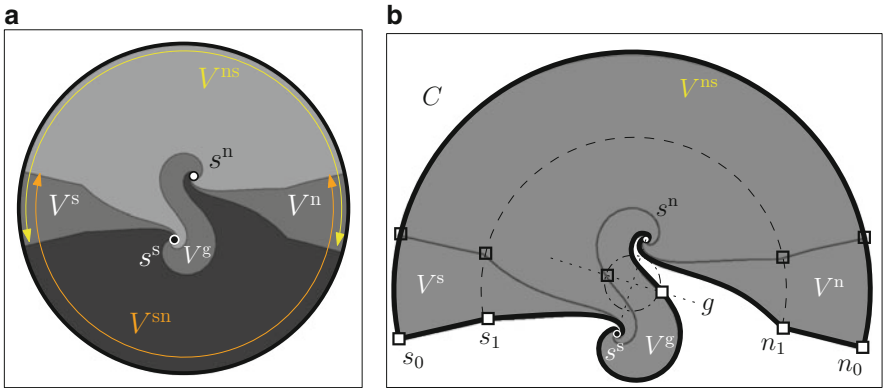


Fig. 22.15 (**a**) A sectorial covering by squid sectors. (**b**) Structure of a single squid sector. *White squares* represent construction points of V_s^{ns}, while *unfilled squares* do those of V_s^{sn}

In particular ϑ (s) *is a node-direction (*resp. *a saddle-direction) for* l_s^+ (resp. l_s^-), *and* Σ *meets no forbidden curve.*

Definition 10. We refer to Fig. 22.15b. Let $r > 2\rho > 0$ and $s \in \Sigma \cup \{0\}$. We recall the vector field

$$\Xi_s = \vartheta(s) P_\lambda \frac{\partial}{\partial x}$$

where ϑ is given by (22.32). The **squid sector** V_s^{ns} of radius r is the domain bounded by

- the forward trajectory of $P_0 \frac{\partial}{\partial x}$ starting from $n_0 := r \exp \frac{-\mathrm{i}\pi}{8}$ till it reaches $2|s|$ at a point n_1, then the complete forward trajectory of Ξ_s,
- the backward trajectory of $P_0 \frac{\partial}{\partial x}$ starting from $s_0 := r \exp \frac{9\mathrm{i}\pi}{8}$ till it reaches $2|s|$ at a point s_1, then the complete backward trajectory of Ξ_s,
- the integral curve of Ξ_s passing through the "outward" point of intersection g between the perpendicular bisector of $[-s, s]$ and a circle of radius small enough not to meet the already built paths (of the order of $|s|(1 + \cos \arg s)$),
- the circular arc $C := r \exp \mathrm{i}\left[-\frac{\mathrm{i}\pi}{8}, \frac{9\mathrm{i}\pi}{8}\right]$.

The squid sector V_s^{sn} of radius r is built in much the same way, replacing the circular arc C by $r \exp \mathrm{i}\left[-\frac{9\mathrm{i}\pi}{8}, \frac{\mathrm{i}\pi}{8}\right]$.

We mention without proof the next descriptive lemma.

Lemma 9. (See [28])

1. *The intersection $V_s^{\mathrm{ns}} \cap V_s^{\mathrm{sn}}$ has three components if $s \in \Sigma$:*
   - *a **saddle-part** V_s^{s} having only s^{s} in its adherence,*
   - *a **node-part** V_s^{n} having only s^{n} in its adherence,*
   - *and if $s \neq 0$, a **gate-part** V_s^{g} having both points in its adherence.*

   When $s = 0$ we define the saddle- or node-part as the components crossing $\mathbb{R}_{<0}$ or $\mathbb{R}_{>0}$, respectively.
2. *As $s \to 0$ in Σ, the squid sectors $V_s^\sharp$ tend (for the Hausdorff distance) to the sector $V_0^\sharp$ associated with the saddle-node Z_0.*
3. *The length of $\partial V_s^\sharp$ is uniformly bounded for $s \in \Sigma$.*

Remark 13. Item (3) above is really important in order to get uniform bounds in $s \in \Sigma$ for functions obtained by integrating over spirals included in $V_s^\sharp$, which includes almost all the upcoming material.

22.11 Local Classification

We work here within a fixed formal class $\mu_\bullet$. We take r, $r' > 0$ and $\rho > 0$ sufficiently small so that

$$\Sigma \times \mathcal{U} := \left\{ s \,:\, 0 < |s| < \rho \,,\ |\arg s| < \frac{2\pi}{3} \right\} \times \left(r\mathbb{D} \times r'\mathbb{D} \right)$$

is a domain on which:

- every data appearing in the preparation Theorem 6 is holomorphic and bounded,

- $(s,x) \mapsto \frac{P_\lambda(x)}{1+\mu_\lambda x}$ is holomorphic and bounded,
- Lemma 8 holds.

The actual values of $\rho,\ r,\ r'$ may be implicitly decreased finitely many times in the course of the construction. For any $s \in \Sigma$ we denote $V_s^\sharp$ the squid sector of radius r associated with s built in Definition 10. We follow now the strategy introduced for the formal classification, only on canonical sectors

$$\mathcal{V}_s^\sharp \subset V_s^\sharp \times r'\mathbb{D} \qquad , \ \sharp \in \{\mathrm{n}\ ,\ \mathrm{ns}\ ,\ \mathrm{s}\ ,\ \mathrm{sn}\ ,\ \mathrm{g}\}\ .$$

Before building the sectors, let us first introduce the space

$$\mathrm{Holo_c}\left(\mathcal{D}\right) := \left\{f_\bullet \in C^0\left(\mathrm{adh}\left(\mathcal{D}\right)\right)\ :\ f_\bullet \in \mathrm{Holo}\left(\mathcal{D}\right)\ ,\ (\forall s \in \mathrm{adh}\left(\Sigma\right))\ f_s \in \mathrm{Holo}\left(\mathcal{V}_s\right)\right\} \tag{22.34}$$

where $\mathcal{D}$ is a subdomain $\Sigma \times \mathbb{C}^n$ of the form

$$\mathcal{D} = \bigcup_{s\in\Sigma} \{s\} \times \mathcal{V}_s\ .$$

The key point is to provide weak sectorial separatrices $\mathfrak{s}_\bullet^\sharp$ and solutions $F_\bullet^\sharp$ to cohomological equations (22.27) which belong to $\mathrm{Holo_c}\left(D^\sharp\right)$ and $\mathrm{Holo_c}\left(\mathcal{D}^\sharp\right)$, respectively, with

$$D^\sharp := \bigcup_{s\in\Sigma} \{s\} \times V_s^\sharp\ , \tag{22.35}$$

$$\mathcal{D}^\sharp := \bigcup_{s\in\Sigma} \{s\} \times \mathcal{V}_s^\sharp\ .$$

22.11.1 Sectorial Weak Separatrices

Theorem 8 (See [28]). *Up to decrease $\rho,\ r,\ r'$ there exist two unique families of functions $\mathfrak{s}_\bullet^\sharp \in \mathrm{Holo_c}\left(D^\sharp\right)$, called in the following **sectorial (weak) separatrices**, such that for any $s \in \Sigma \cup \{0\}$:*

1. $\left\{y = \mathfrak{s}_s^\sharp(x)\right\} \subset V_s^\sharp \times r'\mathbb{D}$ *is an integral curve of* X_λ,
2. $\lim_{x\to\pm s} \mathfrak{s}_s^\sharp(x) = 0$,
3. $\mathfrak{s}_s^{\mathrm{ns}}(x) = \mathfrak{s}_s^{\mathrm{sn}}(x)$ *for all* $x \in V_s^{\mathrm{g}} \cup V_s^{\mathrm{s}}$.

The work of Klimeš [14] offers an other, less geometric (but worth mentioning) approach to the question, based on Borel–Laplace transform of the formal normalizing series.

22.11.2 Asymptotic Paths and Canonical Sectors

Definition 11. Let $s \in \Sigma \cup \{0\}$. An **asymptotic path** over $V_s^\sharp$ ending at $p \in \mathcal{U} \cap \left(V_s^\sharp \times \mathbb{C}\right)$ is a regular, smooth curve $\gamma : \mathbb{R}_{\leq 0} \to V_s^\sharp \times \mathbb{C}$ meeting the next requirements.

- $\gamma(0) = p$.
- $\dot{\gamma} = cX_\lambda \circ \gamma$ for some smooth function c. In other words, γ is tangent to X_λ or, equivalently, its image is contained in a single leaf of $\mathcal{F}_\lambda$.
- $\lim_{t \to -\infty} \gamma(t) = s^{\mathrm{n}}$.

We abusively write $\gamma : \left(s^{\mathrm{n}} \to p\right)$ or $\left(s^{\mathrm{n}} \to p\right)$ to stand for such an asymptotic path ending at p.

Theorem 9 (See [28]). *Up to decrease ρ, r, r' the following properties hold for $s \in \Sigma \cup \{0\}$.*

1. *The sets (called **canonical sectors**)*

$$\mathcal{V}_s^\sharp := \left\{p \in \mathcal{U} \cap \left(V_s^\sharp \times \mathbb{C}\right) : \exists \left(s^{\mathrm{n}} \to p\right)\right\} \qquad , \quad \sharp \in \{\mathrm{ns}\,,\ \mathrm{sn}\}$$

   are domains containing smaller sectors $V_s^\sharp \times \tilde{r}\mathbb{D}$ with $\tilde{r} > 0$ independent on s. The union

$$\mathcal{V}_s := \mathcal{V}_s^{\mathrm{ns}} \cup \mathcal{V}_s^{\mathrm{sn}}$$

   is a pointed neighborhood of $\{x = \pm s\} \cap \mathcal{U}$.
2. *Each leaf of $\mathcal{F}_\lambda|_{\mathcal{V}_s^\sharp}$ is simply connected, and if γ and $\tilde{\gamma}$ are two asymptotic paths ending at $p \in \mathcal{V}_s^\sharp$, then the **asymptotic cycle***

$$-\gamma\tilde{\gamma} := \begin{cases} t \leq 0 & \longmapsto \tilde{\gamma}(t) \\ t \geq 0 & \longmapsto \gamma(-t) \end{cases}$$

   *is trivial, in the sense that there exists an **asymptotic tangential homotopy** $h : [-\infty, 0] \times \mathbb{R} \to \mathcal{V}_s^\sharp$ between $\gamma^{-1}\tilde{\gamma}$ and s^{n}, a mapping such that:*
   - *h is uniformly continuous,*
   - *$h(0, \bullet) = \gamma^{-1}\tilde{\gamma}$,*
   - *$h(-\infty, \bullet) = s^{\mathrm{n}}$,*
   - *for every t real $h(\bullet, t) : s^{\mathrm{n}} \to h(0, t)$,*
   - *for every τ negative real $h(\tau, \bullet) : s^{\mathrm{n}} \to p$.*
3. *Take $p \in \mathcal{V}_s^{\mathrm{ns}} \cap \mathcal{V}_s^{\mathrm{sn}}$ and consider an asymptotic cycle $\gamma(p)$ obtained by concatenating two asymptotic paths ending at p, one of which lies in $\mathcal{V}_s^{\mathrm{ns}}$ and the other one in $\mathcal{V}_s^{\mathrm{sn}}$.*

a. If $p \in \mathcal{V}_s^{\mathrm{g}} \cup \mathcal{V}_s^{\mathrm{n}}$, then $\gamma(p)$ is trivial.
b. If $p \in \mathcal{V}_s^{\mathrm{s}}$, then $\gamma(p)$ is trivial if and only if p belongs to the sectorial weak separatrix. Two such asymptotic cycles $\gamma(p)$ and $\tilde{\gamma}(\tilde{p})$ are tangentially homotopic if (and only if) p and $\tilde{p}$ are in the same (sectorial) leaf.

Proof.

1. Straighten the sectorial weak separatrix to $\{y = 0\}$ by applying $\Psi_s : (x, y) \mapsto \left(x, y + \mathfrak{s}_s^{\sharp}(x)\right)$ beforehand. Along the real trajectories $t \leq 0 \longmapsto (x(t), y(t))$ of $\vartheta(s)\Psi_s^* X_\lambda$, a direct variational computation shows that $\phi(t) := |y(t)|^2 = y(t)\overline{y(t)}$ satisfies the differential relation
$$\dot{\phi} = 2\phi\Re(1 + \cdots) > 0\,, \tag{22.36}$$
where $(\cdots)$ stand for a term smaller than $\frac{1}{2}$ in modulus. Actually ϕ is exponentially increasing therefore $t \mapsto y(t)$ tends to 0 as $t \to -\infty$. By construction of the squid sectors $t \mapsto x(t)$ converges towards s^{n}. The argument works well if x remains in a bounded region, which may not be always the case. This is countered by altering slightly $\vartheta(s)$.
2. Because of the previous variational estimate, we can prove that homotopies in the x-space along the flow of $-\vartheta(s) P_\lambda \frac{\partial}{\partial x}$ lift in the leaf through the projection $\Pi : (x, y) \mapsto x$. Indeed the only obstructions preventing the lift is the tangency between X_λ and Π^{-1} (cst), which does not happen close to 0. Since ϕ is decreasing the trajectory remains close to 0. As in [35], it is then possible to prove that any sectorial asymptotic cycle is trivial, by deforming it using such contracting homotopies.
3. The argument is similar to (2) save for the fact that the flow of $-\vartheta(s) P_\lambda \frac{\partial}{\partial x}$ always drive the x-coordinate away while crossing saddle parts. Therefore candidate trivializing homotopies of $\gamma(p)$ in the x-variable cannot be lifted all the way if $p \in \mathcal{V}_s^{\mathrm{s}}$.

22.11.3 Sectorial Solutions to Cohomological Equations

Theorem 10 (See [28]). *For $s \in \Sigma$ define*
$$\mathcal{V}_s^{\mathrm{nsn}} := \mathcal{V}_s \setminus \left(\mathcal{V}_s^{\mathrm{g}} \cup \mathcal{V}_s^{\mathrm{n}}\right)$$
$$\mathcal{D}^{\mathrm{nsn}} := \bigcup_{s \in \Sigma} \{s\} \times \mathcal{V}_s^{\mathrm{nsn}}$$
and take $G_\bullet \in \mathrm{Holo_c}(\mathcal{D}^{\mathrm{nsn}})$ such that $(s, x) \mapsto \frac{G_s(x,0)}{P_\lambda(x)}$ is bounded. For $s \in \Sigma$ and $p \in \mathcal{V}_s^{\sharp}$ define

$$F_s^\sharp(p) := \int_{(s^n \to p)} G_s \frac{\mathrm{d}x}{P_\lambda}$$

where $(s^n \to p)$ *is an asymptotic path of* $X_\bullet$ *ending at* p. *The following properties hold.*

1. $F_s^\sharp(p)$ *is an absolutely convergent integral.*
2. $F_\bullet^\sharp$ *is the unique family of solutions of the cohomological equation* $X_\bullet \cdot F_\bullet = G_\bullet$ *which belongs to* $\mathrm{Holo_c}(\mathcal{D}^\sharp)$ *and vanishes at each* s^n. *Another such solution differs from* $F_\bullet$ *by the addition of a function* $f_\bullet \in \mathrm{Holo_c}(\Sigma)$.
3.
$$F_s^{sn}(p) - F_s^{ns}(p) = \begin{cases} 0 & \text{if } p \in \mathcal{V}_s^n \cup \mathcal{V}_s^g \\ \int_{\gamma(p)} G_s \frac{\mathrm{d}x}{P_\lambda} & \text{if } p \in \mathcal{V}_s^s \end{cases}$$
   where $\gamma(p)$ *is an asymptotic cycle passing through* p.
4. *The following properties are equivalent.*
   a. *There exists* $F_\bullet \in \mathrm{Holo_c}(\mathcal{D}_s)$ *such that* $X_\bullet \cdot F_\bullet = G_\bullet$.
   b. *For all* $s \in \Sigma$ *and* $p \in \mathcal{V}_s^s$
$$\int_{\gamma(p)} G_s \frac{\mathrm{d}x}{P_\lambda} = 0\,.$$

Proof.

1. This follows from the estimate (22.36) since for $t \leq 0$ close to ∞
$$\frac{\mathrm{d}x(t)}{P_\lambda(x(t))} = \vartheta(s)\,\mathrm{d}t$$
$$x(t) = s^n + \mathrm{O}\left(\frac{1}{t}\right)$$
$$y(t) = \mathrm{O}(\exp t)\ .$$

2. The fact that $F_s^\sharp$ is a solution of the cohomological equation can be understood by covering $(s^n \to p)$ with flow-boxes. In such a local rectifying chart $\Psi : \mathcal{D} \to (\mathbb{C}^2, \gamma(t))$ one has to solve
$$\frac{\partial F_s^\sharp \circ \Psi}{\partial x} = G_\lambda \circ \Psi$$
   so that for $p_*,\ q_* \in \mathcal{D}$ in the same leaf of $\mathcal{F}_{\frac{\partial}{\partial x}}|_{\mathcal{D}}$ we have

$$F_s^\sharp \circ \Psi(q_*) - F_s^\sharp \circ \Psi(p_*) = \int_{p_* \to q_*} G_\lambda \circ \Psi \mathrm{d}x$$

and vice versa.

3. By construction $F_s^{\mathrm{sn}}(p) - F_s^{\mathrm{ns}}(p) = \int_{\gamma(p)} G_s \frac{\mathrm{d}x}{P_\lambda}$. Since all asymptotic cycles are trivial when $p \in \mathcal{V}_s^{\mathrm{n}} \cup \mathcal{V}_s^{\mathrm{g}}$ (Theorem 9) then the integral $\int_{\gamma(p)} G_s \frac{\mathrm{d}x}{P_\lambda}$ vanishes because of Cauchy formula.
4. As a consequence of the previous items, $\int_{\gamma(p)} G_s \frac{\mathrm{d}x}{P_\lambda}$ encodes the additive monodromy of the analytic continuation of $F_s^\sharp$ so that (a)⇒(b) Riemann's removable singularity theorem yields the converse implication.

22.11.4 Sectorial Normalization and Space of Leaves

From Theorem 8 and Proposition 4 we obtain a vector field X_s and a right-hand side $G_s := -P_\lambda R_s^\sharp$ with

$$R_s^\sharp : (x,y) \in \mathcal{V}_s^\sharp \longmapsto \frac{R_\lambda(x,y) - R_\lambda\left(x, \mathfrak{s}_s^\sharp(x)\right)}{y - \mathfrak{s}_s^\sharp(x)} \tag{22.37}$$

satisfying the hypothesis of Theorem 10. The orbital normalization equation (22.29) therefore admits solutions $O_\bullet^\sharp \in \mathrm{Holo_c}\left(\mathcal{D}^\sharp\right)$. The situation is the same for the temporal normalization with $\widehat{X}_s := X_\lambda$ and $G_\lambda := \frac{1}{U_\lambda} - \frac{1}{\mathrm{u}_\lambda}$, and (22.30) admits solutions $T_\bullet^\sharp \in \mathrm{Holo_c}\left(\mathcal{D}^\sharp\right)$.

Corollary 3. *$Z_\bullet$ is conjugate to its formal model $\mathrm{u}_\bullet \mathrm{X}_\bullet^\infty$ on $\mathcal{D}^\sharp$. It is orbitally conjugate to $\mathrm{X}_\bullet^\infty$ by (λ, x)-fibered transformations.*

The model $\mathrm{X}_\lambda^\infty$ has a first integral $H_\lambda^{\infty,\sharp} \in \mathrm{Holo}\left(\mathcal{V}^\sharp\right)$ with connected fibers

$$H_\lambda^{\infty,\sharp}(x,y) := y P_\lambda^{-\frac{\mu_\lambda}{2}}(x) \left(\frac{x+s}{x-s}\right)^{\frac{1}{2s}} \quad , \quad s \neq 0$$

$$H_0^{\infty,\sharp}(x,y) := y x^{-\mu} \exp \frac{1}{x},$$

with $\lim_{\lambda \to 0} H_\lambda^{\infty,\sharp} = H_0^{\infty,\sharp}$ uniformly on compact subsets; for the remaining of the section, we only deal with $s \neq 0$ to lighten notations, although everything is valid for $s = 0$ by continuity. Recalling notations and conventions of Sect. 22.4.3, we choose determinations of the multivalued functions involved above:

- $\left(\frac{\bullet+s}{\bullet-s}\right)^{\frac{1}{2s}} = \mathfrak{g}_\lambda$ coincides with the principal determination on V_s^{n} (the one given in Remark 6 for real s); note that $\mathfrak{g}_\lambda \in \mathrm{Holo}\left(\mathbb{C}\backslash V_s^{\mathrm{g}}\right)$,
- $P_\lambda^{-\frac{\mu_\lambda}{2}}$ agrees with the principal determination on V_s^{n}.

Invoking Corollary 3 above and formulas given in Sect. 22.8.3, we obtain **canonical sectorial first integrals** with connected fibers

$$H_s^\sharp(x,y) := H_\lambda^{\infty,\sharp}\left(x, \left(y - \mathfrak{s}_s^\sharp(x)\right) \exp O_s^\sharp(x,y)\right) \tag{22.38}$$
$$= \left(y - \mathfrak{s}_s^\sharp\right) P_\lambda^{-\frac{\mu_\lambda}{2}}(x)\, \mathfrak{g}_\lambda(x) \exp O_s^\sharp(x,y) \ .$$

The next result can be proved by studying the linearized system on squid sectors as in Lemma 7.

Theorem 11 (See [28]). *For $s \in \Sigma$ define the **sectorial spaces of leaves** as*

$$\Omega_s^\sharp := H_s^\sharp\left(\mathcal{V}_s^\sharp\right) \qquad , \sharp \in \{\mathrm{ns}\ ,\ \mathrm{sn}\}$$
$$\Omega_s^\flat := H_s^{\mathrm{ns}}\left(\mathcal{V}_s^\flat\right) \qquad , \flat \in \{\mathrm{n}\ ,\ \mathrm{g}\ ,\ \mathrm{s}\} \ .$$

1. $\Omega_s^{\mathrm{n}} = \Omega_s^{\mathrm{ns}} = \Omega_s^{\mathrm{sn}} = \Omega_s^{\mathrm{g}} = \mathbb{C}$, *the sectorial weak separatrices corresponding to* 0.
2. $\Omega_s^{\mathrm{s}} = (\mathbb{C}, 0)$, *the sectorial weak separatrices corresponding to* 0. *The size of Ω_s^{s} goes to* 0 *as r or r' does.*

22.11.5 Classification

22.11.5.1 Orbital Necklace

Take $s \in \Sigma \cup \{0\}$. A given point $p \in \mathcal{V}_s^{\mathrm{ns}} \cap \mathcal{V}_s^{\mathrm{sn}}$ corresponds to a point $h^{\mathrm{ns}} \in \Omega_s^{\mathrm{ns}}$ and a point $h^{\mathrm{sn}} \in \Omega_s^{\mathrm{sn}}$. These points must be identified in order to encode the local orbital class of X_λ. Because each $H_s^\sharp$ has connected fibers, the (holomorphic) identifications must be injective (Fig. 22.16).

- If $s \neq 0$ and $p \in \mathcal{V}_s^{\mathrm{g}}$, then every function involved in (22.38) for $\sharp = \mathrm{ns}$ and $\sharp = \mathrm{sn}$ coincides at p except $\mathfrak{g}_\lambda P_\lambda^{-\frac{\mu_\lambda}{2}}$. The monodromy of $\mathfrak{g}_\lambda P_\lambda^{-\frac{\mu_\lambda}{2}}$ around s^{n} acts as

$$h^{\mathrm{sn}} = h^{\mathrm{ns}} \exp \mathrm{i}\pi \left(\frac{1}{s} + \mu_\lambda\right) =: \psi_s^{\mathrm{g}}\left(h^{\mathrm{ns}}\right) \ .$$

- If $p \in \mathcal{V}_s^{\mathrm{n}}$, then every function involved in (22.38) for $\sharp = \mathrm{ns}$ and $\sharp = \mathrm{sn}$ coincides at p except $\mathfrak{s}_s^\sharp$. Because $\Omega_s^{\mathrm{n}} = \mathbb{C}$ the mapping ψ_s^{n} must be affine, and

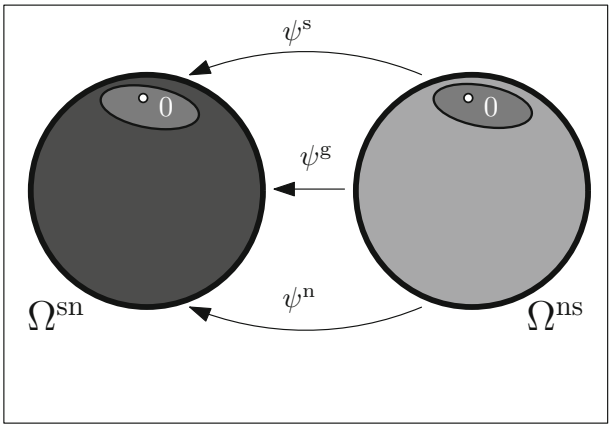


Fig. 22.16 Orbital sectorial identifications

it is not hard to check it is a translation:

$$h^{\mathrm{sn}} = h^{\mathrm{ns}} + \varphi_s^n =: \psi_s^{\mathrm{n}}\left(h^{\mathrm{ns}}\right) . \tag{22.39}$$

- If $p \in \mathcal{V}_s^{\mathrm{s}}$, then every function involved in (22.38) for $\sharp = \mathrm{ns}$ and $\sharp = \mathrm{sn}$ coincides at p except $P_\lambda^{-\frac{\mu_\lambda}{2}} \exp O_s^\sharp$. Because 0 corresponds to both sectorial weak separatrices, and because those two functions agree on V_s^{s} the mapping ψ_s^{s} must fix the point 0:

$$h^{\mathrm{sn}} = h^{\mathrm{ns}} \exp 2\mathrm{i}\pi\mu_\lambda + \mathrm{o}\left(h^{\mathrm{ns}}\right) =: \psi_s^{\mathrm{s}}\left(h^{\mathrm{ns}}\right) \in \mathrm{Diff}\left(\mathbb{C}, 0\right) . \tag{22.40}$$

Definition 12. Let $\mu_\bullet \in \mathbb{C}\{\lambda\}$ be given.

1. Take

$$\varphi_\bullet := \left(\varphi_\bullet^{\mathrm{n}}, \varphi_\bullet^{\mathrm{s}}\right) \in \mathrm{Holo_c}\left(\Sigma\right) \times \mathrm{Holo_c}\left(\Sigma \times (\mathbb{C}, 0)\right)$$

with $\varphi_\bullet^{\mathrm{s}}(0) = 0$. We call **orbital necklace** associated with $\varphi_\bullet$ (and $\mu_\bullet$ implicitly) the complex manifold $\Omega(\varphi_\bullet)$ obtained by the analytic atlas consisting in two copies Ω^{ns} and Ω^{sn} of $(\Sigma \cup \{0\}) \times \mathbb{C}$, with transition maps

$$\begin{aligned} \psi_\bullet^{\mathrm{n}} \ : \ \Omega^{\mathrm{ns}} &\longrightarrow \Omega^{\mathrm{sn}} \\ (s, h) &\longmapsto \left(s, h + \varphi_s^{\mathrm{n}}\right) \end{aligned}$$

$$\psi_\bullet^{\mathrm{s}} \ : \ \left(\Omega^{\mathrm{ns}}, 0\right) \longrightarrow \left(\Omega^{\mathrm{sn}}, 0\right)$$

$$(s,h) \longmapsto \left(s, h\exp\left(2\mathrm{i}\pi\mu_\lambda + \varphi_s^{\mathrm{s}}(h)\right)\right) .$$

2. A **diffeomorphism** between two necklaces $\Omega(\varphi_\bullet)$ and $\Omega(\tilde{\varphi}_\bullet)$ is the data $\left(\Psi_\bullet^{\mathrm{ns}}, \Psi_\bullet^{\mathrm{sn}}\right)$ of s-fibered injective, holomorphic mappings inducing a conjugacy between atlases:

$$\Psi_\bullet^{\mathrm{ns}} \circ \psi_\bullet^{\flat} = \tilde{\psi}_\bullet^{\flat} \circ \Psi_\bullet^{\mathrm{sn}} \quad , \quad \flat \in \{\mathrm{n}\ ,\ \mathrm{s}\}$$

and such that for $s \neq 0$

$$\Psi_\bullet^{\mathrm{ns}} \circ \psi_\bullet^{\mathrm{g}} = \psi_\bullet^{\mathrm{g}} \circ \Psi_\bullet^{\mathrm{sn}} \tag{22.41}$$

where

$$\psi_s^{\mathrm{g}}(h) := h\exp\mathrm{i}\pi\left(\frac{1}{s} + \mu_\lambda\right) .$$

We say that the necklaces are **analytically conjugate**.

3. Let $Z_\bullet$ be a generic unfolding of formal orbital class $\mu_\bullet$ as in Theorem 7. We call **orbital class** of $Z_\bullet$ the necklace $\mathrm{Orb}(Z_\bullet) := \Omega(\varphi_\bullet)$ where $\varphi_\bullet$ is built from the mappings (22.39) and (22.40).

Remark 14. We do not take the gate-part identification into account to build the necklace because the construction does not make it at the limit. It is not needed anyway to perform the local classification, because both sectorial normalizing maps always glue on $\mathcal{V}_s^{\mathrm{g}}$. However the actual space of leaves Ω_s on $\mathcal{V}_s$ is the quotient of the corresponding s-fiber of $\mathrm{Orb}(Z_\bullet)$ by ψ_s^{g} for $s \neq 0$. This is why we must include the condition (22.41) to capture all information relating to orbital conjugacy.

Because holomorphic automorphisms of the complex line are rigid there are not many necklaces diffeomorphisms. Hence the orbital necklace of an unfolding is nearly a local invariant, and we prove in the section that it suffices to characterize its local class.

Lemma 10. *$\left(\Psi_\bullet^{\mathrm{ns}}, \Psi_\bullet^{\mathrm{sn}}\right)$ is a diffeomorphism of necklaces if and only if there exists $c_\bullet \in \mathrm{Holo}_{\mathrm{c}}(\Sigma)^\times$ such that $\Psi_s^\sharp(h) = c_s h$.*

Proof. One direction is trivial. Assume then that $\left(\Psi_\bullet^{\mathrm{ns}}, \Psi_\bullet^{\mathrm{sn}}\right)$ is a diffeomorphism between necklaces. For each $s \in \Sigma \cup \{0\}$ the mapping $\Psi_s^\sharp$ is biholomorphic, thus an invertible linear map $h \mapsto c_s^\sharp h$. The conjugacy equation $\Psi_s^{\mathrm{ns}} \circ \psi_s^{\mathrm{g}} = \psi_s^{\mathrm{g}} \circ \Psi_s^{\mathrm{sn}}$ implies $c_s^{\mathrm{ns}} = c_s^{\mathrm{sn}}$ for $s \neq 0$, thus also for $s = 0$ by continuity. □

22.11.5.2 Integral Representation of the Saddle Orbital Invariant

Thanks to Theorem 8 we know the sectorial separatrices $\mathfrak{s}_s^\sharp$ glue to a holomorphic function $\mathfrak{s}_s^{\mathrm{nsn}}$ on V_s^{nsn}, therefore so does $R_s^\sharp$ appearing in (22.37), yielding a function $R_\bullet^{\mathrm{nsn}} \in \mathrm{Holo_c}\left(\mathcal{D}^{\mathrm{nsn}}\right)$.

Proposition 5. *Let $G_\bullet \in \mathrm{Holo_c}\left(\mathcal{D}^{\mathrm{nsn}}\right)$ such that $(s,x) \mapsto \frac{G_s(x,0)}{P_\lambda(x)}$ is bounded, and let $\mathrm{Orb}\left(X_\bullet\right)$ be the orbital necklace of $X_\bullet$. The mapping (given in the chart Ω^{ns})*

$$\begin{aligned}\mathfrak{T}_\bullet\left(G_\bullet\right) \;:\; \left(\mathrm{Orb}\left(X_\bullet\right),0\right) &\longrightarrow \mathbb{C}\\ (s,h) &\longmapsto \frac{1}{2\mathrm{i}\pi}\int_{\gamma(p)} G_s \frac{\mathrm{d}x}{P_\lambda}\;,\end{aligned}$$

where $\gamma(p)$ is an asymptotic cycle defined in Theorem 9 Item (2). such that $H_s^{\mathrm{ns}}(p) = h$, is well defined and vanishes along $\{h=0\}$. This mapping defines the (linear) ***period operator***

$$\mathfrak{T}_\bullet \;:\; G_\bullet \longmapsto \mathfrak{T}_\bullet\left(G_\bullet\right) \in \mathrm{Holo_c}\left(\mathrm{Orb}\left(X_\bullet\right),0\right)\;.$$

Proof. Because $X_\lambda \cdot F_s^{\mathrm{ns}} = X_\lambda \cdot F_s^{\mathrm{sn}} = G_\lambda$ the difference $F_s^{\mathrm{sn}} - F_s^{\mathrm{ns}}$ is a first integral of X_λ, and therefore factors as a map τ_s defined on the sectorial space of leaves Ω_s^{s}. This can also be seen from Theorem 9. Indeed the value of the integral depends only on the asymptotic tangential homotopy class of $\gamma(p)$, as an asymptotic tangential homotopy is uniformly continuous. Hence only asymptotic cycles with $p \in \mathcal{V}_s^{\mathrm{s}}$ contribute to the period, and the value of the integral only depends on the sectorial leaf $\mathcal{F}_\lambda|_{\mathcal{V}_s^{\mathrm{s}}}$, since there is at most one non-trivial homotopy class of asymptotic cycles per leaf. The same argument shows that $\tau_s(0) = 0$, for any asymptotic cycle within the sectorial separatrix is trivial. □

The transition map $\psi_\bullet^{\mathrm{s}}$ of the orbital necklace of $X_\bullet$ obeys the identity

$$H_s^{\mathrm{sn}}(p) = \psi_s^{\mathrm{s}}\left(H_s^{\mathrm{ns}}(p)\right)$$

while at the same time

$$H_s^\sharp(p) = H_s^{\infty,\sharp}(p)\exp O_s^\sharp(p)\;,$$

following (22.38). For $p \in \mathcal{V}_s^{\mathrm{s}}$ we have $H_s^{\infty,\mathrm{sn}}(p) = H_s^{\infty,\mathrm{ns}}(p)\exp 2\mathrm{i}\pi\mu_\lambda$ so that

$$H_s^{\mathrm{sn}}(p) = H_s^{\mathrm{ns}}(p)\exp\left(2\mathrm{i}\pi\mu_\lambda + O_s^{\mathrm{sn}}(p) - O_s^{\mathrm{ns}}(p)\right)\;.$$

Recalling how we just defined the period operator, we obtain an integral representation for the saddle-part of the orbital invariant.

Corollary 4. *Let $\Omega\left(\varphi_\bullet^{\mathrm{n}}, \varphi_\bullet^{\mathrm{s}}\right)$ be the orbital necklace of $X_\bullet$. Then for all $s \in \Sigma \times \{0\}$*

$$\varphi_s^{\mathrm{s}} = -2\mathrm{i}\pi \mathfrak{T}_s \left(P_\bullet R_\bullet^{\mathrm{nsn}}\right) .$$

22.11.5.3 Temporal Invariant

Definition 13. Let $\mu_\bullet \in \mathbb{C}\{\lambda\}$ and $\mathrm{u}_\bullet \in \mathbb{C}\{\lambda\}[x]^\times_{\leq 1}$ be given.

1. The data of an orbital necklace $\Omega(\varphi_\bullet)$ and

$$f_\bullet \in \mathrm{Holo_c}\left(\Omega\left(\varphi_\bullet\right), 0\right)$$

   with $f_\bullet(0) = 0$ is called **temporal necklace.**
2. A **diffeomorphism** between two temporal necklaces $(\varphi_\bullet, f_\bullet)$ and $\left(\tilde{\varphi}_\bullet, \tilde{f}_\bullet\right)$ is the data $\left(\Psi_\bullet^{\mathrm{ns}}, \Psi_\bullet^{\mathrm{sn}}\right)$ of a diffeomorphism between corresponding orbital necklaces and satisfying:

$$f_\bullet = \tilde{f}_\bullet \circ \Psi_\bullet^{\mathrm{sn}} .$$

   We say that the necklaces are **analytically conjugate**.
3. Let $Z_\bullet = U_\bullet X_\bullet$ be a generic unfolding under prepared form (22.26) with formal invariants $(\mu_\bullet, \mathrm{u}_\bullet)$ as in Theorem 7. We call **local class** of $Z_\bullet$ the temporal necklace

$$\mathrm{Class}\left(Z_\bullet\right) := \left(\mathrm{Orb}\left(X_\bullet\right), \mathfrak{T}_\bullet\left(\frac{1}{U_\bullet} - \frac{1}{\mathrm{u}_\bullet}\right)\right)$$

   where the period operator $\mathfrak{T}_\bullet$ is defined in Proposition 5.

22.11.5.4 Classification Theorem

As prompted by Lemma 10 we define the action of $c_\bullet \in \mathbb{C}\{\lambda\}^\times$ on functions $f_\bullet \in \mathrm{Holo_c}\left(\Sigma \times (\mathbb{C}, 0)\right)$ by

$$c_\bullet^* f_\bullet \ : \ (s, x, y) \longmapsto f_s\left(x, c_\lambda y\right) .$$

The action is extended component-wise to tuples of functions.

Theorem 12 (See [28]). *Two generic unfoldings of codimension* 1 *in the same formal class $(\mu_\bullet, \mathrm{u}_\bullet) \in \mathbb{C}\{\lambda\} \times \mathbb{C}\{\lambda\}[x]^\times_{\leq 1}$ are locally (*resp. *orbitally) equivalent if and only if their temporal (*resp. *orbital) necklaces are analytically conjugate. In other words we have local classifications*

$$\mathrm{Orb} \ : \ \mathrm{Mod}^{\mathrm{orb}}_{\mathrm{loc}}(1) \longrightarrow \mathrm{Holo_c}\left(\Sigma\right) \times \mathrm{Holo_c}\left(\Sigma \times (\mathbb{C}, 0)\right)_{/\mathbb{C}\{\lambda\}^\times}$$

$$\text{Class} : \text{Mod}_{\text{loc}}(1) \longrightarrow \text{Holo}_{\text{c}}(\Sigma) \times \text{Holo}_{\text{c}}(\Sigma \times (\mathbb{C},0)) \times \text{Holo}_{\text{c}}(\Sigma \times (\mathbb{C},0))/_{\mathbb{C}\{\lambda\}^\times}.$$

Proof. Let us present only the orbital part of the proof, the temporal part being easier according to Theorem 10 Item (4). and Proposition 4 Item (1). One way is clear: if $Z_\bullet$ and $\widetilde{Z}_\bullet$ are locally orbitally conjugate by Ψ, then the conjugacy factors as a diffeomorphism $\left(\Psi_\bullet^{\text{ns}}, \Psi_\bullet^{\text{sn}}\right)$ between the orbital necklaces $\text{Orb}(Z_\bullet)$ and $\text{Orb}\left(\widetilde{Z}_\bullet\right)$. Because the conjugacy is analytic in λ the induced linear change of coordinate $h \mapsto c_\bullet h$ in leaves space is also analytic in λ. Conversely, assume the existence of a diffeomorphism $\left(\Psi_\bullet^{\text{ns}}, \Psi_\bullet^{\text{sn}}\right)$ conjugating the necklaces $\text{Orb}(Z_\bullet)$ and $\text{Orb}\left(\widetilde{Z}_\bullet\right)$. Invoking Lemma 10 we can apply the rescaling $(s,x,y) \mapsto (s,x,c_s y)$ to $\widetilde{Z}$. Without changing notations, in the new coordinates we must have $\Psi_s^\sharp = \text{Id}$ for each $s \in \Sigma \cup \{0\}$. This particularly implies the identities $\varphi_s^\flat = \tilde{\varphi}_s^\flat$ for $\flat \in \{\text{n}, \text{s}\}$. But these quantities measure the obstruction to glue over $\mathcal{V}_s^{\text{ns}} \cap \mathcal{V}_s^{\text{sn}}$ the transitions between corresponding sectorial normalization mappings $\mathcal{Y}_s^\sharp := \tilde{\mathcal{O}}_s^\sharp \circ \left(\mathcal{O}_s^\sharp\right)^{\circ -1}$. Hence $\Psi_s|_{\mathcal{V}_s^{\text{ns}}} := \mathcal{Y}_s^{\text{ns}}$ and $\Psi_s|_{\mathcal{V}_s^{\text{sn}}} := \mathcal{Y}_s^{\text{sn}}$ defines a holomorphic conjugacy between X_λ and $\widetilde{X}_\lambda$ on $\mathcal{V}_s = \mathcal{V}_s^{\text{ns}} \cup \mathcal{V}_s^{\text{sn}}$. This mapping is bounded, thus extends biholomorphically to $\text{adh}(\mathcal{V}_s)$ by Riemann's removable singularity theorem.

To conclude the proof we only need to check that $\Psi_s = \Psi_{-s}$. Define $\mathcal{S}_s := \Psi_s^{\circ -1} \circ \Psi_{-s}$ which, by construction, is a symmetry of X_λ, that is $\mathcal{S}_s^* X_\lambda = X_\lambda$. A direct computation at a formal level on $\mathcal{S}_s(x,y) = \left(x\,,\ y + \sum_{n+m>1} S_{n,m}(s)\, x^n y^m\right)$ establishes $\mathcal{S}_s = \text{Id}$. □

The classification presented above cannot be complete, as we explain in the upcoming Sect. 22.12.4.

22.12 Dynamical Interpretation of the Orbital Necklace

We describe the relationship between the actual dynamics of X_λ (the holonomy of the underlying foliation $\mathcal{F}_\lambda$) and what could be called the *necklace dynamics*. We define in the first place what we mean by "the dynamics of X_λ" in Sect. 22.12.1. It encodes the "monodromy" of the canonical first integrals $H_s^\sharp$, which really is what the orbital necklace is all about. As X_λ does not depend on the choice of s or $-s$, the splitting of $\mathcal{U}$ into squid sectors is artificially superimposed on the dynamics. By rewording this fact as a relationship between orbital necklaces $\Omega(\varphi_s)$ and $\Omega(\varphi_{-s})$, we derive the orbital compatibility condition in Sect. 22.12.3. The temporal compatibility condition will be derived while performing the temporal realization in Sect. 22.13.1.

We finally use the compatibility condition to characterize even necklaces, corresponding to invariants $\varphi_\bullet = \left(\varphi_\bullet^{\text{n}}, \varphi_\bullet^{\text{s}}\right)$ holomorphic as a function of the parameter λ. We show this configuration to be very rare, underlying the lack of completeness of the classification provided by Theorem (12).

22.12.1 Weak Holonomy

Fix once and for all $x_* \in V_s^{\mathrm{n}} \backslash \rho \overline{\mathbb{D}}$ (by construction of the squid sector in Definition 10 this domain is independent on $s \in \Sigma$) and a transverse disc

$$T := \{x = x_*\} \cap \bigcap_{s \in \Sigma} \mathcal{V}_s^{\mathrm{s}} = \{x_*\} \times (\mathbb{C}, 0) \ .$$

Because X_λ is in the prepared form (22.26) its integral curves are everywhere transverse to the fibers of the natural projection

$$\Pi \ : \ (x, y) \longmapsto x$$

outside $P_\lambda^{-1}(0)$. As a consequence we can lift (smooth) paths in the punctured base $\gamma \ : \ [0, 1] \to \mathbb{D}_\lambda$,

$$\mathbb{D}_\lambda := r\mathbb{D} \backslash P_\lambda^{-1}(0) \ ,$$

through Π into leaves of $\mathcal{F}_\lambda$ and starting from points in T. More precisely, being given $p_* := (x_*, y_*) \in T$ there exists a unique (germ at 0 of a) solution

$$t \longmapsto \gamma_{p_*}(t) = (\gamma(t), y(t))$$

to the constrained flow-system

$$\dot{\gamma}_{p_*} = \frac{\dot{\gamma}}{P_\lambda \circ \gamma} X_\lambda \circ \gamma_{p_*} \quad , \ \gamma_{p_*}(0) = p_* \ .$$

Notice that the image of γ_{p_*} is included in a single leaf of $\mathcal{F}_\lambda$. Of course if γ is "too long" then γ_{p_*} may eventually escape from $\mathcal{U}$. On the contrary if γ_{p_*} is defined on the whole $[0, 1]$ we call

$$\mathfrak{h}_\lambda^\gamma(p_*) := \gamma_{p_*}(1)$$

the image of p_* by the **holonomy** $\mathfrak{h}^\gamma$ of $\mathcal{F}_\lambda$ along the path γ. The holonomy $\mathfrak{h}_\lambda^\gamma$ is holomorphic and locally invertible. When γ is a loop the holonomy defines a germ of a biholomorphic self-map $(T, p_*) \to \left(T, \mathfrak{h}_\lambda^\gamma(p_*)\right)$ of the transversal (Fig. 22.17).

We are particularly interested in the case where γ is a generator γ_s or Γ of the fundamental group $\pi_1(\mathbb{D}_\lambda, x_*)$ when $s \neq 0$.

Definition 14. Let $s \in \Sigma$. Consider a system $\{\Gamma, \gamma_s\}$ of generators of $\pi_1(\mathbb{D}_\lambda, x_*)$ such that $\Gamma = |x_*| \mathbb{S}^1$ and γ_s winds directly once around s and does not around $-s$. The holonomy

$$\mathfrak{h}_s^{\mathrm{n}} := \mathfrak{h}_\lambda^{\gamma_s}$$

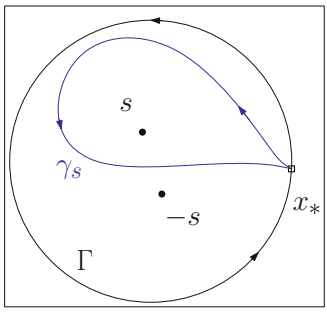


Fig. 22.17 Generators of $\pi_1\left(\mathbb{D}_\lambda, x_*\right)$ for $s \in \Sigma$

is called the (**weak**) **nodal holonomy**. Similarly we name (**weak**) **holonomy** the mapping $\mathfrak{h}_\lambda^\Gamma$.

We state a consequence of Theorem 8.

Lemma 11. *The nodal holonomy $\mathfrak{h}_s^{\mathrm{n}}$ is an injective holomorphic map on a subdomain T_s of T containing both points of intersection of T with the sectorial weak separatrices $\left\{y = \mathfrak{s}_s^{\mathrm{ns}}(x)\right\} \cup \left\{y = \mathfrak{s}_s^{\mathrm{sn}}(x)\right\}$.*

22.12.2 Necklace Holonomy

While walking along a loop $\gamma : [0,1] \to \mathbb{D}_\lambda$ and lifting it in the foliation $\mathcal{F}_\lambda$ to build the holonomy $\mathfrak{h}_\lambda^\gamma$, one may follow what happens in the orbital necklace. More precisely, since the image of γ in the (global) leaves space corresponds with just a point, one may wish to understand the "trajectory" induced by γ in the orbital necklace $\Omega\left(\varphi_\bullet\right)$. We recall that a base-point $x_* = \gamma(0)$ is fixed once and for all now. The restriction of the canonical first integral to the transverse disc T induces an invertible mapping

$$H_s : (x_*, y) \in T_s \longmapsto H_s^{\mathrm{ns}}(x_*, y) \in \left(\Omega^{\mathrm{ns}}, 0\right)$$

whose image contains $\left\{0, \varphi_s^{\mathrm{n}}\right\}$ (Lemma 11). Starting from $h_0 := H_s\left(p_*\right)$ we build a sequence of points $(h_n)_{0 \le n \le 2d}$ such that $h_{2m} \in \Omega_s^{\mathrm{ns}}$ and $h_{2m+1} \in \Omega_s^{\mathrm{sn}}$, the connection between h_{n-1} and h_n being given by the action of $\left(\psi^{\flat_n}\right)^{\circ \pm 1}$ for $\flat_n \in \{\mathrm{n}, \mathrm{g}, \mathrm{s}\}$ corresponding to the connected component $V_s^\flat$ being crossed by γ, the sign being determined by whether the path leaves V_s^{ns} ("+") or enters it ("−"), as long as the partial lift of γ in $\mathcal{F}_\lambda$ is defined. We name Δ_s^γ the **necklace holonomy**

$$\Delta_s^\gamma : h_0 \longmapsto h_{2d}\,.$$

The maps Δ_s^γ and $\mathfrak{h}^\gamma$ represent the same dynamics since they are conjugate:

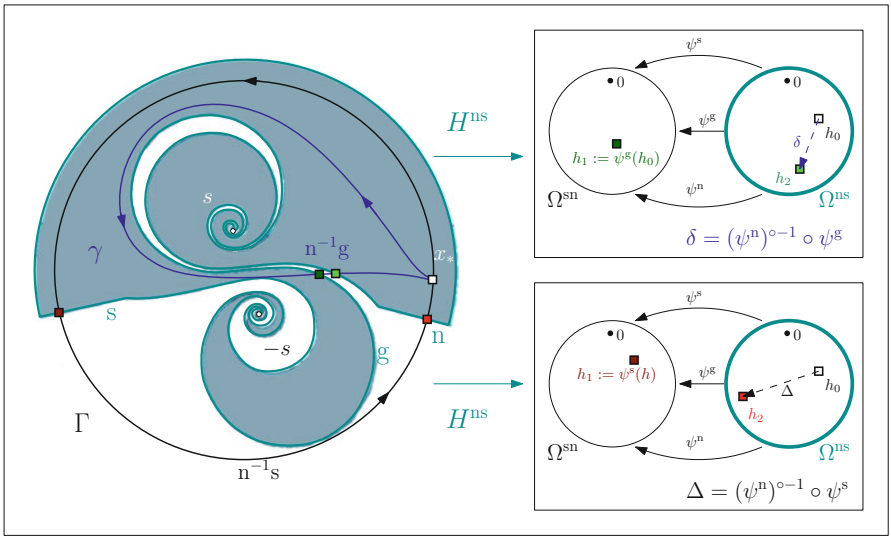


Fig. 22.18 The dynamics induced by the weak holonomies in the orbital necklace

$$H_s^* \, \Delta_s^\gamma = \mathfrak{h}_\lambda^\gamma \, . \tag{22.42}$$

We refer to Fig. 22.18 for a depiction of this construction in the case of the weak holonomies $\mathfrak{h}_s^{\rm n}$ and $\mathfrak{h}_\lambda^\Gamma$.

Let us generalize the construction to abstract orbital necklaces $\Omega\,(\varphi_\bullet)$. Let $\mathfrak{W}$ be the free group on the letters $\{\mathrm{n}, \mathrm{g}, \mathrm{s}\}$. For every $s \in \Sigma$ there exists a group morphism

$$\begin{aligned} \mathfrak{w}_s \;:\; \pi_1\left(\mathbb{D}_\lambda, x_*\right) &\longrightarrow \mathfrak{W} \\ \gamma &\longmapsto \mathfrak{b}_1^{\epsilon_1} \circ \cdots \circ \mathfrak{b}_{2d}^{\epsilon_{2d}} \qquad , \; \mathfrak{b}_j \in \{\mathrm{n}, \mathrm{g}, \mathrm{s}\} \;,\; \epsilon_j \in \{\pm 1\} \end{aligned}$$

defined in such a way that $\left(\mathfrak{b}_j, \epsilon_j\right)_{j \le 2d}$ is the sequence obtained as before: $\mathfrak{b}_j$ corresponds to the connected component $V_s^{\mathfrak{b}_j}$ currently crossed by γ, the sign being determined by whether the path leaves $V_s^{\rm ns}$ or enters it. For instance,

$$\begin{aligned} \mathfrak{w}_s\left(\gamma_s\right) &= \mathrm{n}^{-1}\mathrm{g} =: \hat{\gamma} \\ \mathfrak{w}_s\left(\Gamma\right) &= \mathrm{n}^{-1}\mathrm{s} =: \widehat{\Gamma} \, . \end{aligned}$$

We omit the proof of the following lemma.

Lemma 12.

1. For every $\gamma \in \pi_1\left(\mathbb{D}_\lambda, x_*\right)$ the word $\mathfrak{w}_s\left(\gamma\right)$ has the form

$$\mathfrak{w}_s(\gamma) = \prod_{j=1}^{d} \mathfrak{b}_{j,1}^{-1} \mathfrak{b}_{j,0} \ .$$

2. *The image* $\mathfrak{W}$ *of* $\mathfrak{w}_s$ *is generated by the words* $\mathrm{n}^{-1}\mathrm{g}$ *and* $\mathrm{n}^{-1}\mathrm{s}$*:*

$$\mathfrak{W} = \left\langle \hat{\gamma} \ , \ \widehat{\Gamma} \right\rangle \ .$$

3. *The mapping* $\mathfrak{w}_s \ : \ \pi_1(\mathbb{D}_\lambda, x_*) \to \mathfrak{W}$ *is bijective. For any* $s \in \Sigma$ *we write*

$$\begin{aligned} \mathfrak{p}_s \ : \ \mathfrak{W} &\longrightarrow \pi_1(\mathbb{D}_\lambda, x_*) \\ \hat{\gamma} &\longmapsto \gamma_s \\ \widehat{\Gamma} &\longmapsto \Gamma \end{aligned}$$

its inverse.

Definition 15. Being given an orbital necklace $\Omega(\varphi_\bullet)$ we build a dynamics in the following manner. Take $s \in \Sigma$ and $w = \mathfrak{b}_1^{\epsilon_1} \circ \cdots \circ \mathfrak{b}_{2d}^{\epsilon_{2d}} \in \mathfrak{W}$, then denote $\gamma := \mathfrak{p}_s(w)$. We define the **necklace holonomy** associated with γ (or w) as the following symbolic expression

$$\Delta_s^\gamma := \bigcirc_{1 \leq j \leq 2d} \left(\psi_s^{\mathfrak{b}_j} \right)^{\circ \epsilon_j} \ .$$

Depending on γ the expression Δ_s^γ may not represent an actual germ of a diffeomorphism, because $\mathfrak{h}_\lambda^\gamma$ may not be geometrically defined. For that reason the map $\Delta_s^\bullet$ ranges in the pseudogroup of local diffeomorphisms of Ω_s^{ns}. The **necklace holonomy representation** is the collection $\Delta_\bullet = (\Delta_s)_{s \in \Sigma}$ of (pseudo-) group morphisms

$$\Delta_s \ : \ w \in \mathfrak{W} \longmapsto \Delta_s(w) := \Delta_s^{\mathfrak{p}_s(w)} \ .$$

22.12.3 Orbital Compatibility Condition

Define

$$\begin{aligned} \Sigma^\cap &:= \Sigma \cap (-\Sigma) \\ &= \{ s \in \Sigma \ : \ -s \in \Sigma \} \\ &= \Sigma^+ \cup \Sigma^- \\ \Sigma^\pm &:= \Sigma^\cap \cap \{ \pm \Im(s) > 0 \} \end{aligned}$$

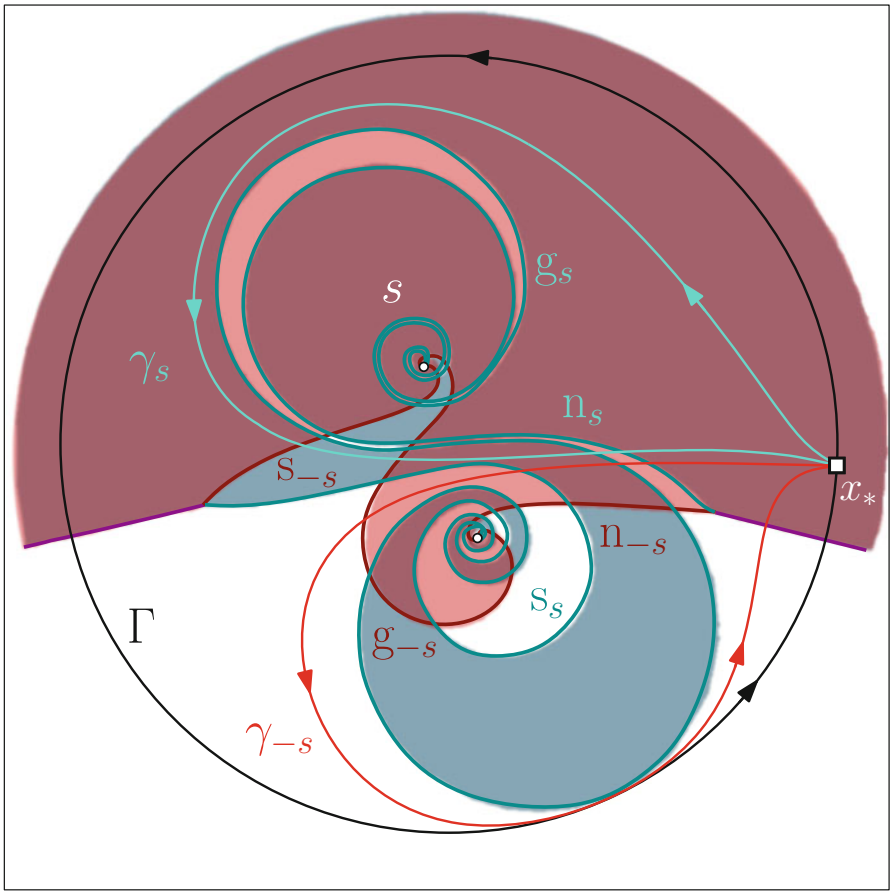


Fig. 22.19 Comparing the necklace dynamics

the union of two domains on which we can compare the necklace dynamics for s and $-s$. According to (22.42) and Fig. 22.19 we have for $s \in \Sigma^{\cap}$ the identities between actual diffeomorphisms

$$\begin{cases} H_s^* \Delta_s \left(\mathrm{n}^{-1}\mathrm{g}\right) = H_{-s}^* \Delta_{-s} \left(\mathrm{g}^{-1}\mathrm{s}\right) & \text{if } s \in \Sigma^+ \\ H_s^* \Delta_s \left(\mathrm{n}^{-1}\mathrm{g}\right) = H_{-s}^* \Delta_{-s} \left(\mathrm{n}^{-1}\mathrm{s}\mathrm{g}^{-1}\mathrm{n}\right) & \text{if } s \in \Sigma^- \\ H_s^* \Delta_s \left(\mathrm{n}^{-1}\mathrm{s}\right) = H_{-s}^* \Delta_{-s} \left(\mathrm{n}^{-1}\mathrm{s}\right) \end{cases} . \tag{22.43}$$

In order to motivate the definition of compatibility condition, we must explicit the bridge between relations on Σ^+ and on Σ^-. We obtain this connection by rewording algebraically the topological fact that $\Gamma = \gamma_{-s}\gamma_s$ if $s \in \Sigma^+$ while $\Gamma = \gamma_s\gamma_{-s}$ if $s \in \Sigma^-$. The monomorphism

$$\begin{aligned} \sigma \;:\; \mathfrak{W} &\longrightarrow \mathfrak{W} \\ \hat{\gamma} &\longmapsto \hat{\gamma}^{-1}\widehat{\Gamma} \\ \widehat{\Gamma} &\longmapsto \widehat{\Gamma} \end{aligned}$$

with inverse

$$\begin{aligned} \sigma^{\circ -1} \;:\; \mathfrak{W} &\longrightarrow \mathfrak{W} \\ \hat{\gamma} &\longmapsto \widehat{\Gamma}\hat{\gamma}^{-1} \\ \widehat{\Gamma} &\longmapsto \widehat{\Gamma} \end{aligned}$$

satisfies

$$\mathfrak{p}_s = \mathfrak{p}_{-s} \circ \sigma^{\circ \pm 1} \qquad \text{for } s \in \Sigma^{\pm} \,.$$

The system (22.43) expresses that the necklace holonomy associated with an actual unfolding is compatible with the latter identity. Yet the system explicitly involves the sectorial first integrals $H_\bullet^{\mathrm{ns}}$ and is therefore not intrinsic to the orbital necklace. The key to resolve this issue is to observe that for $s \in \Sigma^+$ the mapping

$$\Delta_{-s}\left(\mathrm{g}^{-1}\mathrm{s}\right) \in \mathrm{Diff}\,(\mathbb{C}, 0)$$

is hyperbolic and therefore locally analytically linearizable near the fixed-point 0. There exists only one such analytic linearization with prescribed linear part. Because $\Delta_s\left(\mathrm{n}^{-1}\mathrm{g}\right)$ is an affine map the invertible function

$$\eta_s := H_s \circ H_{-s}^{\circ -1} \tag{22.44}$$

is a holomorphic linearization of $\Delta_{-s}\left(\mathrm{g}^{-1}\mathrm{s}\right)$. Hence η_s can be recovered uniquely, up to its linear part, from the knowledge of the orbital necklace.

Definition 16. For an orbital necklace $\Omega\,(\varphi_\bullet)$ recall the symbolic holonomy representation $\Delta_\bullet$. We say that the orbital necklace $\Omega\,(\varphi_\bullet)$ is a **compatible orbital necklace** when there exists $\eta_\bullet \in \mathrm{Holo_c}\left(\Sigma^\cap \times (\mathbb{C}, 0)\right)$ such that for every $s \in \Sigma^+$ (*resp.* $s \in \Sigma^-$) the mapping η_s is a local linearization of $\Delta_{-s}\left(\mathrm{g}^{-1}\mathrm{s}\right)$ (*resp.* $\Delta_{-s}\left(\mathrm{n}^{-1}\mathrm{s}\mathrm{g}^{-1}\mathrm{n}\right)$ satisfying the next properties.

- For every $s \in \Sigma^\cap$

$$\eta_s'\,(0) = 1 \,.$$

- For every $s \in \Sigma^\cap$

$$\eta_s \circ \eta_{-s} = \mathrm{Id} \,.$$

- For every $s \in \Sigma^{\cap}$

$$\eta_s^* \Delta_s \left(\mathrm{n}^{-1}\mathrm{s}\right) = \Delta_{-s} \left(\mathrm{n}^{-1}\mathrm{s}\right) \ ,$$

which is equivalent to the conjugacy of the whole dynamics: $\eta_s^* \Delta_s = \Delta_{-s} \circ \sigma^{\circ \pm 1}$ for all $s \in \Sigma^{\pm}$.

22.12.4 Characterization of Even Purely Convergent Unfoldings

For the sake of concision we only deal with the case $\mu_\bullet = 0$.

Proposition 6. *Take a purely convergent generic unfolding of codimension* 1 *with* $\mu = 0$*, i.e. its orbital necklace* $\varphi_\bullet$ *satisfies* $\varphi_\bullet^{\mathrm{n}} = 0$*. There exist* $p \in \mathbb{N}$ *and* $\alpha_\bullet \in \mathbb{C}\{\lambda\}$ *such that*

$$\varphi_\bullet^{\mathrm{s}}(h) = -\frac{1}{p} \log\left(1 + \alpha_\bullet h^p\right)$$

if and only if $\varphi_\bullet^{\mathrm{s}} = \varphi_{-\bullet}^{\mathrm{s}}$.

Remark 15. A generalization of this result for general $\mu_\bullet$ is performed in [29]. In that case the existence of a non-zero, even $\varphi_\bullet^{\mathrm{s}}$ forces $\mu_\bullet$ to be a rational constant belonging to $p^{-1}\mathbb{Z}$.

Because the local classification for saddle-node vector fields is complete, and since $\varphi_\bullet$ extends continuously at $s = 0$, the configuration presented in the proposition is rather rare. As a consequence the classification presented in Theorem 12 is not complete.

We also mention that such unfoldings are locally orbitally conjugate to a normal form (Theorem 5)

$$\mathrm{X}_\bullet^{\infty} + \lambda^\kappa x^{p+2} y^{p+1} \frac{\partial}{\partial y} \quad , \ \kappa \in \overline{\mathbb{N}} \, .$$

As in [33] it is indeed possible to show that for the above normal form one has

$$\varphi_\bullet^{\mathrm{s}}(h) = -\frac{1}{p} \log\left(1 - 2\mathrm{i}\pi\lambda^\kappa \mathfrak{T}_\bullet\left(xy^p\right)(h)\right) \ ,$$

where the period operator is the one associated with the model $\mathrm{X}_\bullet^{\infty}$. The explicit computations done in Corollary 6 therefore prove our claim, as well as one direction of the proposition since the period $\mathfrak{T}_\bullet\left(xy^p\right)$ is actually even.

Let us prove the other direction. If $\varphi_\bullet^{\mathrm{s}}$ is even, then the orbital compatibility condition writes

$$\eta_s^* \psi_s^{\mathrm{s}} = \psi_s^{\mathrm{s}}$$

for some η_s tangent-to-identity. Therefore $\langle \eta_s, \psi_s^{\mathrm{s}} \rangle < \mathrm{Diff}(\mathbb{C}, 0)$ is Abelian. Since η_s and ψ_s^{s} are tangent-to-identity there consequently exists [18] a formal tangent-to-identity change ϕ_s in the variable h, unique $d \in \mathbb{N}$, $\nu \in \mathbb{C}$ and $t \in \mathbb{C} \setminus \{0\}$ such that

$$\begin{aligned} \hat{\eta}_s &:= \phi_s^* \eta_s = \Phi^1_{Z(d,\nu)} \\ \hat{\psi}_s^{\mathrm{s}} &:= \phi_s^* \psi_s^{\mathrm{s}} = \Phi^t_{Z(d,\nu)} \\ Z(p,\nu) &:= \frac{h^{p+1}}{1+\nu h^p} \frac{\partial}{\partial h} . \end{aligned}$$

Observe that for all $t \in \mathbb{C}$ the diffeomorphism

$$\Phi^t_{Z(p,0)}(h) = \frac{h}{(1-pth^p)^{1/p}}$$

is a ramification of homography, therefore we aim at showing $\nu = 0$. This is ultimately done by applying the next lemma.

Lemma 13. *[5, Assertions 1.1–1.4]In the following ξ is a formal diffeomorphism in the variable h at 0.*

1. *Let Z, $\widetilde{Z}$ be formal vector fields in the variable h at 0. If $\xi^* \Phi^1_Z = \Phi^1_{\widetilde{Z}}$, then $\xi^* Z = \widetilde{Z}$ (the converse is trivial).*
2. *Assume that $\xi^* Z(p,\nu) = aZ(p,\nu)$ with $a \neq 1$. Then $\nu = 0$ and ξ is a ramification of homography (in particular analytic).*

Show now that $\nu = 0$ and ϕ_s itself is a ramification of homography, forcing ψ_s^{s} to be also one. Taking into account the fact that η_s linearizes $\Delta_{-s}\left(\mathrm{g}^{-1}\mathrm{s}\right) = \ell_s \psi_s^{\mathrm{s}}$ for $\ell_s := \exp \frac{\mathrm{i}\pi}{s}$, we obtain the relation

$$L_s^* \hat{\eta}_s = \left(\hat{\psi}_s^{\mathrm{s}}\right)^{\circ -1} \circ \hat{\eta}_s = \Phi^{1-t}_{Z(p,\nu)}$$

where

$$L_s := \phi_s^{\circ -1} \circ (\ell_s \phi_s) .$$

We apply the lemma with $\xi := L_s$ and $a := 1 - t \neq 1$. Hence $\nu = 0$ and L_s is a ramification of homography.

ϕ_s is a formal linearization of L_s which is tangent-to-identity. For values λ of the parameter corresponding to $\ell_s \notin \mathbb{R}$ (say $\Im(\ell_s) > 0$) the fix-point 0 of L_s is hyperbolic: the map ϕ_s is locally holomorphic at 0, unique and therefore given by

$$\phi_s := \lim_{n \to \infty} \frac{L_s^{\circ n}}{\ell_s^{-n}}$$

uniformly on a neighborhood of 0. Since for every $n \in \mathbb{N}$ the map $\frac{L_s^{\circ n}}{\ell_s^{-n}}$ is a ramification of homography the result folds by taking the limit $n \to \infty$.

Remark 16. In the coordinate induced by ϕ_s the subgroup $\langle \eta_s, \psi_s^{\mathrm{s}} \rangle$ is given by

$$\widehat{G} := \left\langle \frac{h}{(1+\xi_\bullet h^p)^{\frac{1}{p}}} \, , \, \frac{h}{(1+\alpha_\bullet h^p)^{\frac{1}{p}}} \right\rangle =: \left\langle \hat{\eta}_\bullet, \hat{\psi}_\bullet^{\mathrm{s}} \right\rangle$$

for $\alpha_\bullet, \ \xi_\bullet \in \mathbb{C}\{\lambda\}$. Hence L_s must be of the form

$$L_s(h) = \frac{\ell_\bullet h}{(1+\delta_s h^p)^{\frac{1}{p}}} \, ,$$

and we deduce

$$\xi_\bullet = \frac{\alpha_\bullet}{1-\ell_\bullet^p} \, . \tag{22.45}$$

In this specific configuration we observe explicitly that the transition mapping η_s cannot be defined for all values of $s \in \Sigma$ save when $\alpha_\bullet = 0$, because of (22.45).

22.13 Instances of Complete Classifications

22.13.1 Complete Temporal Classification

We first describe the range of the period operator (Proposition 5). The next theorem is showed in Sect. 22.13.1.1.

Theorem 13. *For a generic unfolding $X_\bullet$ of codimension 1 we recall its orbital necklace $\mathrm{Orb}(X_\bullet) = \Omega(\varphi_\bullet)$ and associated transition map $\eta_\bullet$ as in (22.44).*

1. *Let $f_\bullet \in \mathrm{Hol}_{\mathrm{c}}(\mathrm{Orb}(X_\bullet), 0)$ such that $f_\bullet(0) = 0$. There exists $G_\bullet \in \mathrm{Hol}_{\mathrm{c}}\left(\Sigma \times \left(\mathbb{C}^2, 0\right)\right)$ with $G_\bullet = \mathrm{O}(P_\bullet) + \mathrm{O}(y)$ such that $\mathfrak{T}_\bullet(G_\bullet) = f_\bullet$.*
2. *We can find such a function $G_\bullet$ satisfying $G_\bullet = G_{-\bullet}$ on $\Sigma^\cap$ if, and only if*

$$\left(\forall s \in \Sigma^+\right) \qquad \sum_{n=0}^{\infty} f_s \circ \Delta_s\left(\left(\mathrm{g}^{-1}\mathrm{s}\right)^n\right) \circ \eta_s = \sum_{n=0}^{\infty} f_{-s} \circ \Delta_{-s}\left(\left(\mathrm{s}^{-1}\mathrm{g}\right)^n\right) \, . \tag{22.46}$$

Notice that because $\Delta_s\left(\mathrm{g}^{-1}\mathrm{s}\right)$ and $\Delta_{-s}\left(\mathrm{s}^{-1}\mathrm{g}\right)$ are tangent to $\exp \mathrm{i}\pi\left(-\frac{1}{s} + \mu_\lambda\right)\mathrm{Id}$ and $\exp \mathrm{i}\pi\left(-\frac{1}{s} - \mu_\lambda\right)\mathrm{Id}$ respectively, both sums converge geometrically.

Definition 17. Let $(\varphi_\bullet, f_\bullet)$ be a temporal necklace (Definition 13). We say that it is a **compatible temporal necklace** if $\Omega(\varphi_\bullet)$ is a compatible orbital necklace and if $f_\bullet$ fulfills the condition (22.46). This is equivalent to the following statement: for $s \in \Sigma^+$ let $\phi_{\pm s}$ denote the unique solution vanishing at 0 of the discrete cohomological equations

$$\begin{cases} \phi_s - \phi_s \circ \Delta_s\left(\mathrm{g}^{-1}\mathrm{s}\right) & = f_s \\ \phi_{-s} - \phi_{-s} \circ \Delta_{-s}\left(\mathrm{s}^{-1}\mathrm{g}\right) & = f_{-s} \end{cases},$$

then

$$\eta_s^* \phi_s = \phi_{-s}\,.$$

Remark 17. It is sufficient to ensure the relation holds on Σ^+ because the orbital necklace is compatible. We refer to Remark 18 for more details.

Corollary 5. *Being given $X_\bullet$ define*

$$\mathrm{Fol}\left(X_\bullet\right) := \left\{ U_\bullet X_\bullet \ : \ U_\bullet \in \mathrm{Holo}\left(\mathbb{C}^3, 0\right)^\times \right\}$$

and

$$\mathrm{Compat}\left(X_\bullet\right) := \left\{ f_\bullet \in \mathrm{Holo}_{\mathrm{c}}\left(\Sigma \times (\mathbb{C}, 0)\right) \ : \ f_\bullet \textit{ satisfies (22.46)} \right\}\,.$$

1. *We have a complete classification*

$$\mathrm{Mod}_{\mathrm{loc}}\left(\mathrm{Fol}\left(X_\bullet\right)\right) \simeq {}^{\mathrm{Compat}\left(X_\bullet\right)}\!/_{\mathbb{C}\{\lambda\}^\times}\,.$$

2. *If $X_\bullet = \mathrm{X}_\bullet^\infty$ we have normal forms*

$$\mathrm{NF}_{\mathrm{loc}}\left(\mathrm{Fol}\left(\mathrm{X}_\bullet^\infty\right)\right) := \left\{ \frac{\mathrm{u}_\bullet}{1 + \mathrm{u}_\bullet Q_\bullet} \ : \ \mathrm{u}_\bullet \in \mathbb{C}\{\lambda\}[x]^\times_{\leq 1}\,,\ Q_\bullet \in \mathrm{Section}\left(1, \mathrm{fi}\right) \right\} \mathrm{X}_\bullet^\infty$$

where

$$\tau := \begin{cases} 0 & \textit{if } \mu_0 \notin \mathbb{R}_{\leq 0} \\ 1 + \left\lfloor \frac{-\mu_0}{k+1} \right\rfloor & \textit{otherwise} \end{cases}$$

$$\mathrm{Section}\left(1, \tau\right) := x P_\lambda^\tau y \mathbb{C}\left\{\lambda, P_\lambda^\tau y\right\}\,.$$

Item (1) is a direct restatement of Theorem 13, Theorem 12 and Definition 13 Item (3). We give the proof of Item (2) in Sect. 22.13.1.2, based on the explicit computation of the period operator for the formal model in terms of the Gamma

function. The dominant terms are $\Gamma\left(1+m\left(\tau\left(k+1\right)+\mu_\bullet\right)\right)$ for $m \in \mathbb{Z}_{\geq 0}$: the presence of the monomial x^τ helps keeping far away from poles for small $|s|$.

22.13.1.1 Range of the Period Operator (Proof of Theorem 13)

The proof relies on solving two Cousin problems, one in (x, y)-space and the other one in s-space.

Lemma 14. *Recall the total sectorial spaces* $\mathcal{D}^\sharp$ *as in* (22.35). *Given* $\delta_\bullet \in \mathrm{Holo_c}\left(\mathrm{Orb}\left(X_\bullet\right), 0\right)$ *such that* $s \mapsto \frac{\delta_s}{s}$ *is bounded, there exist two functions* $F_\bullet^\sharp \in \mathrm{Holo_c}\left(\mathcal{D}^\sharp\right)$ *such that*

$$F_s^{\mathrm{sn}} - F_s^{\mathrm{ns}} = \begin{cases} \delta_s \circ H_s^{\mathrm{ns}} & \text{on } \mathcal{V}_s^{\mathrm{s}} \\ 0 & \text{on } \mathcal{V}_s^{\mathrm{g}} \cup \mathcal{V}_s^{\mathrm{n}} \end{cases}.$$

Proof. This is a slight variation on the classical Cauchy–Heine transform. See Fig. 22.20a. For $(x, y) \in \mathcal{V}_s^\sharp$ we set

$$F_s^\sharp(x, y) := \frac{1}{2\mathrm{i}\pi} \int_{\Gamma_s^\sharp} \frac{\delta_s \circ H_s^{\mathrm{ns}}(z, y)}{z - x} \mathrm{d}z,$$

which by hypothesis defines an element of $\mathrm{Holo_c}\left(\mathcal{D}^\sharp\right)$ if $\Gamma_s^\sharp$ is deformed slightly to lie outside $\mathrm{adh}\left(V_s^\sharp\right)$. The rest follows from Cauchy's formula. □

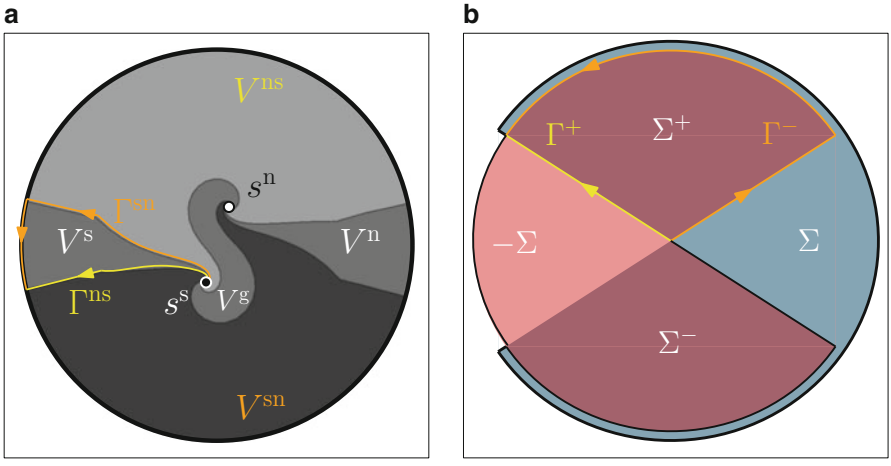


Fig. 22.20 The contour used for the Cauchy–Heine transforms. (**a**) Spatial contours. (**b**) Parametric contours

The same construction applies for the parametric Cousin problem, using the corresponding contours given in Fig. 22.20b.

Lemma 15. *Given* $d_\bullet \in \mathrm{Hol}_{\mathrm{c}}\left(\Sigma^+ \times \mathcal{U}\right)$ *such that* $s \mapsto \frac{d_s}{s}$ *is bounded, there exists a function* $D_\bullet \in \mathrm{Hol}_{\mathrm{c}}\left(\Sigma \times \mathcal{U}\right)$ *such that for all* $s \in \Sigma^+$ *we have* $D_s - D_{-s} = d_s$.

Proof. For $s \in \Sigma$ simply set

$$D_s := \frac{1}{2\mathrm{i}\pi} \int_{\Gamma+} \frac{d_z}{z-s}\mathrm{d}z + \frac{1}{2\mathrm{i}\pi} \int_{\Gamma-} \frac{d_z}{z+s}\mathrm{d}z \,.$$

□

Let us proceed now with the proof of Theorem 13. Lemma 14, applied to the temporal invariant $\delta_s := f_s$, yields two sectorial functions $F_s^\sharp$ with prescribed difference over the saddle sector. Because $F_s^{\mathrm{sn}} - F_s^{\mathrm{ns}}$ is a first integral of X_λ the function defined by

$$G_s := \begin{cases} X_\lambda \cdot F_s^{\mathrm{ns}} & \text{on } \mathcal{V}_s^{\mathrm{ns}} \\ X_\lambda \cdot F_s^{\mathrm{sn}} & \text{on } \mathcal{V}_s^{\mathrm{sn}} \end{cases}$$

is holomorphic and bounded on $\mathcal{V}_s$, therefore extends holomorphically to $\mathcal{U}$. By construction its period coincides with f_s. Because $F_s^\sharp$ has bounded derivatives [28], $G_\bullet \in \mathrm{Hol}_{\mathrm{c}}\left(\Sigma \times \mathcal{U}\right)$ is of the desired form:

$$G_s(x,y) = P_s(x)\left(\frac{\partial F_s^\sharp}{\partial x} + R_\lambda(x,y)\frac{\partial F_s^\sharp}{\partial y}\right) + y\left(1+\mu_\lambda x\right)\frac{\partial F_s^\sharp}{\partial y} \,.$$

This proves Item (1).

In general there is no reason for $G_s = G_{-s}$ to hold when $s \in \Sigma^\cap$. We must therefore modify $G_\bullet$ by adding to it a function of the form $X_\bullet \cdot D_\bullet$ with $D_\bullet \in \mathrm{Hol}_{\mathrm{c}}\left(\Sigma \times \mathcal{U}\right)$, in such a way that $\widetilde{G}_\bullet := G_\bullet + X_\bullet \cdot D_\bullet$ be even. The equation $\widetilde{G}_{-s} = \widetilde{G}_s$ reads

$$G_s - G_{-s} = X_\lambda \cdot (D_{-s} - D_s) \,.$$

According to Theorem 10 Item (4). we need the identity

$$f_s = \mathfrak{T}_s(G_s) = \mathfrak{T}_s(G_{-s})$$

to hold for $s \in \Sigma^\cap$. We prove below the hypothesis of Theorem 13 Item (2) guarantees that very property. Taking this fact for granted, we deduce the existence of $d_\bullet$ such that $G_s - G_{-s} = X_\lambda \cdot d_s$ for all $s \in \Sigma^+$ using Theorem 10 Item (4) again. This function can be so chosen that $\lim_{s\to 0} d_s = 0$, in which case $s \mapsto \frac{d_s}{s}$ is bounded. Then Lemma 15 yields the expected $D_\bullet$, completing the proof of Item (2).

Remark 18. From Theorem 10 Item (4) we know that $\mathfrak{T}_s(G_s - G_{-s}) = 0$ for all $s \in \Sigma^+$ if and only if $\mathfrak{T}_{-s}(G_s - G_{-s}) = 0$ for all $s \in \Sigma^+$. Therefore $f_s = \mathfrak{T}_s(G_{-s})$ for all $s \in \Sigma^\cap$ if and only if the equality holds merely on Σ^+.

Proposition 7. *1. For all $s \in \Sigma^+$ we have*

$$\eta_s^* \sum_{n=0}^{\infty} \mathfrak{T}_s(G_{-s}) \circ \Delta_s\left(\left(\mathrm{g}^{-1}\mathrm{s}\right)^n\right) = \sum_{n=0}^{\infty} f_{-s} \circ \Delta_{-s}\left(\left(\mathrm{s}^{-1}\mathrm{g}\right)^n\right) .$$

2. $f_\bullet = \mathfrak{T}_\bullet(G_{-\bullet})$ if and only if condition (22.46) holds.

Proof. Set $\phi_s := \sum_{n=0}^{\infty} f_s \circ \Delta_s\left(\left(\mathrm{g}^{-1}\mathrm{s}\right)^n\right)$, $\phi_{-s} := \sum_{n=0}^{\infty} f_{-s} \circ \Delta_{-s}\left(\left(\mathrm{s}^{-1}\mathrm{g}\right)^n\right)$ and $\tilde{\phi}_s := \sum_{n=0}^{\infty} \mathfrak{T}_s(G_{-s}) \circ \Delta_s\left(\left(\mathrm{g}^{-1}\mathrm{s}\right)^n\right)$.

1. Take $p_* \in T$. Any asymptotic cycle $\gamma(p_*)$, used to compute the period in Proposition 5, is tangentially asymptotically homotopic to the lift in $\mathcal{F}_\lambda$ of the limit of nested cycles $\lim_{m\to\infty} \gamma_m$

$$\gamma_m := \mathfrak{p}_{-s}\left(\left(\mathrm{s}^{-1}\mathrm{g}\right)^{-m}\left(\mathrm{n}^{-1}\mathrm{s}\right)\left(\mathrm{s}^{-1}\mathrm{g}\right)^m\right) : [-m, m] \longrightarrow \mathbb{D}_\lambda \qquad (22.47)$$

$$\gamma_{m+1}|_{[-m,m]} = \gamma_m .$$

We let $\tilde{\gamma}_m$ be the lift of γ_m in $\mathcal{F}_\lambda$ with $\tilde{\gamma}_m(0) = p_*$. The quantity $F_{-s}^{\mathrm{ns}}(\tilde{\gamma}_m(-m)) + \int_{\tilde{\gamma}_m} G_{-s}\frac{\mathrm{d}x}{P_\lambda}$ represents the analytic continuation of F_{-s}^{ns} along $\tilde{\gamma}_m$. By construction the additive monodromy of the continuation of F_{-s}^{ns} is given by the period f_{-s} when turning around the saddle-like singularity of $\mathcal{V}_{-s}^{\mathrm{ns}}$. Hence

$$\int_{\tilde{\gamma}_m} G_{-s}\frac{\mathrm{d}x}{P_\lambda} + F_{-s}^{\mathrm{ns}}(\tilde{\gamma}_m(-m)) - F_{-s}^{\mathrm{ns}}(\tilde{\gamma}_m(m)) =$$

$$\left(\sum_{n=0}^{m} f_{-s} \circ \Delta_{-s}\left(\left(\mathrm{s}^{-1}\mathrm{g}\right)^{-n}\right) - \sum_{n=1}^{m} f_{-s} \circ \Delta_{-s}\left(\left(\mathrm{s}^{-1}\mathrm{g}\right)^{-n}\mathrm{n}^{-1}\mathrm{s}\right)\right) \circ H_{-s}^{\mathrm{ns}}(p_*) .$$

Because F_{-s}^{ns} extends continuously to $\{x = s\}$ and $\lim_{n\to\infty} \tilde{\gamma}_n(\pm n) = (s, 0)$, taking the limit we obtain in the chart $\Omega_{-s}^{\mathrm{ns}}$

$$\mathfrak{T}_s(G_{-s}) \circ \eta_s = \phi_{-s} - \phi_{-s} \circ \Delta_{-s}\left(\mathrm{s}^{-1}\mathrm{g}\mathrm{n}^{-1}\mathrm{s}\right) .$$

According to Definition 16, for $s \in \Sigma^+$ the identity $\eta_s^* \Delta_s\left(\mathrm{g}^{-1}\mathrm{s}\right) = \Delta_{-s}\left(\mathrm{s}^{-1}\mathrm{g}\mathrm{n}^{-1}\mathrm{s}\right)$ holds, so that summing over all terms $\eta_s^*\left(\mathfrak{T}_s(G_{-s}) \circ \Delta_s\left(\left(\mathrm{g}^{-1}\mathrm{s}\right)^n\right)\right)$ for $n \in \mathbb{Z}_{\geq 0}$ yields the expected result:

$$\eta_s^* \tilde{\phi}_s = \phi_{-s} .$$

2. The direct implication is trivial. Assume conversely that $\eta_s^* \phi_s = \sum_{n=0}^{\infty} f_{-s} \circ \Delta_{-s} \left(\left(\mathrm{s}^{-1} \mathrm{g} \right)^n \right)$, i.e. $\phi_s = \tilde{\phi}_s$. Because $\phi_s - \phi_s \circ \Delta_s \left(\mathrm{g}^{-1} \mathrm{s} \right) = f_s$ and $\phi_s - \tilde{\phi}_s \circ \Delta_s \left(\mathrm{g}^{-1} \mathrm{s} \right) = \mathfrak{T}_s \left(G_{-s} \right)$ we recover $f_s = \mathfrak{T}_s \left(G_{-s} \right)$.

□

22.13.1.2 Computation of the Period of the Model $\mathrm{X}_\bullet^\infty$

Below we compute periods explicitly, extending the calculation of Sect. 22.4.3.

Proposition 8. *Let $G_{n,m}(x,y) := x^n y^m$, $m \in \mathbb{N}$, and $X_\bullet := \mathrm{X}_\bullet^\infty$. Then for all $s \in \Sigma$ we have (in the chart Ω_s^{ns})*

$$\mathfrak{T}_s \left(G_{n,m} \right)(h) = h^m \times \frac{(-m)^{n+m\mu_\lambda}}{\Gamma\left(n + m\mu_\lambda\right)} \times t_{\lambda,n,m} \times T_{s,m}$$

$$t_{\lambda,n,m} := \frac{1}{2^n} \sum_{p+q=n} \binom{n}{p} \prod_{j=0}^{p-1} \left(1 - s \left(\mu_\lambda + \frac{2j}{m} \right) \right) \prod_{j=0}^{q-1} \left(1 + s \left(\mu_\lambda + \frac{2j}{m} \right) \right)$$

$$T_{s,m} := \frac{\left(-\frac{2s}{m} \right)^{m\mu_\lambda}}{1 + s\mu_\lambda} \times \frac{\Gamma\left(-\frac{m}{2s} + \frac{m\mu_\lambda}{2} \right)}{\Gamma\left(-\frac{m}{2s} - \frac{m\mu_\lambda}{2} \right)}.$$

For given s small enough, the period is zero if and only if $n + m\mu_\lambda \in \mathbb{Z}_{\leq 0}$. The period is a holomorphic function of λ if and only if $m\mu_\bullet \in \mathbb{Z}$ (in which case μ is a rational constant).

Remark 19. 1. Notice that taking the limit $s \to 0$ in Σ leads to

$$\lim_{s \to 0} T_{s,m} = \lim_{\lambda \to 0} t_{\lambda,n,m} = 1$$

recovering classical computations [8, 32, 33] performed for $\lambda = 0$.

2. The eventual lack of evenness of the period comes from the term $T_{s,m}$. Since it is independent on n, any period $\mathfrak{T}_\bullet \left(y^m g(x) \right)$, $g \in \mathbb{C}\{\lambda, x\}$, is even in s if and only if $m\mu \in \mathbb{Z}$.

Proof. We perform the computation over $V_s^{\mathrm{nsn}} = \mathbb{D}_\lambda \setminus \left(V_s^{\mathrm{g}} \cup V_s^{\mathrm{n}} \right)$. Because

$$H_s(x,y) := y (x-s)^{-\frac{1}{2s} - \frac{\mu_\lambda}{2}} (x+s)^{\frac{1}{2s} - \frac{\mu_\lambda}{2}}$$

is constant on the leaves of $\mathcal{F}_\lambda$ we can parameterize an asymptotic path as

$$x \in \gamma_\infty \longmapsto \left(x, h (x-s)^{\frac{1}{2s} + \frac{\mu_\lambda}{2}} (x+s)^{-\frac{1}{2s} + \frac{\mu_\lambda}{2}} \right)$$

where $h = H_s^{\mathrm{ns}}(p_*)$ and $\gamma_\infty = \lim_{p\to\infty}\gamma_p$ [as in (22.47)] is the projection on $\{y=0\}$ of the asymptotic cycle $\gamma(p_*)$. This projection does not depend on the choice of $p_* \in T$.

Remembering the computations performed in Proposition 3, we introduce the Pochhammer contour $\mathcal{P}_s \in \pi_1(\mathbb{D}_\lambda, x_*)$ whose encoding in the dynamics necklace is given by

$$\mathfrak{w}_s(\mathcal{P}_s) = \mathrm{n}^{-1}\mathrm{g}\mathrm{s}^{-1}\mathrm{n}\mathrm{g}^{-1}\mathrm{s}\,.$$

For $h \in \left(\Omega_s^{\mathrm{ns}}, 0\right)$ let $\tilde{\mathcal{P}}_s(h)$ be the lift in $\mathcal{F}_\lambda$ of $\mathcal{P}_s$ starting from $p_* \in T$ with $h = H_s(p_*)$. Both necklace holonomies $\Delta_s(\hat{\gamma})$ and $\Delta_s\left(\widehat{\Gamma}\right)$ are linear in the same coordinate $H_s|_T$, and therefore commute. Because the Pochhammer contour is a commutator we have

$$\Delta_s\left(\mathrm{n}^{-1}\mathrm{g}\mathrm{s}^{-1}\mathrm{n}\mathrm{g}^{-1}\mathrm{s}\right) = \mathrm{Id}\,.$$

Hence $\tilde{\mathcal{P}}_s(h)$ is a (non-trivial) element of the fundamental group of the corresponding leaf of $\mathcal{F}_\lambda$. As a matter of consequence

$$\oint_{\tilde{\mathcal{P}}_s(h)} G_{n,m}\frac{\mathrm{d}x}{P_\lambda} = \mathfrak{T}_s(G_{n,m}) - \mathfrak{T}_s(G_{n,m})\circ\Delta_s\left(\mathrm{g}^{-1}\mathrm{s}\right)\,.$$

Summing over the forward orbit of $\Delta_s\left(\mathrm{g}^{-1}\mathrm{s}\right)$ we obtain

$$\begin{aligned}
\mathfrak{T}_s(G_{n,m})(h) &= \sum_{\ell=0}^{\infty}\oint_{\tilde{\mathcal{P}}_s\circ\Delta_s\left(\left(\mathrm{g}^{-1}\mathrm{s}\right)^\ell\right)(h)} G_{n,m}\frac{\mathrm{d}x}{P_\lambda}\\
&= \sum_{\ell=0}^{\infty}\left(\frac{\Delta_s\left(\left(\mathrm{g}^{-1}\mathrm{s}\right)^\ell\right)(h)}{h}\right)^m\oint_{\tilde{\mathcal{P}}_s(h)} G_{n,m}\frac{\mathrm{d}x}{P_\lambda}\\
&= \frac{1}{1-\exp \mathrm{i}m\pi\left(-\frac{1}{s}+\mu_\lambda\right)}\oint_{\tilde{\mathcal{P}}_s(h)} G_{n,m}\frac{\mathrm{d}x}{P_\lambda}
\end{aligned}$$

because

$$\oint_{\tilde{\mathcal{P}}_s(h)} G_{n,m}\frac{\mathrm{d}x}{P_\lambda} = h^m\oint_{\mathcal{P}_s} x^n(x-s)^{\frac{m}{2s}+\frac{m\mu_\lambda}{2}-1}(x+s)^{-\frac{m}{2s}+\frac{m\mu_\lambda}{2}-1}\,\mathrm{d}x\,.$$

We recognize the integral representation of the Beta function. Up to the presence of μ_λ and m, the remaining computations are done identically to those in Proposition 3.

The function is even if and only if $T_{s,m} = T_{-s,m}$. Using the reflection formula we find

$$\frac{T_{s,m}}{T_{-s,m}} = (-1)^{m\mu_\lambda} \frac{\sin\left(\pi\frac{m}{2s} + \pi\frac{m\mu_\lambda}{2}\right)}{\sin\left(\pi\frac{m}{2s} - \pi\frac{m\mu_\lambda}{2}\right)} = \frac{\exp\left(\mathrm{i}\pi m\left(\frac{1}{2s} + \mu_\lambda\right)\right) - \exp\left(-\frac{\mathrm{i}\pi m}{2s}\right)}{\exp\left(\mathrm{i}\pi m\left(\frac{1}{2s} - \mu_\lambda\right)\right) - \exp\left(-\frac{\mathrm{i}\pi m}{2s}\right)} .$$

This quantity equals 1 if and only if $m\mu_\lambda \in \mathbb{Z}$. □

Remark 20. The fact that the complete invariant $\varphi^n_\bullet$ introduced in Sect. 22.4.3 for affine unfoldings is obtained from the above proposition for $\mu_\bullet := 0$ and $m := -1$ is not fortuitous and can be explained very much like in [33]. The best heuristics is the relation

$$X_\lambda = \mathrm{X}^\infty_\lambda - a_\lambda P_\lambda \frac{\partial}{\partial y} = \mathrm{X}^\infty_\lambda - \frac{a_\lambda P_\lambda}{y} \times y\frac{\partial}{\partial y}$$

so that locally conjugating $\mathrm{X}^\infty_\bullet$ to $X_\bullet$ is somehow equivalent to solving analytically

$$\mathrm{X}^\infty_\lambda \cdot F_\lambda = \frac{a_\lambda P_\lambda}{y} .$$

A more precise approach would require to study the generic saddle-node unfolding near $(0, 0, \infty)$ whose node- and saddle-parts correspond to saddle- and node-parts near $(0, 0, 0)$.

We deduce from this proposition the following result, concluding the proof of Corollary 5.

Corollary 6. *Recall the notations of Corollary 5. The operator*

$$\begin{aligned} \mathrm{Section}\,(1,\tau) &\longrightarrow \{f_\bullet \in \mathrm{Holo}_{\mathrm{c}}\,(\Sigma \times \mathbb{C}, 0) \ : \ f_\bullet\,(0) = 0\} \\ Q_\bullet &\longmapsto \mathfrak{T}_\bullet\,(Q_\bullet) \end{aligned}$$

is bijective.

In the next paragraph we generalize this result for all purely convergent unfoldings.

Proof. This is a consequence of the following two facts:

- the period operator sends $y^m x^n$ to some monomial $h^m T_{s,n,m}$,
- $T_{s,n,m} = 0$ for $s \in (\Sigma, 0)$ if and only if $T_{0,n,m} = 0$, that is $n + m\mu_\lambda \in \mathbb{Z}_{\leq 0}$.

The choice of τ prevents $T_{s,m\tau(k+1),m}$ to vanish so that the operator is formally invertible. The fact that the inverse operator maps convergent power series $f_\bullet$ to convergent power series is a consequence of $\lim_{s\to 0} T_{s,1+\tau(k+1)m,m} = T_{0,1+\lambda m,m} \neq 0$ and of estimates established in [33] for $T_{0,1+m\lambda,m}$. □

22.13.2 Normal Forms for Pure Convergence

Here we assume that $Z_\bullet$ is a generic unfolding of codimension 1 whose orbital invariant has null node-part:

$$\varphi^{\mathrm{n}}_\bullet = 0\,.$$

We say that the unfolding is **purely convergent**. For $\tau \in \mathbb{Z}_{\geq 0}$ define

$$\text{Convergent}\,(1,\tau) := \{Z_\bullet \;:\; \varphi^{\mathrm{n}}_\bullet = 0\,,\ \mu_0 \notin \mathbb{R}_{\leq \tau}\}$$

and let

$$\text{Convergent}\,(1) := \bigcup_{\tau \geq 0} \text{Convergent}\,(1,\tau)$$

the set of all convergent unfoldings.

Theorem 14. *Recall the definition of* $\text{Section}\,(1,\tau)$ *given in Corollary 5 Item (2). The collection*

$$\mathrm{NF}_{\mathrm{loc}}\,(\text{Convergent}\,(1,\mathrm{fi})) := \left\{ \frac{\mathrm{u}_\bullet}{1+\mathrm{u}_\bullet Q_\bullet} \left(\mathrm{X}^\infty_\bullet + y R_\bullet \frac{\partial}{\partial y} \right) \;:\; Q_\bullet,\ R_\bullet \in \text{Section}\,(1,\tau) \right\}$$

is a family of normal forms for $\text{Convergent}\,(1,\tau)$. *Two vector fields in normal forms are locally analytically conjugate if and only if there exists* $c_\bullet \in \mathbb{C}\{\lambda\}^\times$ *such that*

$$\begin{cases} \widetilde{R}_\lambda\,(x,y) & = R_\lambda\,(x, c_\lambda y) \\ Q_\lambda\,(x,y) & = Q_\lambda\,(x, c_\lambda y) \end{cases}.$$

Remark 21.

1. Notice that the normal forms are *not* in prepared form (22.26).
2. This collection of normal forms depends analytically on the parameter, hence we have a presentation

$$\mathrm{Mod}_{\mathrm{loc}}\,(\text{Convergent}\,(1)) \simeq \text{Section}\,(1,\tau) \times \text{Section}\,(1,\tau)\,/_{\mathbb{C}\{\lambda\}^\times}$$

where the invariants (the functions $Q_\bullet$ and $R_\bullet$) depend analytically on the parameter. This is very much the opposite case of Theorem 12 (compare with Sect. 22.12.4).

We show this result in two steps, following the strategy presented in [29, 30] for $\lambda = 0$. For the sake of clarity we only deal here with the case $\Re\,(\mu_0) > 0$, particularly implying $\tau = 0$.

The initial data is a formal class $(\mu_\bullet, \mathfrak{u}_\bullet)$ and a compatible temporal necklace $(\varphi_\bullet, f_\bullet)$, as in Definition 17, with $\varphi_\bullet^{\mathrm{n}} = 0$.

Proposition 9. *One can find $r' > 0$ and a covering of $(\mathbb{C}\setminus\{\pm s\}) \times r'\mathbb{D}$ into two modified, infinite canonical sectors $\mathcal{V}_s^\sharp$, $\sharp \in \{\mathrm{ns}, \mathrm{sn}\}$, such that the following properties hold.*

1. *There exists a unique collection of holomorphic vector fields $X_\bullet \in \mathfrak{X}(\Sigma \times \mathbb{C} \times r'\mathbb{D})$ of the form*

$$X_s = \mathrm{X}_\lambda^\infty + yxR_s\frac{\partial}{\partial y} \quad , R_\bullet \in \mathrm{Holo}_{\mathrm{c}}\left(\Sigma \times r'\mathbb{D}\right) , \tag{22.48}$$

   and such that the associated canonical first integrals $H_\bullet^\sharp \in \mathrm{Holo}\left(\mathcal{D}^\sharp\right)$ have connected fibers and satisfy $H_s^{\mathrm{sn}} = \varphi_s^\flat \circ H_s^{\mathrm{ns}}$ for $\flat \in \{\mathrm{n}, \mathrm{g}, \mathrm{s}\}$ and $s \in \Sigma$.
2. *The action of holomorphic automorphisms of the orbital necklace $\Omega(\varphi_\bullet)$ by $h \mapsto c_\bullet h$ induces an action*

$$c_\bullet^* \ : \ R_\bullet \longmapsto R_\bullet \circ (c_\bullet \mathrm{Id}) \ .$$

   In other words, two vector fields as above are locally orbitally conjugate for all $\lambda \in \Lambda$ if and only if $\widetilde{R}_\bullet(y) = R_\bullet(c_\bullet y)$ for some $c_\bullet \in \mathbb{C}\{\lambda\}^\times$.

It might happen that $R_\bullet \neq R_{-\bullet}$ in Item (1), preventing $X_\bullet$ to be a generic unfolding. The compatibility condition precisely guarantees that $R_\bullet = R_{-\bullet}$ so that $\mathrm{X}_\bullet^\infty + yxR_\bullet$ is the expected normal form in that case. The following lemma hence completes orbital realization with normal forms.

Lemma 16. *Let $X_\bullet$ be a collection of vector fields in the form (22.48), with associated orbital necklace $\Omega(\varphi_\bullet)$. Then $R_\bullet = R_{-\bullet}$ if and only if $\Omega(\varphi_\bullet)$ is compatible.*

The temporal realization is a straightforward consequence of the next concluding result, as was done to prove Corollary 5. This is a generalization of Corollary 6 to generic unfoldings under normal forms.

Proposition 10. *Let $X_\bullet$ be a generic unfolding of codimension 1 in normal form of Theorem (14). The operator*

$$\begin{aligned} \mathrm{Section}(1, \tau) &\longrightarrow \{f_\bullet \in \mathrm{Holo}_{\mathrm{c}}(\Sigma \times \mathbb{C}, 0) \ : \ f_\bullet(0) = 0\} \\ Q_\bullet &\longmapsto \mathfrak{T}_\bullet(Q_\bullet) \end{aligned}$$

is bijective.

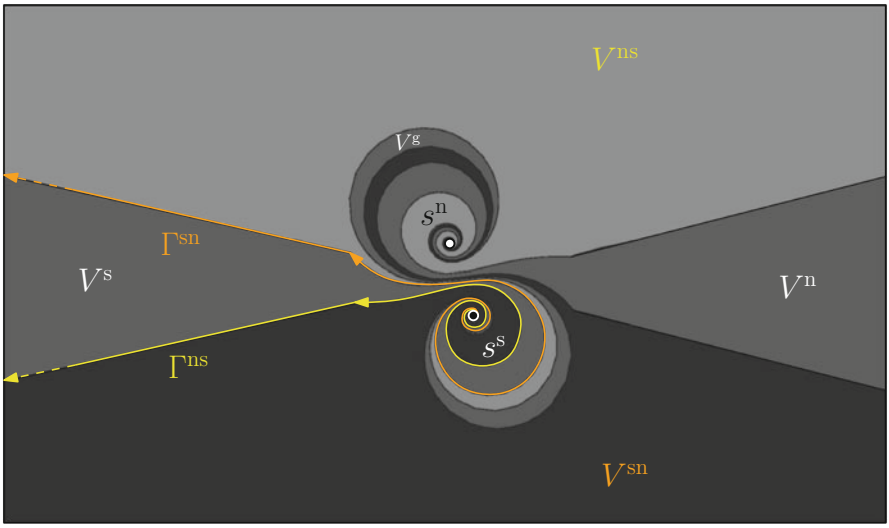


Fig. 22.21 The modified (unbounded) squid sectors and associated Cauchy–Heine contours

22.13.2.1 Orbital Realization on Σ (Proof of Proposition 9)

The proof is achieved by iterating a Cauchy–Heine integral transformation solving a certain sectorial Cousin problem, like in Lemma 14, to obtain $H^{\sharp}_{\bullet}$. In fact we seek two functions $\mathcal{O}^{\sharp}_{\bullet}$ such that

$$H^{\sharp}_{\bullet} := H^{\infty,\sharp}_{\bullet} \exp\left(\mathcal{O}^{\sharp}_{\bullet}\right)$$

as in (22.38).

Definition 18. Take $\rho, r > 0$ such that $r < \frac{1}{|\mu_\lambda|}$ for every $|s| < \rho$. We refer to Fig. 22.21.

1. The **modified squid sector** V^{ns}_s is obtained from the union of a squid sector of radius r as in Definition 10 and the half-rays $\exp \frac{-\mathrm{i}\pi}{8}\mathbb{R}_{\geq r}$ and $\exp \frac{9\mathrm{i}\pi}{8}\mathbb{R}_{\geq r}$. The construction is analogous for V^{sn}_s, as well as their saddle-, gate-, and node-parts.
2. Let $r' > 0$ be given. We call **modified canonical sector** $\mathcal{V}^{\sharp}_s$ the product $V^{\sharp}_s \times r'\mathbb{D}$.
3. We say that a triple $\left(r', \mathcal{O}^{\mathrm{sn}}_{\bullet}, \mathcal{O}^{\mathrm{ns}}_{\bullet}\right)$, with $\mathcal{O}^{\sharp}_{\bullet} \in \mathrm{Holo}_{\mathrm{c}}\left(\mathcal{D}^{\sharp}\right)$, is **adapted** to a domain $\Omega = (\mathbb{C}, 0)$ if $H^{\mathrm{ns}}_s\left(\mathcal{V}^{\mathrm{s}}_s\right) \subset \Omega$ for all $s \in \Sigma$.

The next result is the basis of the construction. We omit the proof, which is a straightforward generalization of its counterpart in [30] for $s = 0$ and can be found in [29]. Instead we focus on the constructions involved, stressing the few steps where the case $s = 0$ does not extend straightforwardly.

Theorem 15. *Consider some* $\left(r', \mathcal{O}_\bullet^{\mathrm{sn}}, \mathcal{O}_\bullet^{\mathrm{ns}}\right)$ *adapted to* Ω. *Take any* $f_\bullet \in \mathrm{Holo_c}\left(\Sigma \times \Omega\right)$ *vanishing along* $\{h = 0\}$ *and with bounded derivative* $f'_\bullet = \frac{\partial f_\bullet}{\partial h}$ *on* $\Sigma \cup \{0\}$, *then define* $F_\bullet^\sharp$ *by*

$$F_s^\sharp(x,y) := \frac{1}{2\mathrm{i}\pi} \int_{\Gamma_s^\sharp} \frac{f_s\left(H_s^\sharp(z,y)\right)}{z-x} \mathrm{d}z \ , \ s \in \Sigma \ , \ (x,y) \in \mathcal{V}_s^\sharp \ , \tag{22.49}$$

where the paths were described in Definition 18. The following properties hold.

1. $F_\bullet^\sharp \in \mathrm{Holo_c}\left(\mathcal{D}^\sharp\right)$.
2. $F_s^{\mathrm{sn}} - F_s^{\mathrm{ns}} = f_s \circ H_s^{\mathrm{ns}}$ *on* $\mathcal{V}_s^{\mathrm{s}}$ *and vanishes on* $\mathcal{V}_s^{\mathrm{n}} \cup \mathcal{V}_s^{\mathrm{g}}$.
3. *For* $|s| < \rho$ *one has estimates*

a.

$$\sup_{\mathcal{V}_s^\sharp} \left|F_s^\sharp\right| \leq r' K \sup_{\Omega} \left|f_s'\right| \exp \sup_{\mathcal{V}_s^\sharp} \left|\mathcal{O}_s^\sharp\right|$$

b.

$$\sup_{\mathcal{V}_s^\sharp} \left|y \frac{\partial F_s^\sharp}{\partial y}\right| \leq r' K \sup_{\Omega} \left|f_s'\right| \exp \sup_{\mathcal{V}_s^\sharp} \left|\mathcal{O}_s^\sharp\right| \left(1 + \sup_{\mathcal{V}_s^\sharp} \left|y \frac{\partial \mathcal{O}_s^\sharp}{\partial y}\right|\right)$$

c.

$$\sup_{\mathcal{V}_s^\sharp} \left|x \frac{\partial F_s^\sharp}{\partial x}\right| \leq r' K \sup_{\Omega} \left|f_s'\right| \exp \sup_{\mathcal{V}_s^\sharp} \left|\mathcal{O}_s^\sharp\right| \left(1 + \sup_{\mathcal{V}_s^\sharp} \left|x \frac{\partial \mathcal{O}_s^\sharp}{\partial x}\right|\right)$$

with some constant $K > 0$ *depending only on* μ_0, ρ *and* r.

Remark 22. We mention the fact that the assumption $\Re\left(\mu_\lambda\right) > 0$ guarantees the convergence of the integrals near ∞, since $f_s\left(H_s^\sharp(z,y)\right) \sim_\infty Cz^{-\mu_\lambda}$ for fixed $y \in r'\mathbb{D}$.

For $r' > 0$ and $\mathcal{O} := \left(\mathcal{O}_\bullet^{\mathrm{sn}}, \mathcal{O}_\bullet^{\mathrm{ns}}\right)$ adapted to Ω, we write $\mathcal{E}(\mathcal{O})$ the pair $\left(F_\bullet^{\mathrm{ns}}, F_\bullet^{\mathrm{sn}}\right)$ built in the previous theorem for $f_\bullet := \varphi_\bullet^{\mathrm{s}}$. Define the recursive sequence $(\mathcal{O}_n)_{n \in \in \mathbb{Z}_{\geq 0}}$ starting with $\mathcal{O}_0 := (0,0)$ and

$$\mathcal{O}_{n+1} := \mathcal{E}(\mathcal{O}_n) \ , \ n \geq 0 \ . \tag{22.50}$$

Show it converges in the Banach space $\mathcal{H} := \mathrm{Holo_c}\left(\mathcal{D}^{\mathrm{ns}}\right) \times \mathrm{Holo_c}\left(\mathcal{D}^{\mathrm{sn}}\right)$ (equipped with the sup-norm). For the sake of clarity we write H_n the canonical first integral H_s^{ns} built from $\mathcal{O}_s^{\mathrm{ns}}$. We can assume that all φ_s^{s} are holomorphic and have bounded derivatives on some disc $\eta\mathbb{D} = (\mathbb{C}, 0)$. Then we choose

$$\rho \leq \frac{\eta}{M} \exp\left(-\frac{\eta}{M} K \sup_{\eta\mathbb{D}} \left| \frac{\mathrm{d}\varphi_s^{\mathrm{s}}}{\mathrm{d}h} \right| \right)$$

where

$$M = M(\mu_\bullet) := \sup_{|s|<\rho\ ,\ z\in V_s^{\mathrm{s}}} \left| P_\lambda^{-\frac{\mu_\lambda}{2}} \mathfrak{g}_s \right| \exp 2\pi\, |\mu_\lambda|\ ,$$

and K is the constant appearing in Theorem 15. We wish to ensure that

$$(\forall s \in \Sigma)\ (\forall n \in \mathbb{N})\ \left(\forall y \in r'\mathbb{D}\right)\ \left(\forall z \in V_s^{\mathrm{s}}\right) \qquad |H_n(z,y)| \leq \eta\,. \quad (22.51)$$

By construction of H_n we have for $(z,y) \in \mathcal{V}_s^{\mathrm{ns}}$

$$|H_n(z,y)| \leq r' M \exp \sup_{\mathcal{V}_s^{\mathrm{ns}}} \left| \mathcal{O}_s^{\mathrm{ns}} \right|\ .$$

Therefore if for some $n \in \mathbb{N}$ we have $\sup_{\mathcal{V}_s^{\mathrm{ns}}} \left|\mathcal{O}_n^{\mathrm{ns}}\right| \leq \frac{2\pi}{M}\eta K \sup_{\eta\mathbb{D}} \left|\frac{\mathrm{d}\varphi_s^{\mathrm{s}}}{\mathrm{d}h}\right|$ then we first find that

$$|H_n(z,y)| < r' M \exp\left(\frac{\eta}{M} K \sup_{\eta\mathbb{D}} \left| \frac{\mathrm{d}\varphi_s^{\mathrm{s}}}{\mathrm{d}h} \right| \right) = \eta,$$

i.e., $(r', \mathcal{O}_n)$ is adapted to $\eta\mathbb{D}$ and then, using Theorem 15 Item (4a), we obtain

$$\begin{aligned} \sup_{\mathcal{V}_s^{\mathrm{ns}}} \left|\mathcal{O}_{n+1}^{\mathrm{ns}}\right| &\leq r' K \sup_{\eta\mathbb{D}} \left| \frac{\mathrm{d}\varphi_s^{\mathrm{s}}}{\mathrm{d}h} \right| \exp \sup_{\mathcal{V}_s^{\mathrm{ns}}} \left|\mathcal{O}_n^{\mathrm{ns}}\right| \\ &\leq \frac{\eta}{M} K \sup_{\eta\mathbb{D}} \left| \frac{\mathrm{d}\varphi_s^{\mathrm{s}}}{\mathrm{d}h} \right|\ . \end{aligned}$$

These estimates show by induction on n that, with the above choice of r', the relation (22.50) defines a bounded sequence $(\mathcal{O}_n)_n \subset \mathcal{H}$. It so happens that the components of the sequence $(\mathcal{O}_n)_n$ converge for the Krull topology in $\mathrm{Holo_c}\left(D^\sharp\right)[[y]]$, almost by construction, hence we can ensure it converges in the space $\mathrm{Holo_c}\left(\mathcal{D}^\sharp\right)$ by borrowing the argument of [30].

Lemma 17. *[30] Let $\mathcal{D}$ be a domain in $\mathbb{C}^m$ and consider a bounded sequence $\left(f_p\right)_{p\in\mathbb{N}}$ of $\mathrm{Holo_c}(\mathcal{D})$ satisfying the additional property that there exists some point $\mathbf{z}_0 \in \mathcal{D}$ such that the corresponding sequence of Taylor series $\left(T_p\right)_{p\in\mathbb{N}}$ at $\mathbf{z}_0$ is convergent in $\mathbb{C}[[\mathbf{z}-\mathbf{z}_0]]$ (for the projective topology). Then $\left(f_p\right)_p$ converges uniformly on compact sets of $\mathcal{D}$ towards some $f_\infty \in \mathrm{Holo_c}(\mathcal{D})$.*

The cited reference likewise provides the remaining claims of the upcoming proposition.

Proposition 11. *Let $\Omega = (\mathbb{C}, 0)$ be a domain and a collection $\varphi_\bullet^{\mathrm{s}} \in \mathrm{Holo}_{\mathrm{c}}(\Sigma \times \Omega)$ be given. Then there exists $\left(r', \mathcal{O}_\bullet^{\mathrm{sn}}, \mathcal{O}_\bullet^{\mathrm{ns}}\right) \in \mathbb{R}_{>0} \times \mathcal{H}$ adapted to Ω such that*

1.

$$H_s^{\mathrm{sn}} = H_s^{\mathrm{ns}} \exp\left(2\mathrm{i}\pi\mu_\lambda + \varphi_s^{\mathrm{s}}\left(H_s^{\mathrm{ns}}\right)\right) ,$$

2.

$$\sup_{\mathcal{V}_s^\sharp} \left| y \frac{\partial \mathcal{O}_s^\sharp}{\partial y} \right| < 1 \quad , \quad \sup_{\mathcal{V}_s^\sharp} \left| x \frac{\partial \mathcal{O}_s^\sharp}{\partial x} \right| < 1 .$$

We can now complete the proof of the first item of Proposition 9.

Corollary 7. *Let $r' > 0$ and $\mathcal{O}_\bullet^\sharp$ be given by Proposition 11.*

1. For $s \in \Sigma$ the vector field

$$X_s^\sharp := \mathrm{X}_s^\infty - y \frac{\mathrm{X}_s^\infty \cdot \mathcal{O}_s^\sharp}{1 + y \frac{\partial \mathcal{O}_s^\sharp}{\partial y}} \frac{\partial}{\partial y} \quad , \ \sharp \in \{\mathrm{ns}, \mathrm{sn}\}$$

is holomorphic on $\mathcal{V}_s^\sharp$ and admits $H_s^\sharp$ as first integral.

2. The vector fields $X_s^\sharp$ are restrictions to the sectors $\mathcal{V}_s^\sharp$ of a vector field

$$X_s(x, y) = \mathrm{X}_s^\infty(x, y) + R_s(y)\, y \frac{\partial}{\partial y}$$
$$R_\bullet \in \mathrm{Holo}_{\mathrm{c}}\left(\Sigma \times r'\mathbb{D}\right) \ , \ R_\bullet(0) = 0 .$$

Proof. Define

$$R_s^\sharp := -\frac{\mathrm{X}_s^\infty \cdot \mathcal{O}_s^\sharp}{1 + y \frac{\partial \mathcal{O}_s^\sharp}{\partial y}} . \tag{22.52}$$

1. It is a straightforward consequence of Proposition 11.
2. On the one hand, we have

$$X_s^\sharp \cdot H_s^{\mathrm{sn}} = X_s^\sharp \cdot \left(H_s^{\mathrm{sn}} \exp\left(2i\pi\mu_\lambda + \varphi_s^{\mathrm{s}}\left(H_s^{\mathrm{sn}}\right)\right)\right) = 0 .$$

On the other hand, a short calculation shows that

$$X_s^\sharp \cdot H_s^{\mathrm{sn}} = H_s^{\mathrm{sn}} \left(X_s^\infty \cdot \mathcal{O}_s^{\mathrm{sn}} + \left(1 + y \frac{\partial \mathcal{O}_s^{\mathrm{sn}}}{\partial y} \right) R_s^{\mathrm{ns}} \right) .$$

Therefore the functions $R_s^\sharp$ glue other the intersection of canonical sectors to a holomorphic function $\widehat{R}_s$ on $(\mathbb{C} \setminus \{\pm s\}) \times r'\mathbb{D}$, bounded near $\{x = \pm s\}$. Hence $\widehat{R}_s$ is holomorphic on $\mathbb{C} \times r'\mathbb{D}$ by Riemann's removable singularity theorem. From (22.52) follows, for $|x| > 1$,

$$\left| \widehat{R}_s \right| \leq \frac{\left| \frac{P_\lambda}{x} \times x \frac{\partial \mathcal{O}_s^\sharp}{\partial x} \right| + (1 + |\mu_\lambda x|) \left| y \frac{\partial \mathcal{O}_s^\sharp}{\partial y} \right|}{1 - \left| y \frac{\partial \mathcal{O}_s^\sharp}{\partial y} \right|} \leq C\,|x|$$

for some constant $C > 0$ whose existence is guaranteed by Proposition 11 Item (2). Therefore for any fixed $y \in r'\mathbb{D}$ the partial function $x \mapsto \widehat{R}_s(x, y)$ is affine. A change of coordinates of the form

$$(\lambda, x, y) \longmapsto (\lambda, x, y \exp N_\lambda(y))$$

allows to get rid of the term $\widehat{R}_s(0, y)$, so we may as well assume that R_s only depends on y, concluding the proof.

□

So far we have proven Proposition 9 Item (1). The second item can be shown in exactly the same way as in [29, 30], so we shall skip additional details.

22.13.2.2 Gluing Antipodal Realizations (Proof of Lemma 16)

Assume that R_λ is holomorphic with respect to λ. The orbital necklace $\Omega(\varphi_\bullet)$ is compatible: $\eta_\bullet := H_\bullet \circ H_{-\bullet}$ conjugates the necklace dynamics $\Delta_\bullet$ and $\Delta_{-\bullet}$, while $\eta_\bullet^{\circ -1} = \eta_{-\bullet}$.

Conversely, if $\Omega(\varphi_\bullet)$ is compatible, then $\psi_\bullet := H_\bullet^{\circ -1} \circ \eta_\bullet \circ H_{-\bullet}$ is a self-map of T conjugating the holonomy representation of $R_\bullet$ and $R_{-\bullet}$ on T. This map is tangent-to-identity in the y-variable. Because the union of leaves $\mathcal{L}_p$ of $\mathcal{F}_s$ for $p \in T$ contains a uniform, connected neighborhood $\widehat{\mathcal{U}}$ of $\{(\pm s, 0)\}$, there exists a family of paths $\gamma(x)$ linking x to x_* so that $\mathfrak{h}_s^{\gamma(x)}(x, y) \in T$ for every $(x, y) \in \widehat{\mathcal{U}}$. The map built à la Mattei–Moussu [22]

$$\Psi_s \ : \ (x, y) \longmapsto \left(x \, , \ \mathfrak{h}_s^{-\gamma(x)} \circ \psi_s \circ \mathfrak{h}_s^{\gamma(x)}(x, y) \right)$$

is therefore well defined, biholomorphic, and locally conjugates $X_\bullet$ and $X_{-\bullet}$. The conclusion follows from the uniqueness clause of Proposition 9 Item (1), as the linear part of Ψ_s in the y-variable is 1. More details can be found in [29].

22.13.2.3 A Section to the Period Operator (Proof of Proposition 10)

This is really Theorem 15. Being given $X_\bullet \in \mathrm{Holo_c}(\Sigma \times \mathbb{C} \times r'\mathbb{D})$ in normal form, the functions $F_\bullet^\sharp$ of (22.49) induce a function $Q_\bullet := X_\bullet \cdot F_\bullet^\sharp \in \mathrm{Holo_c}(\Sigma \times \mathbb{C} \times r'\mathbb{D})$ with prescribed period as in Sect. 22.13.1.1. Applying again the arguments of Corollary 7 Item (2), we can give a polynomial bound on the growth of $x \mapsto Q_s(x, y)$, so that it must be of the expected form after a final correction in the y-variable to normalize $Q_s(0, y) = 0$. The claim follows from the same reasoning as in Theorem 13 in order to perturb $Q_\bullet$ so that the resulting function is even in s and has same period.

References

1. Birkhoff, G.D.: Déformations analytiques et fonctions auto-équivalentes. Ann. Inst. H. Poincaré **9**, 51–122 (1939)
2. Branner, B., Dias, K.: Classification of complex polynomial vector fields in one complex variable. J. Differ. Equ. Appl. **16**(5–6), 463–517 (2010)
3. Briot, C., Bouquet, C.: Recherches sur les propriétés des fonctions définies par des équations différentielles. Journal de l'École Polytechnique **36**, 133–198 (1856)
4. Bruno, A.D.: Local Methods in Nonlinear Differential Equations. Springer Series in Soviet Mathematics. Springer, Berlin (1989). Part I. The local method of nonlinear analysis of differential equations. Part II. The sets of analyticity of a normalizing transformation. Translated from the Russian by William Hovingh and Courtney S. Coleman, With an introduction by Stephen Wiggins
5. Cerveau, D., Moussu, R.: Groupes d'automorphismes de $(\mathbf{C}, 0)$ et équations différentielles $ydy + \cdots = 0$. Bull. Soc. Math. Fr. **116**(4), 459–488 (1988)
6. Dulac, H.: Recherches sur les points singuliers des équations différentielles. Journal de l'École Polytechnique **2**(9), 1–125 (1904)
7. Dulac, H.: Sur les points singuliers d'une équation différentielle. Ann. Fac. Sci. Toulouse Sci. Math. Sci. Phys. (3) **1**, 329–379 (1909)
8. Elizarov, P.M.: Tangents to moduli maps. In: Nonlinear Stokes Phenomena. Advances in Soviet Mathematics, vol. 14, pp. 107–138. American Mathematical Society, Providence, RI (1993)
9. Fauvet, F.: Résurgence et bifurcations dans des familles à un paramètre. C. R. Acad. Sci. Paris Sér. I Math. **315**(12), 1283–1286 (1992)
10. Glutsyuk, A.A.: Confluence of singular points and the nonlinear Stokes phenomenon. Tr. Mosk. Mat. Obs. **62**, 54–104 (2001)
11. Hukuhara, M., Kimura, T., Matuda, T.: Equations différentielles ordinaires du premier ordre dans le champ complexe. In: Publications of the Mathematical Society of Japan, vol. 7. Mathematical Society of Japan, Tokyo (1961)

12. Hurtubise, J., Lambert, C., Rousseau, C.: Complete system of analytic invariants for unfolded differential linear systems with an irregular singularity of Poincaré rank k. Mosc. Math. J. **14**(2), 309–338, 427 (2014)
13. Klimeš, M.: Analytic classification of families of linear differential systems unfolding a resonant irregular singularity. Preprint (2013). http://arxiv.org/abs/1301.5228
14. Klimeš, M.: Confluence of singularities of nonlinear differential equations via Borel–Laplace transformations. J. Dyn. Control. Syst. **22**(2), 285–324 (2016)
15. Lambert, C., Rousseau, C.: The Stokes phenomenon in the confluence of the hypergeometric equation using Riccati equation. J. Differ. Equ. **244**(10), 2641–2664 (2008)
16. Lambert, C., Rousseau, C.: Complete system of analytic invariants for unfolded differential linear systems with an irregular singularity of Poincaré rank 1. Mosc. Math. J. **12**(1), 77–138, 215 (2012)
17. Lambert, C., Rousseau, C.: Moduli space of unfolded differential linear systems with an irregular singularity of Poincaré rank 1. Mosc. Math. J. **13**(3), 529–550, 553–554 (2013)
18. Loray, F.: Dynamique des groupes d'automorphismes de $\mathbf{C}, 0$. Bol. Soc. Mat. Mexicana (3) **5**(1), 1–23 (1999)
19. Mardešić, P., Roussarie, R., Rousseau, C.: Modulus of analytic classification for unfoldings of generic parabolic diffeomorphisms. Mosc. Math. J. **4**(2), 455–502, 535 (2004)
20. Martinet, J.: Remarques sur la bifurcation nœud-col dans le domaine complexe. Astérisque **150–151**, 131–149, 186 (1987). Singularités d'équations différentielles (Dijon, 1985)
21. Martinet, J., Ramis, J.-P.: Problèmes de modules pour des équations différentielles non linéaires du premier ordre. Inst. Hautes Études Sci. Publ. Math. **55**, 63–164 (1982)
22. Mattei, J.-F., Moussu, R.: Holonomie et intégrales premières. Ann. Sci. École Norm. Sup. (4) **13**(4), 469–523 (1980)
23. Remoundos, G.: Contribution à la théorie des singularités des équations différentielles du premier ordre. Bull. Soc. Math. Fr. **36**, 185–194 (1908)
24. Rousseau, C.: Modulus of orbital analytic classification for a family unfolding a saddle-node. Mosc. Math. J. **5**(1), 245–268 (2005)
25. Rousseau, C.: Normal forms for germs of analytic families of planar vector fields unfolding a generic saddle-node or resonant saddle. In: Nonlinear Dynamics and Evolution Equations. Fields Institute Communications, pp. 227–245. American Mathematical Society, Providence, RI (2006)
26. Rousseau, C.: The moduli space of germs of generic families of analytic diffeomorphisms unfolding of a codimension one resonant diffeomorphism or resonant saddle. J. Differ. Equ. **248**(7), 1794–1825 (2010)
27. Rousseau, C., Christopher, C.: Modulus of analytic classification for the generic unfolding of a codimension 1 resonant diffeomorphism or resonant saddle. Ann. Inst. Fourier (Grenoble) **57**(1), 301–360 (2007)
28. Rousseau, C., Teyssier, L.: Analytical moduli for unfoldings of saddle-node vector fields. Mosc. Math. J. **8**(3), 547–614, 616 (2008)
29. Rousseau, C., Teyssier, L.: Analytical normal forms for purely convergent unfoldings of complex planar saddle-node vector fields (preprint, 2015)
30. Schäfke, R., Teyssier, L.: Analytic normal forms for convergent saddle-node vector fields. Ann. Inst. Fourier **65**(3), 933–974 (2015)
31. Teyssier, L.: Analytical classification of singular saddle-node vector fields. J. Dyn. Control Syst. **10**(4), 577–605 (2004)
32. Teyssier, L.: Équation homologique et cycles asymptotiques d'une singularité nœud-col. Bull. Sci. Math. **128**(3), 167–187 (2004)

33. Teyssier, L.: Examples of non-conjugated holomorphic vector fields and foliations. J. Differ. Equ. **205**(2), 390–407 (2004)

34. Teyssier, L.: Analyticity in spaces of convergent power series and applications. Mosc. Math. J. **15**(3), 527–592 (2015)

35. Teyssier, L.: Germes de feuilletages présentables du plan complexe. Bull. Braz. Math. Soc. **46**(2), 275–329 (2015)

36. Voronin, S.M., Meshcheryakova, Y.I.: Analytic classification of germs of holomorphic vector fields with a degenerate elementary singular point. Vestnik Chelyab. Univ. Ser. 3 Mat. Mekh. Inform. **3**(9), 16–41 (2003)

Chapter 23
The CD45 Case Revisited

Henryk Żołądek

Abstract In Żołądek (Nonlinearity 8:843–860, 1995) the existence of 11 small amplitude limit cycles in a perturbation of some special cubic plane vector field with center was demonstrated. Here we present a new and corrected proof of that result.

Keywords Center • Limit cycle • Melnikov integral

2000 *Mathematics Subject Classification.* Primary 34C05; Secondary 58F21

23.1 Introduction

The center variety of polynomial vector fields in $\mathbb{R}^2$ of degree n consists of vector fields, of degree n, with a center, a singular point surrounded by a family of closed phase curves. When we consider only elementary centers (with the linear part having eigenvalues $\pm i\gamma$) then we get a semi-algebraic variety consisting of several irreducible components. However, only for $n \leq 2$ we know completely the structure of the corresponding center variety: there is one component for $n = 1$ and four components for $n = 2$ (see [11]).

It is conjectured that for $n = 3$ all components of the center variety are divided into two groups: rationally reversible and Darboux integrable. A rationally reversible center of a vector field V is obtained by a pull-back of a polynomial vector field W by means of a rational map Ψ with a fold singularity; the center of V is a preimage of a tangency point of a phase curve of W to the curve of critical values of Ψ. Vector fields with center of the second class have Darboux type first integral $F = F_1^{a_1} \dots F_k^{a_k}$, with polynomials F_j of degrees d_j and constants a_j. For more details we refer the reader to [12, 14], where preliminary lists of reversible (denoted CR_j)

H. Żołądek (✉)
Institute of Mathematics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
e-mail: zoladek@mimuw.edu.pl

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8_23

and Darboux (denoted $CD_{d_1,\dots,d_k}$) components are given.[1] Plausibly these lists are not complete; this follows from numerical investigations of H. C. Graf von Bothmer [7, 8] (who has moved some components CR_j into the Darboux side).

In what follows we adopt the language of Pfaff equations $\Omega = 0$, instead of vector fields: if the vector field is $V = P\partial/\partial x + Q\partial/\partial y$, then $\Omega = Q\mathrm{d}x - P\mathrm{d}y$ is the corresponding 1-form.

Another important problem associated with the center problem is the question of cyclicity of singular points of center or focus type for polynomial vector fields. One asks about the number of limit cycles appearing near the singular point after perturbation $\Omega_\varepsilon = 0$ of a given Pfaff equation $\Omega_0 = 0$, with Ω_0 and Ω_ε of degree n. This constitutes an important element of the bifurcational approach to the 16th Hilbert problem (about the number of limit cycles of polynomial vector fields).[2]

In the case $n = 3$ the space of vector fields has dimension 20. On this space the group of affine automorphisms of $\mathbb{R}^2$ and time rescaling acts; this group has dimension 7. Therefore each component of the center variety and dimension ≥ 7 (and codimension ≤ 13). But the experience shows that each component of the center variety (and for any degree n) has a continuous modulus, its dimension is > 7.[3] Therefore, in order to get an example of a singular point of large cyclicity for $n = 3$ one should take a component of the center variety of maximal codimension, equal 12, and generate 11 limit cycles in some its perturbation. This was the main idea of the work [13].

The chosen component of the cubic center variety was $CD_{4,5}$. In some coordinates one has a first integral of the form

$$F = F(x, y; a) = F_1^{-5} F_2^4 \tag{23.1}$$

where $F_1 = F_1(x, y; a)$ and $F_2 = F_2(x, y; a)$ are some polynomials in x, y and a is a parameter; the precise formulas are given in the next section. One has

$$\mathrm{d}F = M\Omega, \tag{23.2}$$

where $M = -20F/(F_1F_2)$ is an integrating factor and $\Omega = \Omega(x, y; a)$ is a cubic 1-form (corresponding in natural way to a cubic vector field). The component $CD_{4,5}$ is obtained by applying the changes $F \longmapsto \alpha\Phi^*F$ (or $\Omega \longmapsto \alpha\Phi^*\Omega$), with $\alpha \in \mathbb{R}\setminus 0$ and with affine diffeomorphisms Φ of the plane.

[1] When $n > 3$ there exist reversible centers but not rationally, only algebraically, i.e., with algebraic map Ψ (see [3]). Also the class Darboux type integrals should be replaced the with so-called Liouvillian first integrals of the form $\int M\Omega$, where M is an integrating multiplier of the Darboux form and Ω is a polynomial 1-form associated with the vector field.

[2] More precisely, by estimating the cyclicity of the so-called limit sets (like a center or a focus or a polycycle) one can prove the existence of an upper bound for the number of limit cycles for any vector field of degree n). Their details are in [4].

[3] This is not the case for the so-called $p : -q$ resonant complex center problem. For $p = 1,\ q = 2$ and $n = 2$ some components of the corresponding center variety are without moduli (see [6]).

A generic 12-parameter perturbed vector field corresponds to a Pfaff equation of the form

$$\Omega + \sum_{l=1}^{12} \varepsilon_l \xi_l = 0, \tag{23.3}$$

where ε_l are small parameters, ξ_l are generators of the normal space $N_\Omega CD_{4,5} = \mathbb{R}^{12}/T_\Omega CD_{4,5}$, to the component $CD_{4,5}$ (in terms of 1-forms). After rewriting the latter equation,

$$\mathrm{d}F + \sum \varepsilon_l M \xi_l = 0,$$

and integrating it along a part $\Gamma_\varepsilon(t)$ of phase curve (a segment between two consecutive intersections with a Poincaré section and with initial value $F = t$) one gets the formula

$$\Delta F = 20 \sum \varepsilon_l \int_{\Gamma_\varepsilon(t)} \frac{F}{F_1 F_2} \xi_l \tag{23.4}$$

for the increment of the first integral along $\Gamma_\varepsilon(t)$. Approximating the curve $\Gamma_\varepsilon(t)$ by a corresponding oval $\Gamma(t) = \Gamma_0(t) \subset \{F = t\}$ around a center of $\Omega = 0$ we arrive at the condition

$$\sum \varepsilon_l I_{\xi_l}(t) = 0, \tag{23.5}$$

$$I_\xi(t) = \oint_{\Gamma(t)} \frac{\xi}{F_1 F_2} = \oint_{\Gamma(t)} \tilde{\xi}, \tag{23.6}$$

for bifurcation of a limit cycle from the oval $\Gamma(t)$. The corresponding integrals I_ξ are known as the *first order Melnikov integrals*. If the 12 functions $I_{\xi_l}(\cdot)$ were independent, then the corresponding equation would have 11 solutions corresponding to 11 limit cycles after bifurcation.

But the functions I_{ξ_l} are not independent. The reason for this is a deformation $F_\delta = F_\delta(x, y; a)$, $\delta \in (\mathbb{R}, 0)$, of the function F of the form $F_\delta = F_{1,\delta}^{-4} F_{2,\delta}^5$, with suitable polynomials $F_{1,\delta}$ and $F_{2,\delta}$, such that

$$\mathrm{d}F_\delta = M_\delta \left(\Omega + \delta \xi_{12} + \delta^2 \xi_0\right), \tag{23.7}$$

where ξ_{12} is a cubic 1-form (corresponding to $I_{\xi_{12}} \equiv 0$), ξ_0 is a quartic 1-form and $M_\delta = -20 F_\delta / (F_{1,\delta} F_{2,\delta})$ is an integrating multiplier. This implies that the increment ΔF corresponding to the Pfaff equation $\Omega + \delta \xi_{12} = 0$ equals

$$20 \delta^2 \cdot t \cdot I_{\xi_0}(t), \tag{23.8}$$

where the factor $\delta^2 I_{\xi_0}$ is the *second order Melnikov integral*; we prove this formula in Appendix.

So, we should study the function $t \longmapsto \sum_{l=0}^{11} \varepsilon_l I_{\xi_l}(t)$. Besides the small parameters ε_l (with $\varepsilon_0 = \delta^2$) this function contains another parameter, a; moreover, the functions $I_{\xi_l} = I_{\xi_l}(t; a)$ depend analytically on (t, a), i.e., outside a critical curve in $\mathbb{C}^2$. A natural strategy, used also in [13], is to prolong these functions to a value of a such that the function $F_a(x, y)$ is "simple"; $a = 0$ is such a value. Then the corresponding critical point is a saddle, but the cycles $\Gamma(t)$ are continuously deformed and become cycles vanishing at this saddle point.

Unfortunately, it turns out that the 12 functions $I_l(t; 0)$ are linearly dependent (see Sect. 23.4.5). Some combination $\tilde{\kappa} = \kappa / (F_1 F_2)$ of the forms $\xi_l / (F_1 F_2)$ can be written as $\tilde{\kappa} = \mathrm{d}H + K\mathrm{d}F$, with rational functions H and K, and this leads to a third order Melnikov integral of the form

$$I_\nu(t) = \oint_{\Gamma(t)} \nu, \quad \nu = H\mathrm{d}K. \tag{23.9}$$

The main result of the paper is following.

Theorem 1. *For $a = 0$ the Melnikov functions I_{ξ_l}, $l = 0, 1, \ldots, 11$, and I_ν span a space of dimension 12.*

This implies that, for typical a and any collection $\{t_1, \ldots, t_{11}\}$ of values of F, there exists a perturbation $\Omega + \epsilon\omega = 0$ (with a cubic form ω) of the Pfaff equation $\Omega = 0$ with limit cycles $\Gamma_\epsilon(t_j)$, $j = 1, \ldots 11$, such that $\Gamma_\epsilon(t_j) \to \Gamma(t_j)$ as $\epsilon \to 0$. In particular, at least 11 limit cycles can bifurcate from the center after suitable cubic perturbation of the equation $\Omega = 0$.

The above result was stated in [13] but its proof contained several technical mistakes. Firstly, the basis (denoted $\{\omega_1, \ldots, \omega_{11}\}$in [13]) of the space $N_\Omega CD_{4,5}/\mathbb{R}\xi_{12}$ was chosen incorrectly; some combination of the integrals I_{ω_l} vanishes (see Remark 1).

Next, vanishing of a combination of first and of second order Melnikov functions for $a = 0$ was observed in [13]. But, instead of considering a third order Melnikov integral, like in Eq. (23.9), the derivative with respect to a was applied (see Remark 4). Moreover, no geometrical reasons for such relations were provided.

There were also some mistakes in analysis of the topology of the family of Riemann surfaces $\Pi_t = \{F = t\}$ (see Remark 2).

Moreover, it seems that the analysis of the expansions of the integrals along $\Gamma(t)$ for t near the critical values of F was done with insufficient care and details.

Finally, there appeared the paper [10] of P. Yu and M. Han where the Poincaré–Lyapunov focus quantities in the second order with respect to the small parameter ε were calculated and only nine small amplitude limit cycles were detected in this way. This does not contradict Theorem 1 because we do not say about the Taylor expansion of the Melnikov integrals at the critical value corresponding to

the center.[4] The Yu–Han's work was my main motivation to review my previous analysis of the Melnikov integrals in the case $CD_{4,5}$.

Some useful lesson can be extracted from the below research of the case $CD_{4,5}$. New mechanisms for vanishing of the Melnikov type integrals, i.e., of a rational 1-form along cycles in complex levels of rational functions in $\mathbb{C}^2$, are revealed. The vanishing of $I_{\xi_{12}}(t)$ follows from the fact that the perturbing form ξ_{12} is tangent to a component of a quartic center variety (although ξ_{12} is cubic). Next, the vanishing of $I_\kappa(t)$ follows from the property that the perturbing form κ can be split into a sum of two rational forms, each tangent to different components (with Darboux first integrals) of a center variety in a space of rational Pfaff equations (see Remark 3).

23.2 The Component $CD_{4,5}$ and Its Perturbations

Recall that the first integral in the case $CD_{4,5}$ equals $F = F_1^{-5}F_2^4$ (see Eq. (23.1)) where

$$F_1 = x^4 + 4x^2 + 4y, \quad F_2 = x^5 + 5x^3 + 5xy + 5x/2 + a. \tag{23.10}$$

We have $\mathrm{d}F = -20F_1^{-6}F_2^3\Omega$ (see Eq. (23.2)) where

$$\Omega = \left(ax^3 - 6x^2y + 3x^2 - 4y^2 - 2y + 2ax\right)\mathrm{d}x + \left(x^3 + xy + 5x/2 + a\right)\mathrm{d}y. \tag{23.11}$$

The function F has the critical point

$$p_0 = (-a/2, -a^2/4 - 1/2), \quad t_0 = F(p_0) = -a/2. \tag{23.12}$$

If $|a| > 2^{5/4}$, then p_0 is a local extremum (minimum for $a < -2^{5/4}$ and maximum for $a > 2^{5/4}$) and the corresponding singular point of the Pfaff equation is a center; for $|a| < 2^{5/4}$ it is a saddle. For $|a| > 2^{5/4}$ the ovals $\Gamma(t) \subset \mathbb{R}^2$ of the levels $\{F(x, y) = t\}$ for t close to t_0 are closed curves used in definition of the Melnikov integrals in Eq. (23.6).

The perturbation $F_\delta = F_{1,\delta}^{-5}F_{2,\delta}^4$ is defined by

[4] In [10] the authors refer to the paper [11] as the one where the case $CD_{4,5}$ is studied; they evidently have mixed up two my papers. They make calculations with the help of some computer programs. But their result contradicts analogous computer calculations of the focus quantities in first order with respect to ε made (but not published) by C. Christopher; he found ten small amplitude limit cycles.

In [13] I analyzed the Melnikov integrals using only pen and a sheet of paper, here the MAPLE program has turned out useful.

We note also that Christopher in [2] studied Poincaré–Lyapunov quantities in second order with respect to parameters for perturbations of another component $CD_{4,6}$ of the cubic center variety; he has found 11 cycles. In [9] an example of a cubic family with 13 limit cycles is presented.

$$F_{1,\delta} = x^5+5x^3+5xy+5x/2+a+5\delta x^2y + 5\delta y,\ \ F_{2,\delta} = x^4 + 4x^2 + 4y + 4\delta xy. \tag{23.13}$$

We have $\mathrm{d}F_\delta = M_\delta\left(\Omega + \delta\xi_{12} + \delta^2\xi_0\right)$ (see Eq. (23.7)) with

$$\xi_{12} = \left(-7xy^2 + \frac{21}{2}xy + ay\right)\mathrm{d}x + \left(2x^2y - \frac{3}{2}x^2 + ax + y\right)\mathrm{d}y, \tag{23.14}$$

$$\xi_0 = y^2(5 - 3x^2)\mathrm{d}x + xy(x^2 + 1)\mathrm{d}y. \tag{23.15}$$

The variety $CD_{4,5}$ arises from (23.11) by applying the affine changes of variables. Therefore the tangent space $T_\Omega CD_{4,5}$ to $CD_{4,5}$ at Ω is spanned by the tangent vectors to the curves induced by the following 1–parameter changes:

$$1.\ a \to a + \tau;\ 2.\ x \to x + \tau;\ 3.\ y \to y + \tau; 4.\ x \to (1 + \tau)x;$$
$$5.\ y \to (1 + \tau)y;\ 6.\ t \to (1 + \tau)t;\ 7.\ x \to x + \tau y;\ 8.\ y \to y + \tau x.$$

It follows that this space is generated by the following forms:

$$\begin{aligned}
\eta_1 &= (x^3 + 2x)\mathrm{d}x + \mathrm{d}y,\\
\eta_2 &= (3ax^2 - 12xy + 6x + 2a)\mathrm{d}x + (3x^2 + y + 5/2)\mathrm{d}y,\\
\eta_3 &= -(6x^2 + 8y + 2)\mathrm{d}x + x\mathrm{d}y,\\
\eta_4 &= (4ax^3 - 18x^2y + 9x^2 - 4y^2 - 2y + 4ax)\mathrm{d}x + (3x^3 + xy + 5x/2)\mathrm{d}y,\\
\eta_5 &= (-6x^2y - 8y^2 - 2y)\mathrm{d}x + x^3 + 2xy + 5x/2 + a\mathrm{d}y,\\
\eta_6 &= \left(ax^3 - 6x^2y + 3x^2 - 4y^2 - 2y + 2ax\right)\mathrm{d}x + \left(x^3 + xy + 5x/2 + a\right)\mathrm{d}y = \Omega,\\
\eta_7 &= (3ax^2y - 12xy^2 + 6xy + 2ay)\mathrm{d}x + (ax^3 - 3x^2y + 3x^2 - 3y^2 + y/2 + 2ax)\mathrm{d}y,\\
\eta_8 &= (-5x^3 - 7xy + x/2 + a)\mathrm{d}x + x^2\mathrm{d}y.
\end{aligned}$$

The corresponding Melnikov integrals I_{η_j} vanish identically. Recall also that the form ξ_{12} (see Eq. (23.14)) lies outside of the space $T_\Omega CD_{4,5}$, but the corresponding integral $I_{\xi_{12}}$ also vanishes.

Lemma 1. *The 11 forms*

$$\xi_1 = \mathrm{d}y,\ \ \xi_2 = y\mathrm{d}x,\ \ \xi_3 = x^2\mathrm{d}x,\ \ \xi_4 = x^3\mathrm{d}x,\ \ \xi_5 = y^2\mathrm{d}x,\ \ \xi_6 = xy\mathrm{d}x,$$
$$\xi_7 = x^2y\mathrm{d}x,\ \ \xi_8 = xy^2\mathrm{d}x,\ \ \xi_9 = y^3\mathrm{d}x,\ \ \xi_{10} = xy^2\mathrm{d}y,\ \ \xi_{11} = y^3\mathrm{d}y$$

form a basis of the space $N_\Omega CD_{4,5}/\mathbb{R}\xi_{12}$.

Proof. The idea of the proof is to represent the right-hand sides of the formulas for η_j's and for ξ_l's (with respect to the monomial basis) in form of a bloc-triangular matrix.

Observe that the forms $\xi_{11} = y^3\mathrm{d}y$, $\xi_{10} = xy^2\mathrm{d}y$ and $\xi_9 = y^3\mathrm{d}x$ do not appear in the above formulas for η_j.

The forms $\xi_8 = xy^2\mathrm{d}x$, $y^2\mathrm{d}y$ and $x^2y\mathrm{d}y$ appear only in the expressions for η_7 and ξ_{12}, moreover, in regular way; we choose ξ_8.

Next, the forms $xy\mathrm{d}y$, $x^3\mathrm{d}y$, $\xi_5 = y^2\mathrm{d}x$ and $\xi_7 = x^2y\mathrm{d}x$ appear in η_4, η_5 and η_6 (besides η_7 and η_9). Moreover,

$$\begin{aligned}\frac{1}{2}\eta_4 &= 6\xi_7 - x^3\mathrm{d}y + \text{other terms},\\ \eta_5 &= 4y^2\mathrm{d}x - xy\mathrm{d}y + \text{other terms},\\ \tilde{\eta}_6 &= \eta_6 - \eta_5 - \frac{1}{2}\eta_4 \\ &= -\frac{1}{2}ax^3\mathrm{d}x - 3x^2\mathrm{d}x + 2y\mathrm{d}x - 3ax\mathrm{d}x - \frac{5}{2}x\mathrm{d}y - \frac{a}{2}\mathrm{d}y = \text{other terms}.\end{aligned} \tag{23.16}$$

We choose ξ_7 and ξ_5.

The forms $y\mathrm{d}y$, $x^2\mathrm{d}y$, and $\xi_6 = xy\mathrm{d}x$ appear in the formulas for η_2 and η_8 and we choose ξ_6.

The forms $\mathrm{d}x$, $x\mathrm{d}y$, and $x\mathrm{d}x$ appear in η_3, $\tilde{\eta}_6$, and η_1 with nondegenerate triangular matrix; so, these forms can be deleted.

There remain the forms $\xi_4 = x^3\mathrm{d}x$, $\xi_3 = x^2\mathrm{d}x$, $\xi_2 = y\mathrm{d}x$, and $\xi_1 = \mathrm{d}y$. □

Remark 1. In [13] another system of forms was chosen: $\omega_1 = \mathrm{d}x$, $\omega_2 = x\mathrm{d}x$, $\omega_3 = x^2\mathrm{d}x$, $\omega_4 = x^3\mathrm{d}x$, $\omega_5 = \left(18x^2 + 18y + 5\right)\mathrm{d}x$, and $\omega_l = \xi_l$ for $l \geq 6$ (ξ_0 was denoted as ω_{12} in [13]). That choice was wrong, because of the relation

$$a\omega_4 - \omega_5 = \tilde{\eta}_6 - \frac{5}{2}\eta_3 + \frac{3a}{2}\eta_1$$

where $\tilde{\eta}_6$ is given in Eq. (23.16). Thus $aI_{\omega_4}(t) \equiv I_{\omega_5}(t)$.

23.3 Geometry of the Phase Portrait for $a = 0$

Recall that in the case $a = 0$ the singular point $p_0 = (0, -1/2)$ is an integrable saddle. It is a critical point of F with the critical value $t_0 = 0$. One can join the point $a = 0$ with $a = -4$ (when p_0 is a center) in the complex domain avoiding the bifurcational values of the parameter (which correspond to $a^4 = 32$). Then the cycles $\Gamma(t) = \Gamma_{-4}(t)$, i.e., ovals of the curve $\{F(x, y; -4) = t\} \subset \mathbb{R}^2$ with $t > t_0$, are deformed continuously to a suitable cycles $\Gamma(t) = \Gamma_0(t) \subset \Pi_t = \{F(x, y; 0) = t\} \subset \mathbb{C}^2$ (with t close to $t_0 = 0$).

The reasons why we devote our energy to the case $a = 0$ is the relative simplicity of the family of level surfaces of the function

$$F(x, y) = F(x, y; 0) = \frac{F_2^4}{F_1^5} = \frac{x^4\left(x^4 + 5x^2 + 5y + 5/2\right)^4}{\left(x^4 + 4x^2 + 4y\right)^5}. \tag{23.17}$$

23.3.1 The Phase Portrait

The critical level $\{F = 0\}$ of the function (23.17) consists of two algebraic curves $\{x = 0\}$ and $\left\{x^4 + 5x^2 + 5y + 5/2 = 0\right\}$. Besides the saddle point p_0 there are other five singular points of the Pfaff equation $\Omega = 0$ at this level; they are nodes $p_1, \ldots, p_5$ in the intersection of the curves $\{F_1 = 0\}$ and $\{F_2 = 0\}$; but they are irrelevant in our analysis.

Other critical values of F are $t_1 = 1$ and $t_2 = \infty$. The level $\{F = 1\}$ contains the line at infinity, besides the finite part, with two critical points of the equation $\Omega = 0$: $p_6 = (1 : 0 : 0)$ and $p_7 = (0 : 1 : 0)$ (in the homogeneous coordinates $(x_1 : x_2 : x_3)$ of $\mathbb{CP}^2$). The point p_6 is a $1 : -5$ resonant saddle ($F - 1 \approx 20x_3^5 x_2$) and the point p_7 is highly degenerate.

The level $\{F = \infty\}$ is the curve $\{F_1 = 0\}$.

23.3.2 Normalization Near p_0

The normal form for a nondegenerate critical point of saddle type of a function is Cx_1x_2, where $\{x_1 = 0\}$ and $\{x_2 = 0\}$ are the local separatrices. In the case of the singular point $p_0 = (0, -1/2)$ the considered function is

$$(-F)^{1/4} = x\left(x^4 + 5x^2 + 5y + 5/2\right)\left(-x^4 - 4x^2 - 4y\right)^{-5/4}$$

with one separatrix $\{x = 0\}$. Therefore we put $x_1 = x$ and denote x_2 by z, which equals

$$z = \left[\left(y + \frac{1}{2}\right) + x^2 + \frac{1}{5}x^4\right] \cdot \left[1 - 2\left(y + \frac{1}{2}\right) - 2x^2 - \frac{1}{2}x^4\right]^{-5/4}; \qquad (23.18)$$

we have also $C = 5 \cdot 2^{-5/4}$. Introducing the notation

$$r = \left(-2^5 5^{-4} t\right)^{1/4} \qquad (23.19)$$

we rewrite the equation $F(x, y) = t$ as follows:

$$xz = r. \qquad (23.20)$$

Below we need expansion of y in powers of x and z. For this we use the relation (equivalent to Eq. (23.18))

$$y + \frac{1}{2} = -x^2 - \frac{1}{5}x^4 + z\left[1 - 2\left(y + \frac{1}{2}\right) - 2x^2 - \frac{1}{2}x^4\right]^{5/4}.$$

We expand the $5/4$ power term in powers of x and $y + \frac{1}{2}$ and solve the above equation by iterations.

Lemma 2. *We have*

$$y = -\frac{1}{2} + z - x^2 - \frac{5}{2}z^2 + \frac{55}{8}z^3 - \frac{1}{5}x^4 - 20z^4 - \frac{1}{8}x^4 z + \frac{7735}{128}z^5 + \frac{3}{8}x^4 z^2 - \frac{3003}{16}z^6 + \dots \tag{23.21}$$

where the dots mean term of degree ≥ 7.

The continuation of the ovals around the center (say, for $a = -4$) to the value $a = 0$ leads to the following cycle:

$$\begin{aligned} \Gamma(t) &= \left\{(x,z) = \left(r^{1/2}e^{\mathrm{i}\theta}, r^{1/2}e^{-\mathrm{i}\theta}\right) : 0 \leq \theta \leq 2\pi\right\} \\ &= \left\{x = r^{1/2}e^{\mathrm{i}\theta}, y = -\frac{1}{2} + r^{1/2}e^{-\mathrm{i}\theta}\right\}, \quad r = \mathrm{const}\cdot t^{1/4}. \end{aligned} \tag{23.22}$$

23.3.3 New Variables

We introduce new variables u and v by the formulas

$$x = \frac{t^{1/4}}{u}, \quad y = \frac{v^4}{4u^4} - \frac{t^{1/2}}{u^2} - \frac{t}{4u^4}. \tag{23.23}$$

From this we get the following characterization of the level curves of the function F:

$$G(u,v) := 10u^4 - 4v^5 + 5v^4 = t. \tag{23.24}$$

Indeed, since $v^4 = 4u^4\left(y + x^2 + x^4/4\right) = u^4 F_1$ we get $F_1 = v^4/u^4$ and $F_2 = t^{1/4}F_1^{5/4} = t^{1/4}v^5/u^5$. But $F_2 = x\left(F_1 + x^2 + y + 5/2\right)$ which yields the relation $t^{1/4}v^5/u^5 = \left(t^{1/4}/u\right)\left[v^4/u^4 + t^{1/2}/u^2 + v^4/(4u^4) - t^{1/2}/u^2 - t/(4u^4) + 5/2\right]$, equivalent to Eq. (23.24).

Equations (23.23) define a family of maps (morphisms or four-to-one ramified coverings)

$$\Theta_t : \Sigma_t \longmapsto \Pi_t \tag{23.25}$$

between the algebraic curves [5]

[5] We can describe the formulas (23.23)–(23.24) in terms of some algebraic correspondence. In the space $\mathbb{C}^5 = \mathbb{C}^2 \times \mathbb{C}^2 \times \mathbb{C}$, with the coordinates (u,v), (x,y) and t, we define a complex algebraic surface $\mathcal{S}$ by the formulas: $F(x,y) = t$, $(ux)^4 = t$, $(vx)^4 = t\left(x^4 + 4x^4 + 4y\right)$. Then the intersections $\mathcal{S}_t$ of $\mathcal{S}$ with the hyperplanes $\{t = \mathrm{const}\}$ define a family of correspondences between the curves Σ_t and Π_t via the projections of the curves $\mathcal{S}_t$ to the corresponding two-dimensional spaces $\mathbb{C}^2_{u,v}$ and $\mathbb{C}^2_{x,y}$. This correspondence is a morphism (in one direction) although the analogous correspondence defined by the same projections of $\mathcal{S}$ to $\mathbb{C}^2_{u,v}$ and $\mathbb{C}^2_{x,y}$ is not a morphism.

Moreover, change (23.23) applies to the general case, i.e., for arbitrary a, as well. Then Eq. (23.24) is replaced with $10u^4 - 4v^5 + 5v^4 + 4at^{-1/4}u^5 = t$.

$$\Sigma_t = \{G(u,v) = t\} \text{ and } \Pi_t = \{F(x,y) = t\}.$$

Note also that the complex curves Σ_t are the Riemann surfaces of the algebraic function

$$u = \sqrt[4]{\frac{2}{5}v^5 - \frac{1}{2}v^4 + \frac{t}{10}} =: (P(v;t))^{1/4} = \sqrt[4]{P(v)}. \tag{23.26}$$

We denote the projection $(u,v) \longmapsto v$ by π.

23.3.4 Topological Properties of the Function G

The function G has two finite critical points:

$$\tilde{p}_0 = (0,0), \quad t_0 = G(\tilde{q}_0) = 0,$$
$$\tilde{p}_1 = (u,1), \quad t_1 = G(\tilde{q}_1) = 1.$$

The point $\tilde{p}_0$ corresponds to the saddle point p_0 of $\Omega = 0$ (see below) and the point $\tilde{p}_1$ corresponds to the line at infinity (with two singular points $p_6 = (1:0:0)$ and $p_7 = (0:1:0)$ of $\Omega = 0$).

The complex levels $\Sigma_t = \{G(u,v) = t\} \subset \mathbb{C}^2$ are open Riemann surfaces; after adding one point at infinity, $(1:0:0)$ one gets closed Riemann surfaces. The topology of these surfaces can be studied in two ways.

Firstly, we have $G(u,v) = Q(u) + R(v)$. Thus, by a theorem of Thom and Sebastiani (see [1, 15]) we have the following expression of the first homology group of these surfaces:

$$H_1(\Sigma_t, \mathbb{Z}) \simeq \tilde{H}_0(\Phi_q, \mathbb{Z}) \otimes \tilde{H}_0(\Psi_r, \mathbb{Z}),$$

where $\Phi_q = \{Q(u) = q\}$ and $\Psi_r = \{R(v) = r\}$, $q + r = t$ and $\tilde{H}_0$ denotes the reduced zeroth homology group (the sum of coefficients is zero). Moreover, we assume that $t \neq 0, 1$, $q \neq 0$ and $r \neq 0, 1$.

Let $\Phi_q = \{u_1, \ldots, u_4\}$, $u_j = \text{const} \cdot q^{1/4} e^{2\pi \mathrm{i} j/4}$, and $\Psi_q = \{\tilde{v}_1, \ldots, \tilde{v}_5\}$. Then the generators of $\tilde{H}_0(\Phi_q, \mathbb{Z})$ can be chosen as $\{u_2\} - \{u_1\} := \mathbf{2-1}$, $\{u_3\} - \{u_2\} =: \mathbf{3-2}$, $\{u_4\} - \{u_3\} =: \mathbf{4-3}$ and the generators of $\tilde{H}_0(\Psi_r, \mathbb{Z})$ are of the form $\{\tilde{v}_2\} - \{\tilde{v}_1\} := \mathbf{2-1}, \ldots, \{\tilde{v}_5\} - \{\tilde{v}_4\} =: \mathbf{5-4}$. Therefore

$$\left(\mathbf{j}' - \mathbf{j}\right) \otimes \left(\mathbf{k}' - \mathbf{k}\right), \quad j' = j+1, \quad k' = k+1,$$

generate $H_1(\Sigma_t, \mathbb{Z}) \simeq \mathbb{Z}^{12}$.

The function $Q(u)$ has one critical point $u = 0$, with the critical value $q = 0$, and the cycles $\mathbf{j}' - \mathbf{j}$, $j' = j+1$, vanish for $q \to 0$. The function $R(v)$ has two critical

points $v = 0$, with the critical value $r = 0$, and $v = 1$ with the critical value $r = 1$. We can order the points $\tilde{v}_k$ such that $\tilde{v}_1, \tilde{v}_2, \tilde{v}_3, \tilde{v}_4 \to 0$ ($\tilde{v}_k \approx \text{const} \cdot e^{2\pi \mathrm{i}k/4} \cdot r^{1/4}$) and $\tilde{v}_5 \to 5/4$ as $r \to 0$ and $\tilde{v}_4, \tilde{v}_4 \to 1$ as $r \to 1$ (after prolongation for $r \in (0, 1)$); then the cycles $\mathbf{2} - \mathbf{1}$, $\mathbf{3} - \mathbf{2}$, and $\mathbf{4} - \mathbf{3}$ vanish as $r \to 0$ and the cycle $\mathbf{5} - \mathbf{4}$ vanishes as $r \to 1$. It follows that the cycles $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{k}' - \mathbf{k})$, $j = 1, 2, 3$, $k = 1, 2, 3$ vanish for $t = 0$ and the cycles $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{5} - \mathbf{4})$, $j = 1, 2, 3$, vanish for $t = 1$.

Now, let us look at the complex curve Σ_t as the Riemann surface of the algebraic function $(P(v; t))^{1/4}$ (see Eq. (23.26)), with the branching points $v_k = v_k(t)$, $k = 1, \ldots, 5$.

Lemma 3. *The cycles* $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{k}' - \mathbf{k})$ *can be represented as suitable lifts of the loops* Υ_k *in the* v*–plane surrounding the points* $v_{k'} = v_{k+1}$ *and* v_k *in different directions:* Υ_k *form the shape of the digit eight with* v_{k+1} *surrounded in negative direction and with values of* u *above the intersection point being* $u_{j'} = u_{j+1} = \mathrm{i}u_j$ *and* u_j.

Proof. This can be seen after deforming the loops Υ_k to collection of four segments joining the point $v = 0$ with v_k and v_{k-1} and a suitable choice of the branch of the function $(P(v)^{1/4}$ above each segment. The result is the "wedge product" $\{u_j, u_{j+1}\} \wedge \{v_k, v_{k+1}\}$. We also use the fact that, after surrounding a ramification point v_j in positive (respectively, negative) direction, the argument of $P(v)^{1/4}$ increases (respectively, decreases) by $\pi/4$.

There are four lifts of one loop Υ_k but only three of them are homologically independent. □

Let us localize the cycle $\Gamma(t)$ from Eq. (23.22) in the new coordinates. Thus we consider a loop $\tilde{\Gamma}(t)$ in $\Theta_t^{-1}(\Gamma(t))$. From Eq. (23.22) we find that, as $t = \text{const} \cdot r^4 \to 0$, we have $u = \text{const} \cdot t^{1/4}/\left(r^{1/2} e^{\mathrm{i}\theta}\right) = \text{const} \cdot t^{1/8} e^{-\mathrm{i}\theta}$ and $v \approx \left(4yt/x^4\right)^{1/4} \approx \text{const} \cdot r^{1/2} e^{-\mathrm{i}\theta}$ along $\tilde{\Gamma}(t)$. We get

$$\tilde{\Gamma} = \tilde{\Gamma}(t) = \left\{u = \text{const} \cdot t^{1/8} e^{-\mathrm{i}\theta},\ v \approx \text{const} \cdot t^{1/8} e^{-\mathrm{i}\theta} : 0 \le \theta \le 2\pi\right\}. \quad (23.27)$$

This implies the following result.

Lemma 4. *The loop* $\tilde{\Gamma}(t) \subset \Theta_t^{-1}(\Gamma(t))$ *in the* (u, v)*-space is a lift to the Riemann surface of the function* $P^{1/4}(v)$ *of the loop* $\pi\left(\tilde{\Gamma}(t)\right)$ *in the* v*-plane which surrounds (in the negative direction) the four ramification points* $v_1, \ldots, v_4$ *which vanish at* $v = 0$ *as* $t \to 0$. *Moreover, the map* $\Theta_t|\tilde{\Gamma}(t) : \tilde{\Gamma}(t) \longmapsto \Gamma(t)$ *is one-to-one.*

In fact, there are four such lifts, $\tilde{\Gamma}_1, \tilde{\Gamma}_2, \tilde{\Gamma}_3, \tilde{\Gamma}_4$, such that

$$u|_{\tilde{\Gamma}_{k+1}} = \mathrm{i} \cdot u|_{\tilde{\Gamma}_k},$$

and $\tilde{\Gamma}(t) = \tilde{\Gamma}_1$ is one of them. We see also that $\tilde{\Gamma}(t)$ vanishes at the critical point $\tilde{q}_0$. $\tilde{\Gamma}(t)$ can be expressed in the basis $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{k}' - \mathbf{k})$, $j, k = 1, 2, 3$, of vanishing cycles at $\tilde{q}_0$, but we do not need it.

We should say few words about the action of the monodromy group on $H_1(\Sigma_{t_*}, \mathbb{Z})$, $t_* \neq 0, 1$. It is induced by the monodromy diffeomorphisms of the fixed fibre Σ_{t_*} of the Milnor bundle $G : \mathbb{C}^2 \setminus G^{-1}\{0, 1\} \to \mathbb{C} \setminus \{0, 1\}$. There are two monodromy operators $\mathcal{M}_0$ and $\mathcal{M}_1$ corresponding to variation of the value of t around $t = 0$ and $t = 1$. They act on the corresponding groups of vanishing cycles at the points $\tilde{p}_0$ and $\tilde{p}_1$. By the Thom–Sebastiani theorem (see [1, 15]) these operators are of the form $\mathcal{M}_j^Q \otimes \mathcal{M}_j^R$, $j = 0, 1$. Here $\mathcal{M}_0^Q = \mathcal{M}_1^Q$ is induced by the cyclic permutation of the set $\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$, $\mathcal{M}_0^R$ is induced by the cyclic permutation of $\{\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}\}$ (while $\mathbf{5}$ is fixed) and $\mathcal{M}_1^R$ is induced by the transposition $(\mathbf{4}, \mathbf{5})$ (while $\mathbf{1}, \mathbf{2}, \mathbf{3}$ are fixed). More precisely, we have

$$\mathcal{M}_0 \left(\mathbf{j} - \mathbf{j}''\right) \otimes \left(\mathbf{k} - \mathbf{k}''\right) = \left(\mathbf{j}' - \mathbf{j}\right) \otimes \left(\mathbf{k}' - \mathbf{k}\right), \tag{23.28}$$

$j, k = 1, 2, 3, 4$, $j' = j+1$, $j'' = j-1$, $k' = k+1$, $k'' = k-1$ (where $5 = 1$, $0 = 4$, etc.);

$$\mathcal{M}_1 \left(\mathbf{j} - \mathbf{j}''\right) \otimes (\mathbf{5} - \mathbf{4}) = -\left(\mathbf{j}' - \mathbf{j}\right) \otimes (\mathbf{5} - \mathbf{4}), \tag{23.29}$$

$j = 1, 2, 3$, $j' = j + 1$, $j'' = j - 1$.

From Lemma 4 it follows that the cycle $\tilde{\Gamma}(t)$ from Eq. (23.27) is invariant with respect $\mathcal{M}_0$. To see the action of the second monodromy operator on it we note that it has one point of transversal intersection with one of the cycles $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{5} - \mathbf{4})$, realized as one of the lifts to the Riemann surface Σ_t of the eight shape loop Υ_4 around the branching points v_4 and v_5; we can assume that the latter loop is

$$\Delta = \Delta(t) = (\mathbf{2} - \mathbf{1}) \otimes (\mathbf{5} - \mathbf{4}). \tag{23.30}$$

Similarly, the cycles $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{5} - \mathbf{4})$ intersect some cycles vanishing at $\tilde{p}_0$. An analogue of the Picard–Lefschetz formula [1, 15] implies the following formulas:

$$\mathcal{M}_1 \tilde{\Gamma} = \tilde{\Gamma} + (\mathbf{2} - \mathbf{1}) \otimes (\mathbf{5} - \mathbf{4}),$$
$$\mathcal{M}_0 \left(\mathbf{j}' - \mathbf{j}\right) \otimes (\mathbf{5} - \mathbf{4}) = \left(\mathbf{j}' - \mathbf{j}\right) \otimes (\mathbf{5} - \mathbf{4}) + \left(\mathbf{j}' - \mathbf{j}\right) \otimes (\mathbf{4} - \mathbf{1}).$$

The above implies the following result which we use in the sequel.

Lemma 5. *We have*

(a) $\mathcal{M}_0 \tilde{\Gamma} = \tilde{\Gamma}$,
(b) $\mathcal{M}_1 \tilde{\Gamma} = \tilde{\Gamma} + \Delta$,
(c) $\mathcal{M}_1^{4n} \tilde{\Gamma} = \tilde{\Gamma} + 2n \cdot (\mathbf{2} - \mathbf{1} + \mathbf{4} - \mathbf{3}) \otimes (\mathbf{5} - \mathbf{4})$,
(d) $\mathcal{M}_0^{4n} \Delta = \Delta - n\Xi$, *where* $n = 0, 1, 2, \ldots$ *and*

$$\begin{aligned} \Xi = {} & (\mathbf{2} - \mathbf{1}) \otimes (\mathbf{1} - \mathbf{4}) + (\mathbf{3} - \mathbf{2}) \otimes (\mathbf{2} - \mathbf{1}) + (\mathbf{4} - \mathbf{3}) \otimes (\mathbf{3} - \mathbf{2}) \\ & + (\mathbf{1} - \mathbf{4}) \otimes (\mathbf{4} - \mathbf{3}) = \tilde{\Gamma}_1 - \tilde{\Gamma}_4. \end{aligned} \tag{23.31}$$

The second equality in Eq. (23.31) can be seen after applying $\mathcal{M}_0^4$ to Δ. The corresponding variation is a loop which "begins" and "ends" at v_4 and runs twice around the points $v_1,\ v_2,\ v_3$ in opposite directions; it is $\tilde{\Gamma}_1 - \tilde{\Gamma}_4$, where $u|_{\tilde{\Gamma}_k} = \mathrm{i}^{k-1} \cdot u|_{\tilde{\Gamma}}$.

23.3.5 *Normalizations for $t \to 1$*

In this subsection we are interested in the behavior of the cycles $\Gamma(t)$, $\tilde{\Gamma}(t)$, and $\Delta(t)$ as $t \to 1$. The first becomes large (and we describe it in the (x, y)-coordinates) and the second vanishes at $\tilde{q}_1 : u = 0, v = 1$.

The equation $F(x, y) = 1$ takes the form $F_1^5 - F_2^4 = 0$. Near "infinity" we find

$$F_1^5 - F_2^4 = 20x^{14}y + 10x^{12}y^2 + \dots$$

where the dots denote lower order terms with respect to $\deg x + 2 \deg y$. We put

$$t = 1 + 10s \tag{23.32}$$

and assume the following normalization:

$$x = X/s^{1/4},\ \ y = Y/s^{1/2}. \tag{23.33}$$

Lemma 6. *As $s \to 0$ the curve $F(x, y) = t$ becomes*

$$2X^2Y + Y^2 + X^8 = O(s^{1/4}).$$

This implies that the cycle $\Gamma(t)$, with t like in Eq (23.32), behaves like some cycle in the hyperelliptic curve

$$\left(Y + X^2\right)^2 = X^4(1 - X^8); \tag{23.34}$$

the projection of this cycle onto the X-plane is a loop with vertex at $X = 0$ which runs around the point $X = 1$, it can also be taken as the part of the real curve defined in Eq. (23.34) which lies in the half-plane $\{X \geq 0\}$.

Proof. The first statement follows directly from the formulas (23.32)–(23.33).

To determine the behavior of the cycle $\Gamma(t)$ near infinity, we use its description before Lemma 5 (i.e., for the cycle $\tilde{\Gamma}(t)$ in the (u, v)-coordinates).

Recall that $\tilde{\Gamma}(t)$ has one-point intersection with the cycle $\Delta(t)$, vanishing at the point $\tilde{q}_1 : u = 0, v = 1$ (which corresponds to the point $p_1 = (1 : 0 : 0)$ at infinity). Near this point we have the equation:

$$u^4 - (v - 1)^2 (1 + \dots) = s$$

(see Eq. (23.35) below). The projection $\pi(\Delta(t))$ onto the v-plane is the loop Υ_4 with the eight digit shape around the points $v_4 \approx 1 - s^{1/2}$ and $v_5 \approx 1 + s^{1/2}$. These points correspond to $u = 0$. The point $v = 1$ in Υ_4 corresponds to two points $u = s^{1/4}, v = 1$ and $u = -\mathrm{i}s^{1/4}, v = 1$ in $\Delta(t)$. The part of the cycle $\tilde{\Gamma}(t)$ near $\Delta(t)$ has projection onto the v-plane such that it passes through the point $v = 1$ and intersects the segment $\left[1 - s^{1/2}, 1 + s^{1/2}\right]$ transversally at the point $v = 1$. $\tilde{\Gamma}(t)$ passes only through one of the two points $\left(s^{1/4}, 1\right)$ and $\left(-\mathrm{i}s^{1/4}, 1\right)$ above $v = 1$, it is the point $\left(s^{1/4}, 1\right)$.

By Eqs. (23.23) we have $x \approx 1/u = 1/s^{1/4}$ and hence $X = sH^{1/4}x \approx 1$. Therefore, in the (X, Y)-coordinates, the corresponding loop in the X-plane surrounds only one point $X = 1$ (out of four).

Of course, the cycle $\Gamma(t)$ contains also a finite part, i.e., in the (x, y)-space, which we do not claim to control. □

Consider now a neighborhood of the point $\tilde{q}_1 : u = 0, v = 1$. Here the normal form for the function G is the following:

$$G = 10\left(u^4 - w^2\right) + 1. \tag{23.35}$$

The cycle $\Delta(t)$ can be defined via the Riemann surface of the function

$$w = \sqrt{u^4 - s}, \tag{23.36}$$

which follows from the equations $G(u, v) = t = 1 + 10s$. $\Delta(t)$ is a lift of a loop in the u-plane which surrounds the ramification points $u = s^{1/4}$ and $u = \mathrm{i}s^{1/4}$.

The variable v, as a function of w, is calculated as follows. Putting $v = 1 + z$ we get $5v^4 - 4v^5 = 1 - 10z^2\left(1 + 2z + \frac{3}{2}z^2 + \frac{2}{5}z^3\right)$ (see Eq. (23.24)), i.e.,

$$z = w\left(1 + 2z + \frac{3}{2}z^2 + \frac{2}{5}z^3\right)^{-1/2}.$$

We expand the latter power term and solve the obtained equation by iterations. The result is the following.

Lemma 7. *We have*

$$v = 1 + w - w^2 + \frac{7}{4}w^3 - \frac{37}{10}w^4 + \frac{1379}{160}w^5 + O(w^6). \tag{23.37}$$

Moreover, the vanishing cycle $\Delta(t)$, $t = 1 + 10s$, is a lift to the Riemann surface of the function (23.36) of a loop in the u-plane which surrounds the ramification points $u = s^{1/4}$ and $u = \mathrm{i}s^{1/4}$. [6]

[6] One can notice that, in our analysis, we have not considered one more cycle on the curve Π_t (in the x, y coordinates); namely, a cycle vanishing (for $t \to 1$) at the point $p_6 = (1 : 0 : 0)$ at

23.3.6 Normalization for $t \to \infty$

Since here u and v are large, the normal form for G is the following:

$$G = 10u^4 - 4w^5 \tag{23.38}$$

where $4w^5 = 4v^5 - 5v^4$.

Lemma 8. *We have*

$$v = w + a_0 + a_1^{-1} w + \dots \tag{23.39}$$

as $v \to \infty$ (for some constants a_j). Moreover, as $t \to \infty$ the cycle $\tilde{\Gamma}(t)$ is a suitable lift to the Riemann surface of the function $u = \left(\left(t + 4w^5\right)/10\right)^{1/4}$ of a loop in the w-plane which surrounds four (out of five) consecutive ramification points $w_j = e^{2\pi \mathrm{i} j/5} \cdot (-t/4)^{1/5}$, $j = 1, \dots, 4$.

23.4 Expansions of the Melnikov Integrals at the Critical Values

23.4.1 General Properties of the Melnikov Integrals

Firstly, we use the fact that the four-to-one map $\Theta_t : \Sigma_t \longmapsto \Pi_t$, restricted to the loop $\tilde{\Gamma}(t)$, is one-to-one (see Lemma 4). This implies the identity

$$I_\xi(t) = \oint_{\tilde{\Gamma}(t)} \Theta_t^* \tilde{\xi}, \tag{23.40}$$

where

$$\tilde{\xi} = \xi / (F_1 F_2)$$

(see Eq. (23.6)). (Below we usually denote the forms and functions by their old names (without Θ_t^*), when expressed in the u, v variables).

From Eqs. (23.23) we get

$$\mathrm{d}x = -t^{1/4} \frac{\mathrm{d}u}{u^2}, \quad \mathrm{d}y = \frac{v^3}{u^4} \mathrm{d}v - \frac{v^4}{u^5} \mathrm{d}u + 2t^{1/2} \frac{\mathrm{d}u}{u^3} + t \frac{\mathrm{d}u}{u^5} \tag{23.41}$$

infinity. Indeed, with $t = 1 + 10s$ from Eq. (23.32) we find the local equation for Π_t of the form $2x^{14}y = sx^{20} + \dots$, i.e., $x_3^5 (2x_2 + \dots) = s$ (where $x = x_1/x_3$ and $y = x_2/x_3$). The corresponding vanishing cycle is $\{x_3 = s^{1/8} e^{\mathrm{i}\theta},\ x_2 \approx \frac{1}{2} s^{3/8} e^{-5\mathrm{i}\theta} :\ 0 \le \theta \le 2\pi\}$, i.e., $\{x = s^{-1/8} e^{-\mathrm{i}\theta},\ y \approx \frac{1}{2} s^{1/4} e^{-6\mathrm{i}\theta}\}$. In the u, v coordinates we get $\left\{u = s^{1/8} e^{\mathrm{i}\theta},\ v \approx 1 + s^{1/4} e^{2\mathrm{i}\theta}\right\}$.

We see that the projection of this cycle onto the v-plane surrounds two times the ramification points $v_{4,5} \approx 1 \pm s^{1/2}$. Therefore, this cycle is a combination of the cycles $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{5} - \mathbf{4})$.

and $F_1 = \frac{v^4}{u^4}$, $F_2 = t^{1/4}F_1^{5/4} = t^{1/4}\frac{v^5}{u^5}$. Therefore

$$\tilde{\xi} = \frac{\xi}{F_1 F_2} = t^{-1/4}\frac{u^9}{v^9}\xi. \tag{23.42}$$

We can express the latter forms as linear combinations of the monomial forms

$$u^j v^k \mathrm{d}u, \quad u^j v^k \mathrm{d}v.$$

Using the integration by parts and the relations

$$u^4 = \frac{2}{5}v^5 - \frac{1}{2}v^4 + \frac{t}{10}, \quad 2u^3\mathrm{d}u = v^4\mathrm{d}v - v^3\mathrm{d}v, \tag{23.43}$$

which follow from the equations $G(u,v) = t$ and $\mathrm{d}G(u,v) = 0$, we get the following basic integrals:

$$\oint u^j v^k \mathrm{d}v, \quad \oint v^k \frac{\mathrm{d}u}{u};$$

above the range of the powers k and l can be restricted (due to Eqs. (23.43)), but in not unique way (we skip the details).

For further use we introduce a $\mathbb{Z}_2$-**grading** for the monomial forms as follows. We say that the form $u^j v^k \mathrm{d}v$ is **odd** if $j = 1 \pmod 2$ and is **even** otherwise; the form $v^k \mathrm{d}u/u$ is **even**.[7]

By Lemma 1 and formulas (23.23), (23.41), and (23.42) the forms $\tilde{\xi}_j = \xi_j/(F_1F_2)$ are also homogeneous with respect to the $\mathbb{Z}_2$ grading.

Lemma 9. *The forms* $\tilde{\xi}_l$, $l = 1, 4, 6, 8, 11$, *are odd and the forms* $\tilde{\xi}_l$, $l = 0, 2, 3, 5, 7, 9, 10$, *are even.*

If a form $\tilde{\xi}_l$ is odd, then we can represent it as follows:

$$\tilde{\xi}_l = u^{-1}\varphi_l(v)\mathrm{d}v + u\psi_l(v)\mathrm{d}v = \tilde{\xi}_{l,-} + \tilde{\xi}_{l,+}, \tag{23.44}$$

with Laurent polynomials φ_l and ψ_l. If $\tilde{\xi}_l$ is even then we rewrite it as follows:

$$\tilde{\xi}_l = \{\chi_l(v)\mathrm{d}v + \rho_l(v)\mathrm{d}u/u\} + u^2\sigma(v)\mathrm{d}v = \tilde{\xi}_{l,0} + \tilde{\xi}_{l,2}. \tag{23.45}$$

For a cycle Λ in Σ_t and a form $\tilde{\xi}$ we define the integral

$$J_{\tilde{\xi}}^{\Lambda} = J_{\tilde{\xi}}^{\Lambda}(t) = \oint_{\Lambda} \tilde{\xi}; \tag{23.46}$$

thus $I_\xi(t) = J_{\tilde{\xi}}^{\tilde{\Gamma}}$. The integrals of even (odd) forms are called even (odd) integrals.

[7] This grading reflects the anti-symmetry of the unperturbed form Ω with respect to the reflection $x \longmapsto -x$.

Let us now study the monodromy properties of the Melnikov integrals.

Firstly, Lemma 5(a) implies that the integrals $I_\xi(t)$ are single valued functions of r, as $t = \text{const} \cdot r^4 \to 0$ and have Taylor expansions. In fact, the coefficients of these Taylor series are directly related with the saddle analogues of the Poincaré–Lyapunov quantities.

Next, we recall that the cycles $(\mathbf{j}' - \mathbf{j}) \otimes (\mathbf{5} - \mathbf{4})$, $j' = j + 1$, vanishing at $\tilde{p}_1$, are lifts to the Riemann surface of the function $\sqrt[4]{P(v)}$ of the loop Υ_4 (surrounding the ramification points v_4 and v_5 in opposite directions). Thus the values of u at different lifts differ by suitable powers of $\mathrm{i} = \sqrt{-1}$.

If $\tilde{\xi}_l = \tilde{\xi}_{l,-} + \tilde{\xi}_{l,+}$ is odd, then

$$J^{(\mathbf{j}'-\mathbf{j})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}_{l,\pm}} = (\pm\mathrm{i})^{j-1} J^{\Delta}_{\tilde{\xi}_{l,\pm}},$$

where $\Delta = (\mathbf{2} - \mathbf{1}) \otimes (\mathbf{5} - \mathbf{4})$. Hence

$$J^{(\mathbf{2}-\mathbf{1})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}_l} + J^{(\mathbf{4}-\mathbf{3})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}_l} = J^{\Delta}_{\tilde{\xi}_l} + (\pm\mathrm{i})^2 J^{\Delta}_{\tilde{\xi}_l} = 0. \tag{23.47}$$

If $\tilde{\xi}_l = \tilde{\xi}_{l,0} + \tilde{\xi}_{l,2}$ is even, then the integrals $J^{\Lambda}_{\tilde{\xi}_{l,0}}$ are calculated via residua (see Sect. 23.4.4); hence, they are algebraic functions of t. Moreover,

$$J^{(\mathbf{2}-\mathbf{1})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}_{l,2}} + J^{(\mathbf{4}-\mathbf{3})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}_{l,2}} = 2J^{\Delta}_{\tilde{\xi}_{l,2}}.$$

By Lemma 5 we find that the monodromy changes $\mathcal{M}_1$ (corresponding to the change $t \longmapsto 1 + e^{2\pi\mathrm{i}}(t-1)$) of the integrals are the following:

$$\begin{aligned}
&\mathcal{M}_1 : J^{\tilde{\Gamma}}_{\tilde{\xi}} \longmapsto J^{\tilde{\Gamma}}_{\tilde{\xi}} + J^{\Delta}_{\tilde{\xi}},\\
&\mathcal{M}_1 : J^{(\mathbf{j}-\mathbf{j}'')\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}} \longmapsto -J^{(\mathbf{j}'-\mathbf{j})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}},\\
&\mathcal{M}_1^{4n} : J^{\tilde{\Gamma}}_{\tilde{\xi}} \longmapsto J^{\tilde{\Gamma}}_{\tilde{\xi}} + 2n\left(J^{(\mathbf{2}-\mathbf{1})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}} + J^{(\mathbf{4}-\mathbf{3})\otimes(\mathbf{5}-\mathbf{4})}_{\tilde{\xi}}\right)
\end{aligned} \tag{23.48}$$

(with $j' = j + 1$ and $j'' = j - 1$).

If $\tilde{\xi}_l$ is odd, then the latter equations and Eq. (23.47) imply that $J^{\tilde{\Gamma}}_{\tilde{\xi}}(t)$ and $J^{\Delta}_{\tilde{\xi}}(t)$ have algebraic singularities at $t = 1$.[8] If $\tilde{\xi} = \tilde{\xi}_{l,2}$ is even, then $J^{\tilde{\Gamma}}_{\tilde{\xi}_{l,2}}$ has a logarithmic singularity at $t = 1$:

$$J^{\tilde{\Gamma}}_{\tilde{\xi}_{l,2}} = \frac{1}{2\pi\mathrm{i}} J^{\Delta}_{\tilde{\xi}_{l,2}} \cdot \ln s + \Psi_l(s) \tag{23.49}$$

with $J^{\Delta}_{\tilde{\xi}_{l,2}}$ and Ψ single valued functions of $s^{1/4} = ((t-1)/10)^{1/4}$.

[8] One can check directly that the functions $I^{\tilde{\Gamma}}_{\tilde{\xi}_{l,\pm}} + \frac{1}{2}(1 \pm \mathrm{i}) I^{\Delta}_{\tilde{\xi}_{l,\pm}}$ are single valued near $f = 1$ and the functions $I^{\Delta}_{\tilde{\xi}}$ are holomorphic in $s^{1/4} = ((f-1)/10)^{1/4}$.

On the other hand, the odd function $J^{\Delta}_{\tilde{\xi}_l}$ (when nonzero) has a non-algebraic singularity at $t = 0$. Indeed, by Lemma 5(d) we have

$$\mathcal{M}_0^{4n} : J^{\Delta}_{\tilde{\xi}_l} \longmapsto J^{\Delta}_{\tilde{\xi}_l} - nJ^{\Xi}_{\tilde{\xi}_l}.$$

It means that, for $t = \text{const} \cdot r^{1/4}$, we have

$$J^{\Delta}_{\tilde{\xi}_l} = \text{const} \cdot J^{\Xi}_{\tilde{\xi}_l} \cdot \ln r + \Phi_l(r) \tag{23.50}$$

with $J^{\Xi}_{\tilde{\xi}_l} = J^{\tilde{\Gamma}-\tilde{\Gamma}_4}_{\tilde{\xi}_l}$ and Φ single valued functions of r.

Let us summarize the above.

Lemma 10. *(a) As $r = \text{const} \cdot t^{1/4} \to 0$ any function $I_{\xi_l}(t)$ admit Taylor expansions in r.*

(b) After surrounding the value $t = 1$ one gets (from I_{ξ_l}) the function $I_{\xi_l} + J^{\Delta}_{\tilde{\xi}_l}$.

(c) If ξ_l is odd, then I_{ξ_l} and $J^{\Delta}_{\tilde{\xi}_l}$ have algebraic singularities at $t = 1$.

(d) If $\tilde{\xi}_l = \tilde{\xi}_{l,0} + \tilde{\xi}_{l,2}$ is even, then the function $J^{\tilde{\Gamma}}_{\tilde{\xi}_{l,0}}$ is algebraic and $J^{\tilde{\Gamma}}_{\tilde{\xi}_{l,2}}$ has a logarithmic singularity at $t = 1$ of the form (23.49).

(c) If $\tilde{\xi}_l$ is odd, then one summand $J^{\Delta}_{\tilde{\xi}_l}(t)$ of $I_{\xi_l}(t)$, obtained as result of turning around $t = 1$, has a logarithmic singularity of the form (23.50).

(d) As $t \to \infty$ the functions $J^{\Lambda}_{\tilde{\xi}}(t)$ admit Laurent expansions in powers of $t^{-1/20}$.

Item (d) of the latter lemma follows from the fact that the monodromy operator $\mathcal{M}_\infty = \mathcal{M}_0\mathcal{M}_1$, associated with a loop around $t = \infty$, is a tensor product of two operators corresponding to cyclic permutations of two sets of cardinality 4 and 5.

Remark 2. In [13, Lemma 1] it was stated that formula (23.49) holds true for any form $\tilde{\xi}_l$, even or odd. The mistake followed from a mistake in analysis of the monodromy operator $\mathcal{M}_1$; the sign $-$ in Eq. (23.48) was overlooked.

23.4.2 Behavior of the Integrals I_ξ as $t \to 1$

Recall that, as $t \to 1$ and $s \to 0$, the cycle $\Gamma(t)$ contains a "finite" part and a part near infinity (i.e., near the point $p_1 = (1 : 0 : 0)$) which is controlled in Lemma 6.

Since $x \sim s^{-1/4}$, $\mathrm{d}x \sim s^{-1/4}$, $y \sim s^{-1/2}$, $\mathrm{d}y \sim s^{-1/2}$ and $F_1F_2 \approx x^9 \sim s^{-9/4}$ (see Eqs. (23.10)), from Lemma 1 we get that

$$\tilde{\xi}_l = \xi_l / (F_1F_2) \sim O(s^{1/4});$$

the highest growth is achieved for $\xi_{11} = y^3\mathrm{d}y$. Therefore the integrals $I_{\xi_l}(t)$ are finite for $s \to 0$ ($t \to 1$). For this reason we do not compute the values $I_{\xi_l}(1)$ (as we do not control the finite' part of $\Gamma(t)$).

Instead, we compute the derivatives of these integrals using the below Gelfand–Leray formula, but only in the cases when the corresponding integral is divergent as $s \to 0$.

We have

$$\frac{\mathrm{d}}{\mathrm{d}t} I_\xi(t) = I'_\xi = \oint\limits_{F=t} \frac{\mathrm{d}(\tilde{\xi})}{\mathrm{d}F} = \oint \frac{\mathrm{d}(\tilde{\xi})}{\mathrm{d}x \wedge \mathrm{d}y} \frac{\mathrm{d}x}{\partial F/\partial y}. \tag{23.51}$$

Moreover, $\partial F/\partial y = -20F_1^{-6}F_2^3(x^2 + y + 5/2)x \approx -20s^{3/2}X^{-8}(X^2 + Y) = -20s^{3/2}X^{-8}Z$, where

$$Z = Y + X^2$$

satisfies the equation $Z^2 = X^4(1 - X^4)$ and $X \geq 0$. Since practically we can take $\mathrm{d}\left(x^{-9}\xi\right)$ instead of $\mathrm{d}(\tilde{\xi})$ in Eq. (23.51), we find that the following derivatives are potentially divergent:

$$\begin{aligned}
I'_{\xi_{11}} &\approx \frac{9}{20} s^{-3/4} \oint \frac{Y^3 \mathrm{d}X}{X^2 Z} = \frac{9}{20} s^{-3} \oint \frac{3X^8 - 4X^4}{Z} \mathrm{d}X, \\
I'_{\xi_{10}} &= \frac{8}{20} s^{-1/2} \oint \frac{Y^2 \mathrm{d}X}{XZ} + O(1) = \frac{8}{20} s^{-2} \oint \frac{2X^3 - X^7}{Z} \mathrm{d}X + O(1), \\
I'_{\xi_9} &= \frac{3}{20} s^{-1/2} \oint \frac{Y^2 \mathrm{d}X}{XZ} + O(1), \\
I'_{\xi_0} &= -\frac{1}{20} \oint \frac{\mathrm{d}\left(x^{-6}y\mathrm{d}y - 3x^{-7}y^2\mathrm{d}x\right)}{\mathrm{d}x \wedge \mathrm{d}y} \frac{x^8 \mathrm{d}x}{x^2 + y} + O(1) = O(1), \\
I'_{\xi_8} &\approx \frac{1}{10} s^{-1/4} \oint \frac{Y \mathrm{d}X}{Z} = -\frac{1}{10} s^{-1/4} \oint \frac{X^2 \mathrm{d}X}{Z}.
\end{aligned}$$

The fact that the next term in $I'_{\xi_{10}}$, I'_{ξ_9}, and I'_{ξ_0} is finite (not $O(s^{-1/4})$) follows from a corresponding quasi-homogeneity property of the polynomials F_1 and F_2 and of the form ξ_0, where $\deg x = 1$ and $\deg y = 2$. After expressing the corresponding integrals via the Euler Beta function, we get

$$\begin{aligned}
I'_{\xi_{11}} &\approx -\frac{9}{20}\frac{13}{14} B\left(\frac{5}{4}, \frac{1}{2}\right) s^{-3/4}, \\
I'_{\xi_{10}} &\approx \frac{8}{20}\frac{1}{3} B\left(1, \frac{1}{2}\right) s^{-1/2} \\
3I'_{\xi_{10}} - 8I'_{\xi_9} &= O(1), \\
I'_{\xi_8} &\approx -\frac{1}{10}\frac{1}{2} B\left(\frac{3}{4}, \frac{1}{2}\right) s^{-1/4}.
\end{aligned}$$

We summarize this subsection as follows.

Lemma 11. *As $t = 1 + 10s \to 1$ we have*

$$\begin{aligned} I_{\xi_l} &= A_l + O(s), \quad l = 0, \dots, 7, \\ I_{\xi_8} &= A_8 + B_8 s^{3/4} + O(s), \\ I_{\xi_9} &= A_9 + B_9 s^{1/2} + O(s), \\ I_{\xi_{10}} &= A_{10} + B_{10} s^{1/2} + O(s), \\ I_{\xi_{11}} &= A_{11} + B_{11} s^{1/4} + O(s^{3/4}), \end{aligned}$$

where A_l are some constants, B_l are nonzero constants such that $3B_{10} = 8B_9$. This means that the generators of the space of Melnikov integrals can be chosen as

$$I_{\xi_0}, \dots, I_{\xi_9}, I_{\xi_{11}}, I_\zeta,$$

where

$$\zeta := 3\xi_{10} - 8\xi_9 = 3xy^2 \mathrm{d}y - 8y^3 \mathrm{d}x$$

and I_{ξ_8}, I_{ξ_9}, and $I_{\xi_{11}}$ are independent between themselves and of the integrals $I_{\xi_0}, \dots, I_{\xi_7}, I_\zeta$.

Finally, we note that the above expansions are the same as expansions of the analytic terms $\Psi_l(s)$ from Eq. (23.49).

23.4.3 Even Differential Forms in the u, v Variables

Let us rewrite the even forms $\tilde{\xi}_l = \xi_l / (F_1 F_2)$ in the u, v variables. We use Eqs. (23.23)–(23.24), $u^4 = P(v; t)$, $du^4 = 2v^3(v-1)\mathrm{d}v + \frac{\mathrm{d}G}{10}$, integration by parts and often we write $G = G(u, v)$ in place of t.

We begin with the even forms which consist of two summands: $\tilde{\xi}_l = \tilde{\xi}_{l,0} + \tilde{\xi}_{l,2}$ or $\zeta = \zeta_0 + \zeta_2$. We have

$$\begin{aligned} \tilde{\xi}_{0,0} &= -\tfrac{9}{16}\lambda + \mathrm{d}\left\{-\tfrac{1}{4v} + \tfrac{7G}{8v^4} - \tfrac{9G}{20v^5}\right\} + \left\{-\tfrac{7}{8v^4} + \tfrac{9}{20v^5} - \tfrac{7G}{40v^9}\right\}\mathrm{d}G, \\ \tilde{\xi}_{2,0} &= -\tfrac{1}{8}\tfrac{\mathrm{d}v}{v} + \tfrac{1}{8}\mathrm{d}\left\{-\tfrac{1}{v} - \tfrac{G}{v^4} + \tfrac{4G}{5v^5}\right\} + \tfrac{1}{160}\left\{\tfrac{20}{v^4} - \tfrac{17}{v^5} + 16\tfrac{G}{v^9}\right\}\mathrm{d}G, \\ \tilde{\xi}_{3,0} &= 0, \\ \tilde{\xi}_{5,0} &= \tfrac{-1}{16}\lambda + \mathrm{d}\left\{\tfrac{G}{8v^4} - \tfrac{G}{10v^5}\right\} + \left\{-\tfrac{1}{2v^4} + \tfrac{2}{5v^5} - \tfrac{G}{10v^9}\right\}\mathrm{d}G, \\ \tilde{\xi}_{7,0} &= \mathrm{d}\left\{-\tfrac{G}{8v^4} + \tfrac{G}{10v^5}\right\} + \left\{\tfrac{1}{8v^4} - \tfrac{1}{10v^5} + \tfrac{G}{40v^9}\right\}\mathrm{d}G, \\ \tilde{\zeta}_0 &= -\tfrac{9}{64}\mu + \mathrm{d}\left\{-\tfrac{3G}{5v^5} - 1\tfrac{1}{64v^9u^4}\left(G - v^4\right)^3\right\} + \left\{\tfrac{3}{5v^5} + \tfrac{3}{64v^9u^4}\left(G - v^4\right)^2\right\}\mathrm{d}G \end{aligned} \tag{23.52}$$

where

$$\lambda = \left(t - v^4\right)^2 \frac{\mathrm{d}u}{v^9 u}, \quad \mu = (t - v^4)^3 \frac{\mathrm{d}v}{v^{10} u^4}, \tag{23.53}$$

and

$$\begin{aligned}
\tilde{\xi}_{0,2} &= \tfrac{9}{32}\sigma_0 - \tfrac{3}{2}\sigma_1 + \tfrac{9}{5}\sigma_2 + \mathrm{d}\left\{G^{1/2}\left(\tfrac{u^2}{5v^5} + \tfrac{1}{32u^2v^9}(G - v^4)^2\right)\right\} \\
&\quad + \left\{-\tfrac{u^2}{10v^5} - \tfrac{1}{64u^2v^9}\left(G - v^4\right)\left(5G - v^4\right)\right\}\tfrac{\mathrm{d}G}{G^{1/2}}, \\
\tilde{\xi}_{2,2} &= \tfrac{1}{2}\sigma_3, \\
\tilde{\xi}_{3,2} &= -\tfrac{1}{2}\sigma_3, \\
\tilde{\xi}_{5,2} &= -\tfrac{1}{4}\sigma_1 + \tfrac{1}{4}\sigma_2, \\
\tilde{\xi}_{7,2} &= \tfrac{1}{8}\sigma_1 - \tfrac{1}{8}\sigma_2, \\
\tilde{\zeta}_2 &= -\tfrac{27}{16}\sigma_0 - \sigma_1 + \mathrm{d}\left\{-\tfrac{3G^{1/2}}{16u^2v^9}\left(G - v^4\right)^2\right\} \\
&\quad + \tfrac{3}{32u^2v^9}\left(G - v^4\right)\left(5G - v^4\right)\tfrac{\mathrm{d}G}{G^{1/2}},
\end{aligned} \tag{23.54}$$

where

$$\sigma_0 = t^{1/2}\left(t - v^4\right)^2 \frac{\mathrm{d}v}{v^{10}u^2}, \ \ \sigma_1 = t^{3/2}\frac{\mathrm{d}u^2}{v^9}, \ \ \sigma_2 = t^{1/2}\frac{\mathrm{d}u^2}{v^5}, \ \ \sigma_3 = t^{1/2}\frac{u^4\mathrm{d}u^2}{v^9}. \tag{23.55}$$

23.4.4 Algebraic Parts of the Even Melnikov Integrals

Here we are interested in the algebraic parts of the corresponding Melnikov integrals, i.e., in $J^{\tilde{\Gamma}}_{\tilde{\xi}_{l,0}}$. Recall also that the contour $\tilde{\Gamma}$ is such that its projection onto the v-plane surrounds four ramification points $v_1, \ldots, v_4$ of the algebraic function $P(v;t)^{1/4}$ (in negative direction) and avoids the fifth point

$$v_5 =: \tau. \tag{23.56}$$

Note that $t = 5\tau^4 - 4\tau^5$ and

$$\frac{\mathrm{d}u}{u} = \frac{\mathrm{d}u^4}{4u^4} = \frac{\mathrm{d}P}{4P} = \frac{P'_v \mathrm{d}v}{4P} + \frac{1}{40u^4}\mathrm{d}G, \tag{23.57}$$

where $P'_v = 2v^3(v-1)$. Therefore our integrals can be expressed via the residua at $v = 0$, or at $v = \tau$ and at $v = \infty$. Moreover, the residua at $v = \infty$ turn out equal to zero.

We have

$$\begin{aligned} \oint v^k \tfrac{du}{u} &= \tfrac{\pi \mathrm{i}}{2}\tau^k, && k<-1;\\ \oint u^{4m} v^k \mathrm{d}v &= -2\pi \mathrm{i}\cdot \mathrm{res}_{v=0} P^m(v) v^k dv, && m\geq 0;\\ \oint \tfrac{v^k}{u^4}\mathrm{d}v &= 2\pi \mathrm{i}\tfrac{\tau^k}{P_v'(\tau)} = \pi \mathrm{i}\tfrac{\tau^{k-3}}{\tau-1}, && k<4. \end{aligned} \tag{23.58}$$

It gives the following result.

Lemma 12. *The algebraic parts of the integrals I_{ξ_l} for even forms ξ_l (and of ζ_8) are the following:*

$$J^{\tilde\Gamma}_{\tilde\xi_{0,0}} = -\frac{9\pi \mathrm{i}}{2}\frac{(\tau-1)^2}{\tau} =: -\frac{9}{2}K(t),\ J^{\tilde\Gamma}_{\tilde\xi_{2,0}} = -\frac{\pi \mathrm{i}}{2},\ J^{\tilde\Gamma}_{\tilde\xi_{3,0}} \equiv 0,$$
$$J^{\tilde\Gamma}_{\tilde\xi_{5,0}} = -\frac{1}{2}K(t),\ J^{\tilde\Gamma}_{\tilde\xi_{7,0}} \equiv 0,\ J^{\tilde\Gamma}_{\tilde\zeta_0} = 9K(t).$$

More precisely, we have the relation (which follows from Eq. (23.57))

$$8\lambda+\mu = \mathrm{d}\left\{-\frac{10G^2}{9v^9}+\frac{4G}{5v^5}-\frac{10}{v}\right\} + \left\{\frac{1}{5v^9u^4}\left(G-v^4\right)^2+\frac{20G}{9v^9}-\frac{4}{v^5}\right\}\mathrm{d}G. \tag{23.59}$$

between the forms from Eqs. (23.53), which explains the relation between the integrals $J^{\tilde\Gamma}_{\tilde\xi_{0,0}}$ and $J^{\tilde\Gamma}_{\tilde\zeta_0}$.

23.4.5 Third Order Melnikov Integral

Recall that, by Lemma 11 the logarithmic behavior of the integrals near $t=1$ occurs only in the case of even forms $\tilde\xi_l$ and take the form $\frac{2}{\pi \mathrm{i}}J^{\Delta}_{\tilde\xi_{l,2}}\ln(t-1)$ (see Eq. (23.49)). Therefore we should consider the integrals $J^{\Delta}_{\tilde\xi_{l,2}}(t)$. By Lemma 12 it is enough to consider integrals associated with the four forms

$$\tilde\theta_{0,2} := 2\tilde\xi_{0,2}+\tilde\zeta_2 \sim -\frac{9}{8}\sigma_0 - 4\sigma_1 + \frac{18}{5}\sigma_2,$$
$$\tilde\theta_{5,2} := 18\tilde\xi_{5,2}+\tilde\zeta_2 \sim -\frac{27}{16}\sigma_0 - \frac{11}{2}\sigma_1 + \frac{9}{2}\sigma_2,$$

$\tilde\xi_{3,2}\sim -\frac{1}{2}\sigma_3$ and $\tilde\xi_{7,2}\sim \frac{1}{8}\sigma_1-\frac{1}{8}\sigma_2$ (where the relation $\sim$ means that equality modulo exact forms and modulo $\mathrm{d}G$.

The following result follows from Eqs. (23.43).

Lemma 13. *We have the relation* $5\sigma_1 - 9\sigma_2 + 30\sigma_3 = \mathrm{d}G^{1/2}\left\{\frac{16u^2}{v^5} - \frac{20u^2}{v^4}\right\} + \left\{\frac{10}{G^{1/2}v^4} - \frac{8}{G^{1/2}v^5} + \frac{4G^{1/2}}{v^9}\right\} u^2\mathrm{d}G$. *Thus*

$$\tilde{\xi}_3 = \frac{1}{12}\sigma_1 - \frac{3}{20}\sigma_2 + \mathrm{d}G^{1/2}u^2\left\{\frac{1}{3v^4} - \frac{4}{15v^5}\right\} + \left\{-\frac{1}{6G^{1/2}v^4} + \frac{2}{15G^{1/2}v^5} - \frac{G^{1/2}}{15v^9}\right\} u^2\mathrm{d}G. \tag{23.60}$$

Note now that the above four forms $\tilde{\theta}_{0,2}$, $\tilde{\theta}_{5,2}$, $\tilde{\xi}_{3,2}$, and $\tilde{\xi}_{7,2}$ are expressed via three forms σ_0, σ_1, and σ_2, i.e., when taking into account the relation $\sim$. Define the form

$$\tilde{\kappa} = \kappa/\left(F_1F_2\right) = 3\tilde{\theta}_0 - 2\tilde{\theta}_5 + 12\tilde{\xi}_3 = 6\tilde{\xi}_0 + 12\tilde{\xi}_3 - 36\tilde{\xi}_5 + \tilde{\zeta}, \tag{23.61}$$

where $\tilde{\theta}_0 = 2\tilde{\xi}_0 + \zeta$, $\tilde{\theta}_5 = 18\xi_5 + \zeta$ and $\kappa = \left(-9x^2y^2 + 12x^2 - 8y^3 + 13y^2\right)\mathrm{d}x + 3xy\left(x^2 + y + 1\right)\mathrm{d}y$ (when expressed in the x, y coordinates. We have $\tilde{\kappa} \sim 0$, i.e.,

$$I_\kappa(t) = J_{\tilde{\kappa}}^{\tilde{\Gamma}}(t) \equiv 0,$$

but we need a more subtle statement.

We have

$$\tilde{\kappa} = \mathrm{d}H + K\mathrm{d}G, \tag{23.62}$$

where

$$\begin{aligned} H &= \left\{-\frac{3}{32v} + \frac{3G}{4v^4} + \frac{111G}{80v^5} + \frac{5G^2}{32v^9} - \frac{1}{64v^9u^4}\left(G - v^4\right)^3\right\} + \sqrt{G}u^2\left(\frac{4}{v^4} - \frac{2}{v^5}\right) \\ &= H_0 + H_2, \\ K &= \left\{\frac{179G}{80v^9} + \frac{51}{4v^4} - \frac{843}{80v^5} + \frac{3}{160v^9u^4}\left(G - v^4\right)^2\right\} + \left(\frac{1}{v^5} - \frac{2}{v^4} - \frac{4G}{5v^9}\right)\frac{u^2}{\sqrt{G}} \\ &= K_0 + K_2. \end{aligned}$$

Remark 3. From Eqs. (23.23) and (23.42) we find $u = t^{1/4}/x, v^4 = tF_1/x^4$ and $\frac{1}{v^9} = \frac{1}{t^2}\frac{t^{9/4}}{u^9}\left(t^{-1/4}\frac{u^9}{v^9}\right) = \frac{x^9}{t^2F_1F_2}$.(i.e., for $F(x,y) = G(u.v) = t$). This implies $\frac{1}{v} = v^8\frac{1}{v^9} = \frac{xF_1}{F_2}$, $\frac{t}{v^4} = \frac{x^4}{F_1}$, $\frac{t}{v^5} = t^2\frac{v^4}{t}\frac{1}{v^9} = \frac{x^5}{F_2}$, $\frac{t^2}{v^9} = \frac{x^9}{F_1F_2}$, $\frac{t^{1/2}u^2}{v^5} = \frac{1}{x^2}\frac{t}{v^5} = \frac{x^3}{F_2}$, and $\frac{\left(t-v^4\right)^3}{v^9u^4} = \left(-4t\left(y + x^2\right)/x^4\right)^3\frac{1}{v^9}\frac{x^4}{t} = -64\frac{x\left(y+x^2\right)^3}{F_1F_2}$. Therefore we get the function H expressed in the x, y variables: $H = \frac{x}{F_1F_2}\{-\frac{3}{32}F_1^2 + \frac{3}{4}x^3F_2 + \frac{3}{10}x^4F_1 + \frac{5}{32}x^8 + \frac{6}{5}x^2F_1 + \left(y + x^2\right)^3\}$. Together with the fact that $\Omega/(F_1F_2) = \frac{1}{5}\mathrm{d}\ln F_2 - \frac{1}{4}\mathrm{d}\ln F_1$,

we find that the perturbation $\Omega + \varepsilon\kappa_1 = 0$, with small ε and $\kappa_1 = F_1F_2\mathrm{d}H$, has first integral of the form of a logarithm of a generalized Darboux function: $\frac{1}{5}\ln F_2 - \frac{1}{4}\ln F_1 + \varepsilon H$.

On the other hand, the perturbation $\Omega + \varepsilon\kappa_2 = 0$, with $\kappa_2 = F_1F_2K\mathrm{d}F$ has rational first integral F (with a rational integrating multiplier).

In such situation it is natural to consider the following higher order Melnikov integral:

$$I_\nu(t) := J_\nu^{\tilde\Gamma}(t) = \oint\limits_{\tilde\Gamma(t)} H\mathrm{d}K; \tag{23.63}$$

it is a coefficient in the next term of the expansion of the expansion of the Poincaré return map (see [5]). The form $\nu := H\mathrm{d}K$ is even and can be written as $\nu = \nu_0 + \nu_2$, where $\nu_2 = H_0\mathrm{d}K_2 + H_2\mathrm{d}K_0$ and

$$\nu_0 = H_0\mathrm{d}K_0 + H_2\mathrm{d}K_2. \tag{23.64}$$

(The form ν replaces, in a sense, the form $\tilde\xi_3$). The integrations of ν_0 gives an algebraic function of t.

Lemma 14. *We have*

$$I_\nu(t) = -\frac{27}{80}\frac{(\tau-1)\,(37\tau-49)}{\tau^6}$$

where τ is the same as in Eq. (23.56). This function is independent of the functions appearing in the thesis of Lemma 12.

Proof. As in the previous section we use the residuum theorem with G replaced with the constant t. It is rather easy that $\oint H_2\mathrm{d}K_2 \equiv 0$. Let us write $H_0 = A(v) - \frac{1}{64}B(v)/u^4$ and $K_0 = C(v) + \frac{3}{160}D(v)/u^4$, where $B = \left(t - v^4\right)D = \left(t-v^4\right)^3/v^9$. Again we have $\oint A\mathrm{d}C \equiv 0$.

Next, the integration by parts leads to the integration of the following forms:

$$\tfrac{3}{160}A\mathrm{d}\tfrac{D}{u^4} - \tfrac{1}{64}\tfrac{B}{u^4}\mathrm{d}C = -\tfrac{\left(t-v^4\right)^2}{u^4v^9}\left\{\tfrac{3}{160}\mathrm{d}A + \tfrac{t-v^4}{64}\mathrm{d}C\right\},$$

$$\tfrac{-3}{160\cdot 64}\tfrac{B}{u^4}\mathrm{d}\tfrac{D}{u^4} = -\tfrac{3}{5\cdot 2^{10}}\tfrac{\left(t-v^4\right)^4}{u^8v^{15}}\mathrm{d}v.$$

Like in the proof of Lemma 12 we calculate the residua of these forms at $v = \tau$.

In calculation of the residuum of the first form we use Eqs. (23.58). In calculation of the residuum of the second form we use the expansions

$$u^{-8} = \left[2\tau^3(\tau-1)(v-\tau)\right]^{-2}\left\{1 - \left((4\tau-3)\,/\,(\tau\,(\tau-1))\right)(v-\tau) + \ldots\right\},$$

$$\left(t-v^4\right)^4 = \left[4\tau^4(\tau-1)\right]^4\left\{1 + \left(4/\,(\tau\,(\tau-1))\right)(v-\tau) + \ldots\right\},$$

$$v^{-15} = \tau^{-15}\left\{1 - 15\,(v-\tau)\,/\tau + \ldots\right\},$$

as $v - \tau \to 0$. □

Remark 4. In [13, Lemma 3] a vanishing of some Melnikov integral, like $I_\kappa(t)$, was also observed. But, instead of considering higher order Melnikov function (like above), the derivative $\frac{\partial}{\partial a} I_\kappa|_{a=0}$ of the latter integral with respect to the parameter a was analyzed. That approach seems to be not the best one, because one can imagine a situation when a relation between basic Melnikov functions exists for all values of a and, moreover, with coefficients depending on this parameter.

Note that we could use also the part $\nu_2 = H_0 \mathrm{d}K_2 + H_2 \mathrm{d}K_0$ of the form ν. But the corresponding integral is not an algebraic function and one should rather integrate it over the cycle $\Delta(t)$ (vanishing at $(u, v) = (0, 1)$), like in the next section. We have chosen the algebraic part.

23.4.6 Logarithmic Part of Even Integrals at $t = 1$

Recall that, by Lemma 10 the logarithmic behavior of the integrals near $t = 1$ occurs only in the case of even forms $\tilde{\xi}_l$ and take the form $\frac{2}{\pi i} J^\Delta_{\tilde{\xi}_{l,2}} \ln(t-1)$ (see Eq. (23.49). Here we expand the integrals $J^\Delta_{\sigma_k}(t)$, $k = 0, 1, 2$, in powers of $s = (t-1)/10$. In fact, we shall expand the integrals of the forms σ_0, σ_1, and σ_2 from Eq. (23.55).

With the normal form (23.35), i.e., $u^4 - w^2 = s$, where w is defined via Eq. (23.37), we make the following substitutions:

$$u^2 = \frac{1}{2x}\left(x^2 + s\right), \quad w = \frac{1}{2x}\left(x^2 - s\right). \tag{23.65}$$

We have

$$\frac{\mathrm{d}w}{u^2} = \frac{\mathrm{d}u^2}{w} = \frac{\mathrm{d}x}{x}.$$

Then the corresponding integrals are equal $2\pi\mathrm{i}$ times the residuum of the corresponding form at $x = 0$.

We arrive at the following:

Lemma 15. *As $s \to 0$ we have*

$$\begin{aligned}
\frac{1}{2\pi\mathrm{i}} J^\Delta_{\sigma_0} &= -8s - 12s^2 + \frac{16\,811}{32}s^3 + \dots, \\
\frac{1}{2\pi\mathrm{i}} J^\Delta_{\sigma_1} &= -\frac{9}{2}s - \frac{5409}{32}s^2 - \frac{34\,749}{512}s^3 + \dots, \\
\frac{1}{2\pi\mathrm{i}} J^\Delta_{\sigma_2} &= \frac{5}{2}s + \frac{315}{32}s^2 - \frac{84\,185}{512}s^3 + \dots.
\end{aligned}$$

Therefore the above three functions are independent and hence the corresponding functions $J^{\tilde{\Gamma}}_{\tilde{\theta}_0}$, $J^{\tilde{\Gamma}}_{\tilde{\theta}_5}$, and $J^{\tilde{\Gamma}}_{\tilde{\xi}_7}$ are independent.

23.4.7 Logarithmic Parts of the Odd Integrals Near $t = 0$

For the odd forms we have the following formulas, with $\sim$ meaning the equivalence modulo exact forms and modulo $\mathrm{d}G$:

$$\tilde{\xi}_1 \sim \frac{18}{7}\rho_1 + \frac{9}{5}\rho_2,\ \tilde{\xi}_4 \sim -\frac{9}{5}\rho_2,\ \tilde{\xi}_6 \sim -\frac{9}{5}\rho_2 - \frac{3}{4}\rho_3 + \frac{5}{12}\rho_4, \tag{23.66}$$

where

$$\begin{aligned} \rho_1 &= \tfrac{t^{1/4}u^7\mathrm{d}v}{v^{10}} \sim -\tfrac{7x\mathrm{d}x}{9F_1F_2},\ \rho_2 = \tfrac{t^{3/4}u^5\mathrm{d}v}{v^{10}} \sim -\tfrac{5x^3\mathrm{d}x}{9F_1F_2}, \\ \rho_3 &= \tfrac{t^{5/4}u^3\mathrm{d}v}{v^{10}} \sim -\tfrac{x^5\mathrm{d}x}{3F_1F_2},\ \rho_4 = \tfrac{t^{1/4}u^3\mathrm{d}v}{v^6} \sim -\tfrac{3xF_1\mathrm{d}x}{5F_1F_2} \end{aligned} \tag{23.67}$$

(compare Remark 3).

Recall that the logarithmic singularity at $t = 0$ holds for the parts $J^{\Delta}_{\tilde{\xi}_l}$ of I_ξ which appear after turning around $t = 1$ (see); we have $J^{\Delta} = -\frac{1}{2\pi\mathrm{i}}J^{\Xi}\ln t^{1/4} +$ (regular part) . Here we calculate initial terms of the Taylor expansions of $J^{\Xi}_{\tilde{\xi}_l}(t)$ at $t = 0$.

Since $\Xi = \tilde{\Gamma} - \tilde{\Gamma}_4$ and $u|_{\tilde{\Gamma}_4} = -\mathrm{i}\cdot u|_{\tilde{\Gamma}}$, we have to compute the integrals

$$(1-\mathrm{i})\oint_{\tilde{\Gamma}}\rho_1 = \frac{7\,(1-\mathrm{i})}{9}\oint_{\Gamma}\frac{x\mathrm{d}x}{F_1F_2},\quad (1+\mathrm{i})\oint_{\tilde{\Gamma}}\rho_2 = \frac{5\,(1+\mathrm{i})}{9}\oint_{\Gamma}\frac{x^3\mathrm{d}x}{F_1F_2},$$

$$(1-\mathrm{i})\oint_{\tilde{\Gamma}}\rho_3 = -\frac{1-\mathrm{i}}{9}\oint_{\Gamma}\frac{x^5\mathrm{d}x}{F_1F_2},\quad (1-\mathrm{i})\oint_{\tilde{\Gamma}}\rho_4 = -\frac{3\,(1-\mathrm{i})}{5}\oint_{\Gamma}\frac{xF_1\mathrm{d}x}{F_1F_2}.$$

By the analysis of Sect. 23.3.2 the (complex) levels $\Pi_t = \{F = t\}$ of the first integral near the critical point $p_0 : x = 0, y = -1/2$ can be written in the form $xz = r$, where $r = \left(-2^5 t\right)^{1/4}/5$ and $z = F_2/\left(5x\right)\cdot\left(-F_1/2\right)^{-5/4} = y + \frac{1}{2} + \ldots$ (see Eqs. (23.18)–(23.20)). Moreover, the cycle $\Gamma(t)$ is a lift to Σ_t of a small loop around $x = 0$ in the x-plane. It follows that

$$J^{\tilde{\Gamma}}_{\tilde{\xi}}(t) = 2\pi\mathrm{i}\cdot\mathrm{res}_{x=0}\tilde{\xi}|_{z=r/x}. \tag{23.68}$$

So, our aim is to express the forms ξ_l and the factor $(F_1F_2)^{-1}$ as depending on x and r (with $\mathrm{d}r = 0$).

We have

$$F_1F_2 = -10xz\left(-F_1/2\right)^{9/4} = -10r\left(1 - 2\left(y + 1/2\right) - 2x^2 - x^4/4\right)^{9/4}.$$

Moreover, we have formula (23.21) for y as a function of x and z. We put $x = \sqrt{r}w$, $z = \sqrt{r}/w$ and we get

$$y = -\frac{1}{2} + \frac{1}{w}r^{1/2} - \left(\frac{5}{2w^2} + w^2\right)r + \frac{55}{8w^3}r^{\frac{3}{2}} - \left(\frac{20}{w^4} + \frac{1}{5}w^4\right)r^2 - \left(\frac{1}{8}w^3 - \frac{7735}{128w^5}\right)r^{\frac{5}{2}} + \left(\frac{3}{8}w^2 - \frac{3003}{16w^6}\right)r^3 + \dots,$$

$$-10r\,(F_1F_2)^{-1} = 1 - \frac{2}{w}r^{1/2} + \frac{5}{w^2}r - \frac{55}{4w^3}r^{3/2} + \left(\frac{40}{w^4} + \frac{3}{20}w^4\right)r^2 + \left(\frac{1}{4}w^3 - \frac{7735}{64w^5}\right)r^{3/2} - \left(\frac{3}{4}w^2 - \frac{3003}{8w^6}\right)r^3 + \dots$$

This allows to compute the residua from Eq. (23.68) with the following result.

Lemma 16. *As $t = \mathrm{const}\cdot r^4 \to 0$ we have*

$$\frac{10}{2\pi\mathrm{i}}J^{\Xi}_{\tilde{\xi}_1} = 10\,(1-\mathrm{i})\,r + 40(1+\mathrm{i})r^3 + O(r^5),$$
$$\frac{10}{2\pi\mathrm{i}}J^{\Xi}_{\tilde{\xi}_4} = -40(1+\mathrm{i})r^3 + O(r^5),$$
$$\frac{10}{2\pi\mathrm{i}}J^{\Xi}_{\tilde{\xi}_6} = 7\,(\mathrm{i}-1)\,r + \left(\frac{1001}{32}\,(1-\mathrm{i}) - 40\,(1+\mathrm{i})\right)r^3 + O(r^5).$$

23.4.8 Behavior of Odd Integrals as $t \to \infty$ via Euler Beta Integrals

Recall that we are left with the forms $\tilde{\xi}_1$, $\tilde{\xi}_4$, and $\tilde{\xi}_6$ expressed via the forms ρ_j (see Eqs. (23.66)). Recall also that the Riemann surface Σ_t, for large $|t|$ and large $|u| + |v|$, is equivalent to $\{10u^4 - 4w^5 = f\}$ with $w = v + O(1)$ (see Lemma 8).

Let us describe more precisely the cycle $\tilde{\Gamma}(t)$ for large and real t. When $0 < t < 1$ the function $\sqrt[4]{P(v;t)}$ has three real ramification points, $v_2 < 0 < v_4 < 1 < v_5$ and two non-real ones, v_1 and v_3; the projection $\pi(\tilde{\Gamma})$ of the cycle $\tilde{\Gamma}(t)$ surrounds the points $v_1, \dots, v_4$ in negative direction. When t avoids the critical value $t_1 = 1$ along small semicircle (in $\mathbb{C}$) from above, and next moves along the half-line $\{t > 1\}$, the points v_4 and v_5 become non-real: $\mathrm{Im}\, v_4 = -\mathrm{Im}\, v_5 > 0$. For large real t we get $v_4 \sim (f/4)^{1/5}\epsilon$, $v_1 \sim (f/4)^{1/5}\epsilon^3$, $v_2 \sim -(f/4)^{1/5}$, $v_3 \sim (f/4)^{1/5}\epsilon^{-3}$, $v_5 \sim (f/4)^{1/5}\epsilon^{-1}$, where

$$\epsilon = e^{\pi\mathrm{i}/5}$$

and the loop $\pi(\tilde{\Gamma})$ surrounds the points $v_4,\ v_1,\ v_2$, and v_3. We deform the latter loop to a collection of arcs of circles of small radii around $v_4,\ v_1,\ v_2, v_3$, and $v=0$ and of segments along radii $\{\arg v = \pi/5, 3\pi/5, \pi, -3\pi/5\}$ joining corresponding arcs.

When calculating a leading term of corresponding integrals, which are integrals of monomial forms, we make the following substitutions:

$$v \approx w = (t/4)^{1/5}\, e^{\mathrm{i}\theta_j} V^{1/5},\ \ u = (t/10)^{1/5}\, \gamma_j\, (1-V)^{1/4}\,, \tag{23.69}$$

where $\theta_j = \pi/5,\ 3\pi/5, \pi,\ -3\pi/5$ and γ_j are suitable powers of $\mathrm{i} = e^{\mathrm{i}\pi/2}$. We keep in mind the fact that $u>0$ at the lower ridge of the segment $[0, v_4]$.

We arrive at Euler type integrals $\int V^{\alpha-1}\,(1-V)^{\beta-1}\,\mathrm{d}V$ along the contour just defined. Here the numbers α and β are non-integer and we can use analytic continuation of such integrals as functions of the parameters α and β (one begins with the situation when $\mathrm{Re}\,\alpha>0$ and $\mathrm{Re}\,\beta>0$). The result is $\mathrm{const}\cdot B(\alpha,\beta)$, where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\,\Gamma(\alpha+\beta)$ is the Euler Beta function.

For example, in calculation of $L_1(t) = \oint \rho_2$ we get

$$\begin{aligned} L_2 \sim\ & -t^{3/4}\cdot\frac{1}{5}\left(\frac{t}{4}\right)^{-9/5}\left(\frac{t}{10}\right)^{5/4} \\ & \cdot\epsilon^{-9}\,(1-\mathrm{i})\left(1+\left(\mathrm{i}\epsilon^2\right)+\left(\mathrm{i}\epsilon^2\right)^2+\left(\mathrm{i}\epsilon^2\right)^3\right) B\left(\frac{9}{4},-\frac{9}{5}\right) \\ =\ & Ct^{1/5}\cdot 10^{-5/4}\,\frac{\sin\,(9\pi/5)}{\sin\,(9\pi/20)} B\left(\frac{9}{4},-\frac{9}{5}\right), \end{aligned}$$

where the minus comes from the negative direction of $\pi(\tilde{\Gamma})$ and $C = \frac{\sqrt{2}}{5}4^{9/5}e^{3\mathrm{i}\pi/10}$. Analogously, we calculate other integrals $L_j = \oint \rho_j,\ j=1,3,4$, and we get the following result.

Lemma 17. *As $t\to\infty$ we have*

$$\begin{aligned} L_1 &\approx -Ct^{1/5}\cdot 10^{-7/4}\,\frac{\sin(\pi/5)}{\sin\,(\pi/20)} B\left(\frac{11}{4},-\frac{9}{5}\right) \approx 18.679\cdot Ct^{1/5}, \\ L_2 &\approx -1.0922\cdot Ct^{1/5}, \\ L_3 &\approx -Ct^{1/5}\cdot 10^{-3/4}\,\frac{\sin(\pi/5)}{\sin\,(\pi/20)} B\left(\frac{7}{4},-\frac{9}{5}\right) \approx -0.53368\cdot Ct^{1/5}, \\ L_4 &= O(1). \end{aligned}$$

This, together with Lemma 16, implies that the functions $I_{\xi_1}(t),\ I_{\xi_4}(t)$, and $I_{\xi_6}(t)$ are independent and are independent on the even integrals.

The second statement of the latter lemma involves calculation of the determinant of a corresponding 3×3 matrix.

Corollary 1. *The functions $I_{\xi_0}, \dots, I_{\xi_{11}}, I_\nu$ span a space of dimension 12.*

23.4.9 Proof of Theorem 1

We have two possibilities: either the 12 functions $I_{\xi_0}, \dots, I_{\xi_{11}}$ (from the second order analysis) are independent for a typical value of the parameter a or not.

In the first case, by a variation of the coefficients ε_l in the Melnikov function $\sum_{l=0}^{11} \varepsilon_l I_{\xi_l}(t)$ we can localize its zeroes (maybe not all) in any fixed collection $\{t_1, \dots, t_{11}\}$ of points. By suitable application of Implicit Function Theorem to the Poincaré return map, i.e., for a typical collection $\{t_1, \dots, t_{11}\}$, we get a family of limit cycles $\Gamma_\varepsilon(t_j) \to \Gamma(t_j)$ as $|\varepsilon| \to 0$.

In the second case we have $I_\kappa(t; a) \equiv 0$ for a linear combination $\kappa = \kappa(x, y; a)$ of the forms $\xi_0, \dots, \xi_{11}$ (with coefficients depending on a). In such a case there exists a general formula for a next order Melnikov function

$$M(t; a) = \oint_{\Gamma(t)} H \frac{\mathrm{d}\tilde{\kappa}}{\mathrm{d}F}, \tag{23.70}$$

where $\tilde{\kappa} = \kappa / (F_1 F_2)$,

$$H(P) = \int_{P_0}^{P} \tilde{\kappa}$$

and the integration runs along a path in Π_t from a fixed point P_0 to $P = (x, y)$. For $a = 0$ integral (23.70) reduces to the function $I_\nu(t)$ from Eq. (23.63).

Now, the function $M(t; a)$ depends (locally) analytically on (t, a). By Corollary 1, for $a = 0$, the functions $I_{\xi_0}(t), \dots, I_{\xi_{11}}(t)$ and $I_\nu(t)$ span a 12-dimensional space (of functions of t). Therefore for generic a the corresponding space also has dimension 12.

The further proof is standard and we skip it. □

Appendix: Proof of Formula (23.8)

Recall that Eq. (23.8) calculates the increment ΔF of the first integral along the part $\Gamma_\delta(t)$ of the phase curve (between two consecutive intersections with a Poincaré section and with initial value t of the F) of the Pfaff equation $\Omega + \delta\xi_{12} = 0$.

Since $\mathrm{d}F = M\Omega$, $M = -20M/(F_1F_2)$, we have $\mathrm{d}F + \delta M\xi_{12} \equiv 0$ along $\Gamma_\delta(t)$ and hence $\Delta F = -\delta \int M\xi_{12}$, with the integral along $\Gamma_\delta(t)$ which is close to $\Pi_t = \{F = t\}$; $dist\,(\Gamma_\delta(t), \Pi_t) = O(\delta)$.

On the other hand, we have the relation $\mathrm{d}F_\delta = M_\delta\left(\Omega + \delta\xi_{12} + \delta^2\xi_0\right)$ (see Eq. (23.7)) which means that the phase curves of the Pfaff equation $\Omega + \delta\xi_{12} + \delta^2\xi_0 = 0$ lie in $\Pi_{\delta,t}$. We have dist $(\Gamma_\delta(t), \Pi_{\delta,t}) = O(\delta^2)$, where $\Pi_{\delta,t} = \{F_\delta = t\}$ are the levels of the perturbed first integral F_δ which means that

$$\int_{\Gamma_\delta(t)} M\xi_{12} = \int_{F_\delta = t} M\xi_{12} + O(\delta^2).$$

But Eq. (23.7) implies that

$$\int_{F_\delta = t} M\left(\Omega + \delta\xi_{12} + \delta^2\xi_0\right) = \int_{F_\delta = t} MM_\delta^{-1}\mathrm{d}F_\delta = 0.$$

We have also

$$\int_{F_\delta = t} M\Omega = \oint_{F_\delta = t} \mathrm{d}F = 0.$$

Therefore

$$\begin{aligned}\Delta F &= -\delta \oint_{F_\delta = t} M\xi_{12} + O(\delta^3)\\ &= -\oint_{F_\delta = t} M\left(\Omega + \delta\xi_{12} + \delta^2\xi_0\right) + \delta \oint_{F_\delta = t} M\Omega + \delta^2 \oint_{F_\delta = t} M\xi_0 + O(\delta^3)\\ &= \delta^2 \oint_{F_\delta = t} M\xi_0 + O(\delta^3) = \delta^2 \oint_{F = t} M\xi_0 + O(\delta^3).\end{aligned}$$

Acknowledgements This work was supported by the Polish OPUS Grant No 2012/05/B/ST1/03195.

References

1. Arnold, V.I., Varchenko, A.N., Gusein-Zade, S.M.: Singularities of differential mappings. In: Monographs in Mathematics, vol. 83. Birkhäuser, Boston (1988) [Russian: v. 2, Nauka, Moscow, 1984]
2. Christopher, C.: Estimating limit cycle bifurcations from centers. In: Differential Equations with Symbolic Computations. Trends in Math, pp. 23–35. Birkhäuser, Boston (2005)
3. Christopher, C., Schlomiuk, D.: On general algebraic mechanisms for producing centers in polynomial differential systems. J. Fixed Point Theory Appl. **3**, 331–351 (2008)

4. Dumortier, F., Roussarie, R., Rousseau, C.: Hilbert's 16th problem for quadratic vector fields. J. Diff. Equ. **110**, 86–133 (1994)
5. Françoise, J.-P.: Successive derivatives of a first return map, application to the study of quadratic vector fields. Ergod. Theory Dyn. Syst. **16**, 87–96 (1996)
6. Fronville, A., Sadovskii, A., Żołądek, H.: Solution of the 1:-2 resonant center problem in the quadratic case. Fundam. Math. **157**, 191–207 (1998)
7. Graf v. Bothmer, H.-C.: Experimental results for the Poincaré center problem. Nonlinear Differ. Equ. Appl. **14**, 671–698 (2007)
8. Graf v. Bothmer, H.-C., Kroker, J.: Focal values of plane cubic centers. Qual. Theory Dyn. Syst. **9**, 319–324 (2010)
9. Yang, J., Han, M., Li, J., Yu, P.: Existence conditions of thirteen limit cycles in a cubic system. Int. J. Bifurcation Chaos **20**, 2569–2577 (2010)
10. Yu, P., Han, M.: A study on Zoladek's example. J. Appl. Anal. Comput. **1**, 143–153 (2011)
11. Żołądek, H.: Quadratic systems with center and their perturbations. J. Differ. Equ. **109**, 223–273 (1994)
12. Żołądek, H.: The classification of reversible cubic systems with center. Topol. Methods Nonlinear Anal. **4**, 79–136 (1994)
13. Żołądek, H.: Eleven small limit cycles in a cubic vector field. Nonlinearity **8**, 843–860 (1995)
14. Żołądek, H.: Remarks on: The classification of reversible cubic systems with center. Topol. Methods Nonlinear Anal. **8**, 335–342 (1996)
15. Żołądek, H.: The monodromy group. Monografie Matematyczne, vol. 67. Birkhäuser, Basel (2006)

Index

B. Toni (ed.), *Mathematical Sciences with Multidisciplinary Applications*, Springer Proceedings in Mathematics & Statistics 157, DOI 10.1007/978-3-319-31323-8