Springer Proceedings in Mathematics & Statistics

Andreas Johann
Hans-Peter Kruse
Florian Rupp
Stephan Schmitz *Editors*

Recent Trends in Dynamical Systems

Proceedings of a Conference in Honor of Jürgen Scheurle

Springer

Springer Proceedings in Mathematics and Statistics

Volume 35

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics and Statistics

This book series features volumes composed of selected 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.

Andreas Johann • Hans-Peter Kruse
Florian Rupp • Stephan Schmitz
Editors

Recent Trends in Dynamical Systems

Proceedings of a Conference in Honor of Jürgen Scheurle

Springer

Editors
Andreas Johann
Hans-Peter Kruse
Florian Rupp
Stephan Schmitz
Zentrum Mathematik
Technische Universität München
Garching bei München
Germany

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-0348-0450-9 ISBN 978-3-0348-0451-6 (eBook)
DOI 10.1007/978-3-0348-0451-6
Springer Basel Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013947464

Mathematical Subject Classification (2010): 34-XX, 35-XX, 37-XX, 70-XX, 74-XX, 76-XX, 82-XX, 93-XX

Printed on acid-free paper

Springer Basel is part of Springer Science+Business Media (www.springer.com)

To Jürgen Scheurle

Preface

In January 2012 the International Conference *Recent Trends in Dynamical Systems* was held in Munich on the occasion of Jürgen Scheurle's 60th birthday. As parts of this conference, a scientific colloquium took place at the Carl Friedrich von Siemens Stiftung in Munich from 11th to 13th of January and also a Festkolloquium at the Technische Universität München in the afternoon of January 13th. Besides numerous posters on recent advances in the field of dynamical systems, 25 highly recognized scholars gave plenary talks that were grouped according to the following themes:

- Stability and bifurcation
- Geometric mechanics and control theory
- Invariant manifolds, attractors, and chaos
- Fluid mechanics and elasticity
- Perturbations and multiscale problems
- Hamiltonian dynamics and KAM theory

These themes reflect the broad scientific interests of Jürgen Scheurle and his fascination of applying mathematics to real world situations, in particular from physics and mechanics. The volume at hand is an outgrowth of this conference, containing research articles about exciting new developments in the multifaceted subject of dynamical systems as well as survey articles. We are very happy that the authors accepted the invitation to contribute to this volume in honour of Jürgen Scheurle and we are sure that their exciting articles will be of interest not only to experts in the field of dynamical systems but also to graduate students and scientists from many other fields, including engineering. This is in the spirit of Jürgen Scheurle, who, besides his research activities, always puts a lot of emphasis on conveying the beauty of the Theory of Dynamical Systems and its applicability to real world problems in extremely well-prepared, beautiful lectures.

Munich, Germany
January 2013

Andreas Johann
Hans-Peter Kruse
Florian Rupp
Stephan Schmitz

Short Curriculum Vitae of Jürgen Scheurle

Jürgen Scheurle was born on September 26, 1951, in Schwäbisch Gmünd, Baden-Württemberg. He received his professional education at the University of Stuttgart, where he studied mathematics, physics, and computer science from 1970 until 1974, and finished his diploma degree in mathematics with a thesis entitled "Ein Antikonvergenzprinzip". Some months later, in 1975, he completed his doctorate under the guidance of Klaus Kirchgässner. The title of his Ph.D. thesis is "Ein selektives Iterationsverfahren und Verzweigungsprobleme". In 1981 he presented his Habilitation thesis on "Verzweigung quasiperiodischer Lösungen bei reversiblen dynamischen Systemen".

From 1974 to 1985 Jürgen Scheurle held positions as a postdoctoral researcher, senior researcher, and assistant professor, at the University of Stuttgart. In 1982 he was visiting professor at the Department of Mathematics, University of California, Berkeley (USA), and in 1983 at the Division of Applied Mathematics, Brown University, Providence (USA). In 1985 Jürgen Scheurle moved to Fort Collins (USA), where he became an associate and later full professor at Colorado State University. In 1987 he accepted a full professorship and the Chair of Theory and Applications of Partial Differential Equations at the University of Hamburg. In 1996 Jürgen Scheurle was appointed full professor at the Technische Universität München (TUM) and since then holds the Chair of Advanced Mathematics and Analytical Mechanics. Notable predecessors at this chair were Felix Klein, Walter von Dyck, and Robert Sauer, see Fig. 1, which illustrates the special responsibility of Jürgen Scheurle for the mathematical education of engineering students.

He was the founding director of the Center for Mathematics at TUM and later dean of the Faculty of Mathematics. As dean, he continued the reform-oriented politics of his predecessors. During his term in office, the faculty voluntarily conducted a peer assessment and was awarded the title "Reformfakultät" by the "Stifterverband der Deutschen Wissenschaft". Such assessments are common nowadays but were completely novel 10 years ago. Moreover, far ahead before such procedures were put into law, the Bavarian Ministry of Research and Teaching allowed the faculty to introduce an "Experimentierklausel" to assess prospective for the admission of students.

Jürgen Scheurle was responsible for the introduction of the "Master of Science in Industrial & Financial Mathematics" at the off-shore campus of TUM in Singapore. He was a member of the planning team for the new mathematics building at the research campus Garching and in charge of the relocation from downtown Munich to Garching in 2002. Finally, Jürgen Scheurle was and is member of numerous expert committees appointed by the president of the TUM and the faculty of mathematics. Inter alia he is representative of the "Bayerische Eliteakademie", member of the "Hurwitz-Gesellschaft zur Förderung der Mathematik an der TU München" and its president since 2011.

Jürgen Scheurle authored and co-authored several pioneering publications, and among them the following are highly influential articles:

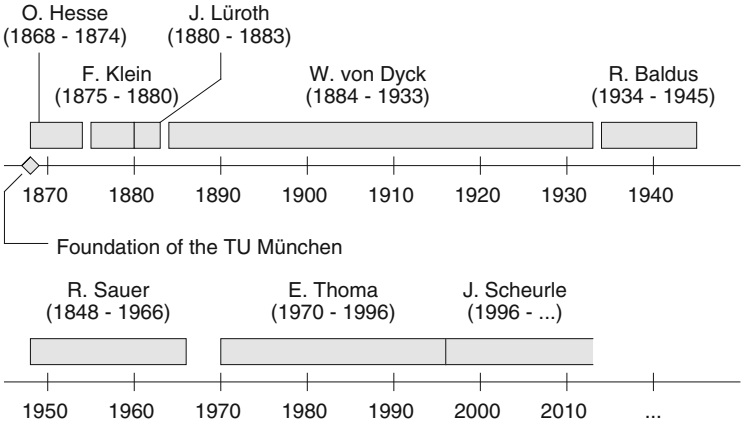

Fig. 1 Genealogy of the chair "Analytische Mechanik und Angewandte Mathematik" at the Technische Universität München

- *On the bounded solutions of a semilinear elliptic equation in a strip* (together with K. Kirchgässner). J. Diff. Equat. 32 (1) (1979), 119–148.
- *Smoothness of bounded solutions of non-linear evolution equations* (together with J. Hale). J. Diff. Equat. 56 (1) (1985), 142–163.
- *Chaotic solutions of systems with almost periodic forcing*. ZAMP 37 (1986), 12–26.
- *The construction and smoothness of invariant manifolds by the deformation method* (together with J. Marsden). SIAM J. Math. Anal. 18 (5) (1987), 1261–1274.
- *Exponentially small splittings of separatrices in KAM theory and degenerate bifurcations* (together with P. Holmes and J. Marsden). Cont. Math. 81 (1988), 213–243.
- *Existence of perturbed solitary wave solutions to a model equation for water waves* (together with J. Hunter). Physica D 32 (1988), 253–268.
- *Lagrangian reduction and bifurcations of relative equilibria of the double spherical pendulum* (together with J. Marsden). ZAMP 44 (1993), 17 - 43.
- *The reduced Euler-Lagrange equations* (together with J. Marsden). Fields Inst. Comm. 1 (1993), 139–164.
- *Pattern evocation and geometric phases in mechanical systems with symmetry* (together with J. Marsden), Dyn. and Stab. of Systems 10 (1995), 315–338.
- *Discretization of homoclinic orbits and "invisible" chaos* (together with B. Fiedler). Memoirs of the AMS vol. 119, nb. 570 (3), Providence 1996.
- *Reduction Theory and the Lagrange-Routh equations* (together with J. Marsden and T. Ratiu). J. Math. Phys. 41(6) (2000), 3379–3429.
- *The orbit space method* (together with M. Rumberger). In Ergodic Theory, Analysis and Efficient Simulation of Dynamical Systems, B. Fiedler edt., Springer-Verlag 2001, 649–689.

- *On the generation of conjugate flanks for arbitrary gear geometries* (together with A. Johann). GAMM-Mitt. 32, No. 1, 2009, 61–79.

His teaching covers a wide spectrum of subjects, ranging from mathematics for engineering students, functional analysis, ordinary differential equations and partial differential equations to dynamical systems, bifurcation theory, hamiltonian dynamics, geometric mechanics, mathematical methods in continuum mechanics, and mathematical modeling in biology and ecology. He supervised more than 20 dissertations and habilitations in these areas.

Jürgen Scheurle was a member of the advisory board of the book series *Dynamics Reported* and an executive editor of the *International Journal of Nonlinear Mechanics*. He is currently a member of the editorial board of the *Journal of Nonlinear Science*, *Nonlinear Science Today*, *Journal of Applied Mathematics and Mechanics* (*ZAMM*), and *Journal of Geometric Mechanics*.

Conference photo in the garden of the Carl Friedrich von Siemens Stiftung at the Schloß Nymphenburg, Munich

Registered Participants in Alphabetic Order

- Wolf-Jürgen Beyn
- Anthony Bloch
- Jörg-Stefan Bock
- Henk W. Broer
- Tomas Caraballo
- David Chillingworth
- Florin Diacu
- Michael Dellnitz
- Jochen Denzler
- Freddy Dumortier
- Dominik Eberlein
- Francesco Fasso
- Peter Giesl
- Christoph Glocker
- John Guckenheimer
- Thomas Hagen
- Heinz Hanßmann
- Karl-Heinz Hoffmann
- Phillip Huber
- Delia Ionescu-Kruse
- Gerard Iooss
- Andreas Johann
- Christopher K.R.T. Jones
- Oliver Junge
- Hansjörg Kielhöfer
- Peter E. Kloeden
- Thorsten Knott
- Peter Koltai
- Carl Friedrich Kreiner
- P.S. Krishnaprasad
- Hans-Peter Kruse
- Tassilo Küpper
- Christian Kühn
- Rainer Lauterbach
- Martin Lehl
- Armin Leutbecher
- Daniel Matthes
- Johannes Mayet
- Alexander Mielke
- James Montaldi
- Horst Osberger
- Kathrin Padberg-Gehle
- Tudor Ratiu
- Geneviève Raugel
- Sebastian Reich
- Michael Renardy
- Mark Roberts
- Marcello Romano
- Matthias Rumberger
- Florian Rupp
- Johannes Rutzmoser
- Björn Sandstede
- Jürgen Scheurle
- Thorsten Schindler
- Günter Schlichting
- Guido Schneider
- Stephan Schmitz
- Svenja Schoeder
- Andreas Schuppert
- Rüdier Seydl
- Andre Vanderbauwhede
- Sebastian Walcher
- Bodo Werner
- Johannes Zimmer

Contents

List of Contributors

Fred C. Adams Department of Physics, University of Michigan, Ann Arbor, MI, USA

Anthony M. Bloch Department of Mathematics, University of Michigan, Ann Arbor, MI, USA

Henk W. Broer Instituut voor Wiskunde en, Informatica, Rijksuniversiteit Groningen, AG Groningen, The Netherlands

T. Caraballo Dpto. Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla, Sevilla, Spain

D.R.J. Chillingworth Department of Mathematics, University of Southampton, Southampton, UK

Freddy Dumortier Hasselt University, Diepenbeek, Belgium

Messoud Efendiev Helmholtz Center Munich, Institute of Biomathematics and Biometry, Neuherberg Germany

Tor Flå Department of Mathematics and Statistics, University of Tromso, Tromso, Norway

Peter A. Giesl Department of Mathematics, University of Sussex, Falmer, UK

Christoph Glocker IMES - Center of Mechanics, ETH Zurich, Zurich, Switzerland

Alexandra Goeke Lehrstuhl A für Mathematik, RWTH Aachen, Aachen, Germany

John Guckenheimer Mathematics Department, Cornell University, Ithaca, NY, USA

Sigurdur Hafstein School of Science and Engineering, Reykjavik University, Reykjavik, Iceland

Heinz Hanßmann Mathematisch Instituut, Universiteit Utrecht, TA Utrecht, The Netherlands

D.D. Holm Mathematics Department, Imperial College London, London, UK

H.A. Hosham Mathematical Institute, University of Cologne, Cologne, Germany

Eric W. Justh Naval Research Laboratory, Washington, DC, USA

Peter E. Kloeden Institut für Mathematik, Goethe-Universität, Frankfurt am Main, Germany

P.S. Krishnaprasad Institute for Systems Research and Department of Electrical and Computer Engineering, University of Maryland, College Park, MD, USA

M. Kružík Institute of Information Theory and Automation of the ASCR, Prague, Czech Republic Faculty of Civil Engineering, Czech Technical University, Prague, Czech Republic

Tassilo Küpper Mathematical Institute, University of Cologne, Cologne, Germany

Jeffrey C. Lagarias Department of Mathematics, University of Michigan, Ann Arbor, MI, USA

Alexander Mielke Weierstraß-Institut für Angewandte Analysis und Stochastik, Berlin, Germany

James Montaldi School of Mathematics, University of Manchester, Manchester, UK

Philip J. Morrison Department of Physics and Institute for Fusion Studies, University of Texas, Austin, TX, USA

J. Munn Eltham College, London, UK

A.B. Nordmark Department of Mechanics, KTH, Stockholm, Sweden

Marius Paicu Univ. Bordeaux, IMB, UMR 5251, Talence, France

Tudor S. Ratiu Department of Mathematics and Bernoulli Center, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Geneviève Raugel CNRS, Laboratoire de Mathématiques d'Orsay, Orsay Cedex, France Univ Paris-Sud, Orsay Cedex, France

Sebastian Reich Universität Potsdam, Institut für Mathematik, Potsdam, Germany

Michael Renardy Department of Mathematics, Virginia Tech, Blacksburg, VA, USA

Florian H.-H. Rupp Lehrstuhl für Höhere Mathematik und Analytische Mechanik, Technische Universität München, Fakultät für Mathematik, Garching, Germany

Björn Sandstede Division of Applied Mathematics, Brown University, Providence, RI, USA

Samuel N. Stechmann Mathematics Department, University of Wisconsin, Madison, WI, USA

Tadashi Tokieda Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, Trinity Hall, Cambridge, UK

André Vanderbauwhede Department of Pure Mathematics, Ghent University, Gent, Belgium

Sebastian Walcher Lehrstuhl A für Mathematik, RWTH Aachen, Aachen, Germany

D. Weiss Mathematical Institute, University of Tübingen, Tübingen, Germany

Clemens Woywod Department of Chemistry, Center for Theoretical and Computational Chemistry (CTCC), University of Tromso, Tromso, Norway

Jiangong You Department of Mathematics, Nanjing University, Nanjing, PR China

Fumin Zhang School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

Anna Zhigun Helmholtz Center Munich, Institute of Biomathematics and Biometry, Neuherberg, Germany

Johannes Zimmer Department of Mathematical Sciences, University of Bath, Bath, UK

Part I
Stability, Bifurcation and Perturbations

Chapter 1
The Birth of Chaos

John Guckenheimer

1.1 Introduction

The word ***chaos*** has become firmly embedded in the literature on dynamical systems. Indeed, James Gleick's book, Chaos Theory [17], established that term as a description of the entire subject in the public mind. Nonetheless, there is no authoritative technical meaning of "chaos" in dynamical systems. Li and Yorke first used the word in the title of their paper "Period three implies chaos" [31], but it does not appear in the text. They refer to trajectories that are "nonperiodic and might be called 'chaotic'." Ruelle and Takens [45] used the longer phrase "sensitive dependence to initial conditions" and the two terms have largely been regarded as synonyms [15, 57]. The informal definition of sensitive dependence to initial conditions is that nearby initial conditions separate; the technical definition is that there are sets of trajectories with positive Lyapunov exponents [15] that measure the exponential rate of separation of nearby trajectories. What is not often specified in the definition is ***how many*** trajectories have positive Lyapunov exponents. For example, if a dynamical system has a saddle point, this point has a positive Lyapunov exponent, but the presence of a single saddle point (or even more complicated normally hyperbolic sets) does not make the system chaotic. There appears to be little consensus on the minimal requirements for sets of trajectories with positive Lyapunov exponents that make a system chaotic, but there is a sufficient criterion formulated by Smale [48] that is often used as a practical test: namely, that the system possesses a "transversal intersection of stable and unstable

J. Guckenheimer (✉)
Mathematics Department, Cornell University, Ithaca, NY 14853, USA
e-mail: jmg16@cornell.edu

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_1,

manifolds of a periodic orbit." This concept is explained below. Such ***homoclinic*** orbits were first discovered by Poincaré in 1890 in a prize winning essay [43] motivated by the question, is the solar system stable? The intriguing history of Poincaré's discovery has been studied and recounted by Barrow-Green [3]. The work of Poincaré and later Birkhoff was directed at ***conservative*** dynamical systems arising in celestial mechanics. Within the setting of systems that preserve a symplectic structure, they investigated the presence of transversal intersections of the stable and unstable manifolds of periodic orbits. The first mathematical analysis of transversal homoclinic orbits in the context of dissipative systems that are not conservative was carried out by Cartwright and Littlewood, beginning during World War II [9–12] and culminating in Littlewood's long two part paper of 1957 [32–34]. The personal aspects of the Cartwright–Littlewood collaboration are also fascinating and have been described by McMurran and Tattersall [37,38] as well as by Cartwright herself [8,50].

The initial presentations of significant mathematical discoveries seldom appear in full clarity. The path to a new discovery is often tortuous, so reformulation is typically needed to distill the essence of new insights. This has been true in dynamical systems theory: the papers of Poincaré and Littlewood cited above are excellent examples. The work of Cartwright–Littlewood has a dual character, containing detailed analysis of the forced Van der Pol differential equation as well as a description of the dynamical consequences of transversal homoclinic orbits in dissipative systems. There was a long period of abstraction and simplification of the arguments of Cartwright–Littlewood that led to piecewise linear vector fields studied by Levinson [30] and later Levi [29], the geometric discrete time Smale horseshoe [47, 49] and the concept of ***hyperbolic invariant sets*** [46]. Figure 1.1 illustrates the horseshoe. These developments provided tremendous insight into chaotic dynamics, but they draw upon only a small portion of the Cartwright–Littlewood analysis of the forced Van der Pol differential equation. Thus, there is a disparity between mathematical awareness of these two aspects of the Cartwright–Littlewood discovery of chaos in dissipative systems. The horseshoe and its symbolic dynamics are a beautiful geometric example of chaotic dynamics, simple enough to be included routinely in undergraduate courses. Littlewood's analysis of the forced Van der Pol equation remains obscure despite its central role in the book of Grasman [18]. This paper visualizes horseshoes in the forced Van der Pol equation from the perspective of ***geometric singular perturbation theory*** and describes recent extensions of the work of Cartwright–Littlewood by myself and collaborators [6,19,20] that culminated in the thesis of Radu Haiduc [22,23]. Haiduc proved that there are parameter values for which the forced Van der Pol equation is structurally stable and possesses a chaotic invariant set. This paper gives an extended outline of this work, presenting the key geometric constructions used in the analysis of the forced Van der Pol equation.

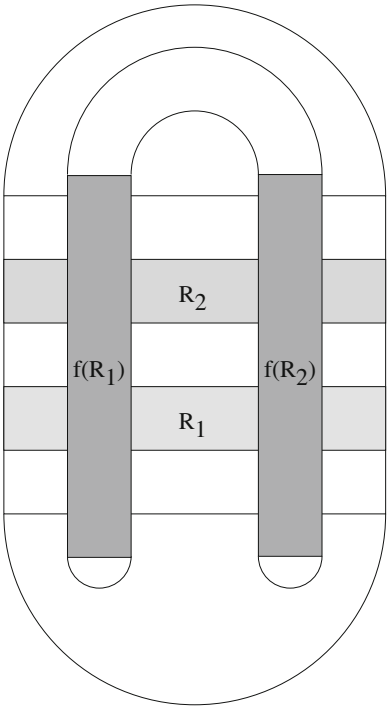

Fig. 1.1 The horseshoe is an invariant set Λ of the discrete map f depicted in this figure. The map f stretches the background oval vertically, compresses it horizontally, and maps it back into itself. The rectangles R_1 and R_2 shaded in *light gray* are mapped rectilinearly into their images shaded in *dark gray*. Inside the intersection $(R_1 \cup R_2) \cap (f(R_1) \cup f(R_2))$, there is an invariant Cantor set Λ consisting of points whose f-trajectories (both forward and backward) remain inside the intersection. The vertical distance between points that lie on different horizontal lines increases until one of the points lands in R_1 at the same time that the other lands in R_2. This expresses the ***sensitivity to initial conditions*** of this map. There is a one-to-one correspondence between points of Λ and bi-infinite sequences of 1 and 2 that encode which rectangle R_j each iterate lies in

1.2 The Forced Van der Pol Equation

The main object of this paper is analysis of the system of differential equations

$$\begin{aligned} \varepsilon \dot{x} &= y + x - \frac{x^3}{3} \\ \dot{y} &= -x + a \sin(2\pi\theta) \\ \dot{\theta} &= \omega \end{aligned} \tag{1.1}$$

where the variable $\theta \in S^1 = \mathbb{R}/\mathbb{Z}$, so we identify θ and $\theta + 1$. We are interested in the parameter regime where $\varepsilon > 0$ is small. The limit $\varepsilon = 0$ produces a system

of ***differential algebraic equations*** that plays a central role in our investigation of the equation. The methods that we use to study this limit come from geometric singular perturbation theory[16, 27] and are reviewed below. In different coordinates, this system is the one studied by Balthasar van der Pol in the 1920s as a model for the dynamics of an electronic vacuum tube subject to a sinusoidal input. In addition to his qualitative analysis of the solutions to this system, Van der Pol built electronic circuits that he studied experimentally. The report of Van der Pol and Van der Mark [53] that they heard noisy output when the output of such a circuit drove a loudspeaker has been long recognized as the first observation of chaos in a physical system [7]. While the observations of Van der Pol and Van der Mark almost certainly did not come from the same chaotic oscillations found by Cartwright and Littlewood in Eq. (1.1), they served as an inspiration for Cartwright and Littlewood.

There are two very different ways of reducing the forced Van der Pol equation to a two-dimensional dynamical system. The first consists of choosing a cross section by fixing a value of θ and examining the return map of this section. Since $\dot{\theta}$ is constant, the section is global; indeed the return map is obtained by flowing for time $1/\omega$. This is the customary perspective of dynamical systems theory, but it is not well adapted to identification of the features that give rise to chaos in the system. Moreover, numerical computation of the return map is problematic because there is a repelling slow manifold along which trajectories separate so rapidly that the return map appears to be discontinuous. Filling the gaps associated with these discontinuities required innovative methods.

Littlewood adopted a different approach in studying the system. His approach is related to investigation of the singular limit $\varepsilon = 0$. Figure 1.2 shows two trajectories of the forced Van der Pol equation when $\varepsilon = 10^{-4}$. The cubic surface S defined by $y + x - \frac{x^3}{3} = 0$ is the critical manifold of the system. Away from S, the flow of the system is almost parallel to the x axis. The projection of S onto the (y, θ) cylinder along the x axis is singular along the two circles defined by $(x, y) = (\pm 1, \mp 2/3)$. These circles are called fold curves. They separate S into three sheets. The outer sheets given by $|x| > 1$ are attracting while the inner sheet given by $|x| < 1$ is repelling. Trajectories are drawn quickly to an $O(\varepsilon)$ neighborhood of the outer sheets, then flow along the sheets until they reach the fold curves. As described in more detail below, they flow very rapidly from the fold curve to the opposite sheet of S where they turn and proceed further along S. Perhaps the most subtle part of the dynamics involves trajectories that approach the fold curves almost tangentially. When $a > 1$, some trajectories approach the fold curve, but then turn back along the attracting slow manifold instead of jumping to the opposite sheet. Littlewood used the terminology "dips" and "slices" to distinguish trajectories that turn back from trajectories that jump. The gap between these regions contains trajectories, called ***canards***, that follow the repelling slow manifold for a substantial time and hence distance before jumping or turning back. Figure 1.3 shows two unstable periodic orbits that each contain a pair of canard segments. Nearby trajectories with canards separate from each other, giving rise to the stretching requisite for the formation of horseshoes and transversal homoclinic orbits. This separation is abrupt when ε is small and evident in Fig. 1.3. Singular perturbation theory identifies ***folded***

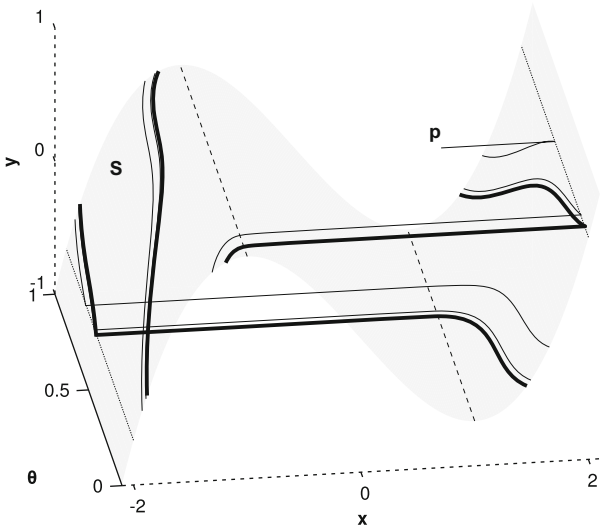

Fig. 1.2 The flow of the forced Van der Pol equation with parameters $(a, \omega, \varepsilon) = (1.1, 1.57, 0.0001)$. The cylindrical phase space is cut along the surface $\theta = 0$, so pairs of points (x, y, θ) with $\theta = 0$ and $\theta = 1$ are to be identified with one another. The critical manifold S is displayed as a *light gray surface* and its fold curves are drawn as *thin dashed lines*. The *dotted lines* are the lines $(x, y) = (\pm 2, \pm 2/3)$ on the critical manifold at the same height as the fold curves. Two trajectories are shown. A stable periodic orbit of period 3 is drawn as a *heavy line*. Starting at $\theta = 0, x < 0$, this trajectory moves up S, reaching $\theta = 1$ with x close to -1. It then crosses the fold curve and jumps almost parallel to the x-axis to the sheet of S with $x > 1$. It turns abruptly and follows this sheet with decreasing y until it reaches the fold curve with $x = 1$. During this portion of the trajectory, θ advances by approximately 1.5. At the fold with $x = 1$, the trajectory jumps back to the sheet of S with $x < -1$, turns abruptly, and returns to its starting position. The second trajectory is drawn as a *thin curve*. It starts at $(x, y, \theta) = (1, 2/3, 0.5)$, marked as the point p. This trajectory approaches the periodic orbit as one follows it forward

singularities, isolated points on the fold curve of the critical manifold that lie at the heart of these dynamics. Analysis of the dynamics of one type of folded singularity, the folded saddle, and its associated canards is the crux of establishing the existence of chaos in the forced Van der Pol equation. The next section describes in more detail the components of geometric singular perturbation theory that are used in analyzing these dynamical features.

1.3 Background on Slow–Fast Dynamical Systems

The forced Van der Pol equation is an example of a ***slow–fast*** dynamical system of the form

$$\begin{aligned} \varepsilon \dot{x} &= f(x, y) \\ \dot{y} &= g(x, y) \end{aligned} \tag{1.2}$$

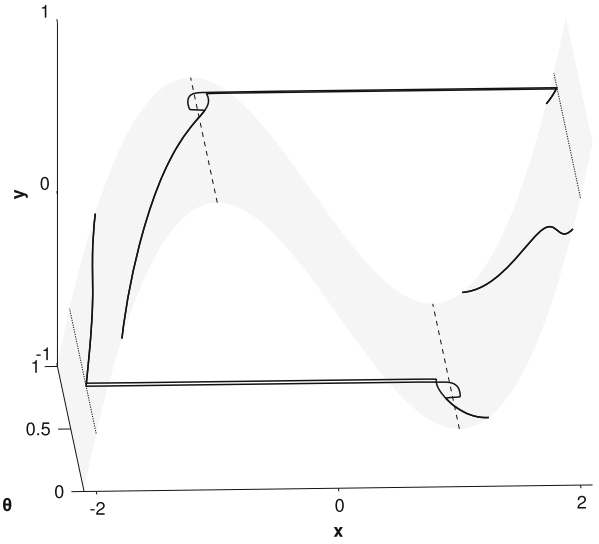

Fig. 1.3 Two periodic orbits of the forced Van der Pol equation containing canards. Parameters are $(a, \omega, \varepsilon) = (1.1, 1.57, 0.0001)$. Over most of their length, these trajectories are very close to one another. After the orbits cross the fold curves at $x = \pm 1$, they remain close to S instead of immediately jumping to the opposite sheet of S. Here they separate as one of the two orbits jumps back to the sheet it was on, while at a later time (i.e., larger θ), the second orbit jumps across to the opposite sheet of S. Both periodic orbits are invariant under the symmetry $(x, y, \theta) \mapsto (-x, -y, \theta + 0.5)$

In our case, $x \in \mathbb{R}$ is the fast variable, $y \in \mathbb{R}^2$ is the slow variable, and ε is a small parameter that represents the ratio of time scales.

The Van der Pol equation with constant forcing a

$$\begin{aligned} \varepsilon \dot{x} &= y - \frac{x^3}{3} + x \\ \dot{y} &= a - x \end{aligned} \tag{1.3}$$

is one of the first slow–fast systems to have been studied mathematically [51, 52], and it has served as a key example in the development of singular perturbation theory for dynamical systems [18]. Figure 1.4 shows a phase portrait of a periodic orbit for this system when $a = 0$ together with the curve S defined by $y - \frac{x^3}{3} + x = 0$ and a pair of horizontal segments. Note that the horizontal segments and two segments of the curve S comprise a good approximation to the periodic orbit. This is apparent from a geometric description of the flow of system (1.3). When $\varepsilon > 0$ is small, trajectories of system (1.3) are almost horizontal outside a neighborhood U of S. The curve S is partitioned into three "branches" S_l, S_m, S_r by the fold points at $x = \mp 1$. Trajectories outside U flow toward S_l and S_r and away from S_m. Trajectories inside U near S_r flow down because $\dot{y} = -x > 0$ and they

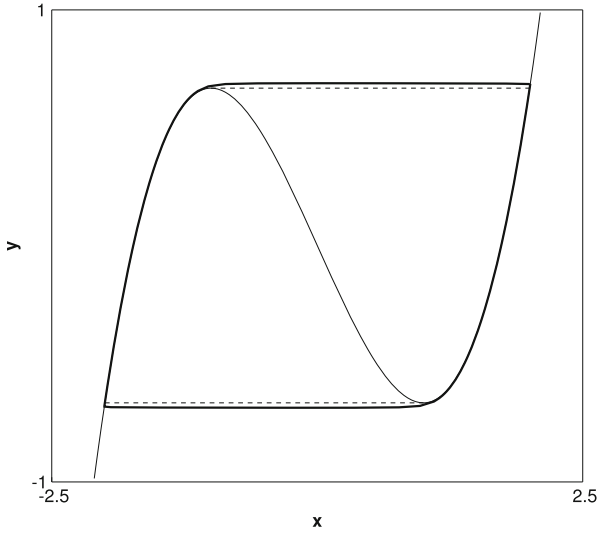

Fig. 1.4 The periodic orbit of the Van der Pol equation with parameter values $(a, \varepsilon) = (0, 0.001)$ is drawn as a *thick curve*. The critical manifold $y = \frac{x^3}{3} - x$ is drawn as a *thin curve*. The singular cycle of the system consists of two segments of the critical manifold together with horizontal segments drawn as *dashed curves* that join the fold points $(\pm 1, \mp 2/3)$ to the opposite branch of the critical manifold. Motion along the periodic orbit is clockwise

flow up inside U near S_l. A key aspect of the dynamics is what happens near the ***fold*** points $p_\pm = (\pm 1, \mp 2/3)$. The vertical component of trajectories near p_+ is negative, and below S the horizontal component points is also negative. Consequently, trajectories near p_+ are carried into the region where the flow is almost horizontal and approaches S_l. Similarly, trajectories near p_- are carried into the region where the flow is almost horizontal and approaches S_r. In numerical simulations with $0 < \varepsilon < 10^{-3}$, the mutual attraction of trajectories near S_r and S_l is so strong that they all appear to merge with one another in less than a single cycle around the periodic orbit. Conceptually, one can think of reducing the system to slow manifolds and fast ***jumps*** that occur when an attracting slow manifold ends at a fold point. Geometric singular perturbation theory provides rigorous foundations for this mathematical intuition [14, 28, 30].

The equation (1.2) no longer defines a system of ordinary differential equations (ODEs) when $\varepsilon = 0$. It gives a system of ***differential algebraic equations*** (DAEs). Unlike ODEs, DAEs may have initial conditions for which solutions do not exist. This happens at the fold points $p_\pm$ of the Van der Pol equation. At p_+, the equations demand that $\dot{y} = -1 < 0$, but y has a local minimum along S at p_+. The solutions of the equation $f = 0$ in (1.2) constitute the critical manifold S of the system. Denote the derivatives of f with respect to x by f_x, etc. If f_x is a regular (nonsingular) matrix at $z = (x, y) \in S$, the implicit function theorem implies that there is a neighborhood U of z with the property that $S \cap U$ is the graph

of a function $x = h(y)$. The DAE then reduces to the ODE $\dot{y} = g(h(y), y)$ which does have solutions. This ODE is often called the ***reduced system*** or the ***slow flow*** of the system (1.2). Convergence of solutions of the "full" system (1.2) to those of the reduced system is a fundamental topic in the theory of slow–fast systems. The results depend upon the ***fast subsystems*** or ***layer equations*** $\dot{x} = f(x)$. Time can be rescaled in Eq. (1.2) to obtain

$$\begin{aligned} x' &= f(x, y) \\ y' &= \varepsilon g(x, y) \end{aligned} \tag{1.4}$$

where the unit time scale is fast rather than slow. The limit $\varepsilon = 0$ of this system is a family of vector fields for x, with y acting as a parameter. The critical manifold S constitutes the set of equilibrium points of this system.

The portion of S where all the eigenvalues of f_x have negative real parts is attracting for the layer equations. Asymptotic methods summarized in the book of Mishchenko and Rozov [40] were used to investigate attractors for the full system near attracting sheets of the critical manifold. Geometric methods were used by Fenichel to study the existence of invariant slow manifolds near sheets of S that are ***normally hyperbolic***. The critical manifold S is normally hyperbolic at a point z if the matrix f_x has no eigenvalue whose real part is zero. Folds occur only at points where f_x has a zero eigenvalue, so normal hyperbolicity is a more stringent condition.

One of the technical difficulties encountered in the search for invariant slow manifolds is that they often do not exist for any $\varepsilon > 0$! This is apparent in the Van der Pol equation, where the critical manifold S is one dimensional. No trajectory remains slow in either forward or backward time. In forward time, trajectories approach an attracting branch of the critical manifold S but then go past a fold point where they become fast as they jump to the opposite branch of S. In backward time, trajectories either become unbounded in finite time or they remain close to the critical manifold where the magnitude $\dot{y} = -x$ is unbounded. Nonetheless, there are long sections of trajectories that remain close to S. Manifolds with boundary formed from these trajectory segments are said to be ***locally invariant***. They cannot be unique since bounded segments of trajectories vary continuously with initial data, but the fast dynamics forces locally invariant manifolds to be ***exponentially close***. This means that, away from their boundaries, the separation of locally invariant manifolds is bounded by a quantity of the form $c_1 \exp(-c_2/\varepsilon)$ for suitable $c_1, c_2 > 0$ that are independent of ε. With these difficulties in mind, Fenichel [16] proved

- The existence of locally invariant manifolds near compact regions of a normally hyperbolic critical manifold
- The convergence of the flow on these invariant manifolds to the slow flow on the critical manifold
- The existence of foliations of strong stable and unstable manifolds of the invariant manifolds

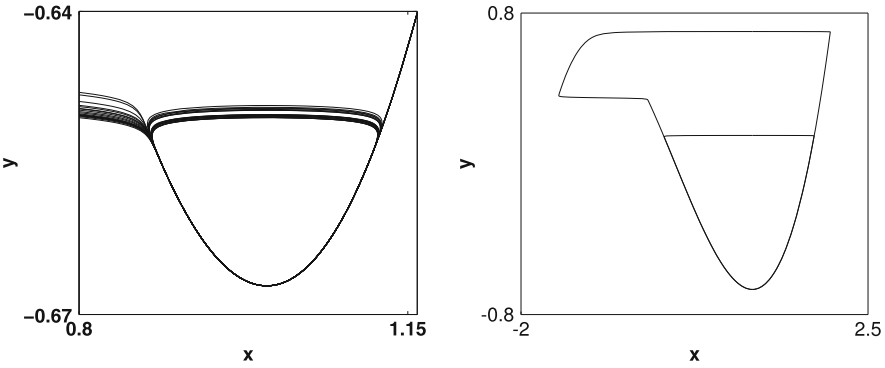

Fig. 1.5 The *left panel* displays a blown-up portion of a numerical trajectory of the Van der Pol equation with parameter values $(a, \varepsilon) = (-0.99987490608059, 0.001)$ and initial condition $(1, -2/3)$. The trajectory is able to follow the repelling branch of the critical manifold only a short distance before jumping right or left. The directions of the successive jumps are erratic, inconsistent with the Poincaré–Bendixson theorem. The *right panel* shows two canard orbits of the Van der Pol equation, computed at the same parameter values. These trajectories were computed by starting with initial conditions on the vertical lines $x = \pm 1$ and integrating in both forward and backward time to the line $x = 1$

The fibers of the stable foliation consist of points whose forward trajectories approach each other at a fast exponential rate until they flow past the boundary of the invariant manifold. Similarly, the fibers of the unstable foliation consist of points whose backward trajectories approach each other at a fast exponential rate. These results comprise Fenichel theory; they give a satisfying description of slow–fast systems in the vicinity of normally hyperbolic portions of their critical manifolds. However, as illustrated by the Van der Pol equation, they are not enough to yield a comprehensive understanding of slow–fast systems where the critical manifolds have folds.

Figure 1.5 illustrates the numerical difficulties of computing solutions with canards using the Van der Pol equation. Forward numerical integration is unable to follow trajectories along the repelling branch of the critical manifold due to its fast transverse instability. We set $\varepsilon = 0.001$ and select the parameter value $a = -0.99987490608059$ so that the attracting and repelling slow manifolds appear to connect near the fold point at $x = 1$. The left panel of the figure shows that simulation of a trajectory at these parameter values is unable to consistently detect the relative locations of the attracting and repelling slow manifolds. The trajectory returns repeatedly to the region of the fold and starts along the repelling manifold. However, it is able to follow this manifold only a short distance before jumping right or left, and the directions of successive jumps are erratic. The figure shows the trajectory only in a small region near the fold. This behavior is inconsistent with the Poincaré–Bendixson theorem, but has been observed with all numerical integration algorithms that have been tried, even with the most stringent error tolerances.

As discovered by a French group of mathematicians [13], periodic orbits of the system undergo a ***canard explosion*** in which they grow from size $O(\varepsilon^{1/2})$ to size $O(1)$ in an exponentially small range of the parameter a. The right panel of the figure shows two of these periodic orbits, one a "canard with head" and the second a "canard without head." The shape of the canard with head motivated the introduction of the term "canard" to denote a special class of periodic orbits, but conventional usage now is that the term refers to any trajectory that contains a segment which spends $O(1)$ (slow) time near a repelling slow manifold. In Fig. 1.5, the canard with head was computed by integrating the initial condition $(-1, 0.35)$ both forward and backward to its intersection with the vertical line $x = 1$ near the fold of the critical manifold. The canard without head was computed in the same way from initial condition $(1, 0.15)$. The terminal points of these four trajectories agree to 15 digits of precision. Thus, numerical integration with double precision floating point arithmetic is unable to resolve the separation of trajectories that arrive at the fold point of this system along its slow manifolds. The exponential convergence (divergence) of trajectories along the attracting (repelling) manifold is so extreme that trajectories starting over a very large region of initial conditions appear to arrive at the same point. Thus, we need to employ our theoretical understanding of the dynamics of canards to obtain qualitative correct simulations from numerical integration. This is also the case for the Van der Pol equation with periodic forcing, where the horseshoes we locate consist of trajectories containing canard segments.

1.4 Folds and Folded Saddles

Singularity theory [1, 2] provides mathematical tools that can be used to investigate slow–fast systems in the vicinity of their folds. The term fold has a technical meaning in singularity theory as the simplest singularity, and folds in this sense are the only type of singularity that appear in the Van der Pol equations with either constant or periodic forcing. We have already determined that the folds of the forced Van der Pol equation (1.1) are the circles defined by $(x, y) = (\pm 1, \mp 2/3)$. At these points, the inequalities that determine that the singularity is a simple fold are that $f_y \neq 0$ and $f_{xx} \neq 0$. The first inequality implies that the critical manifold is indeed a manifold, while the second implies that the tangency of this manifold with the surface of constant x is of the lowest possible order.

One of the principal strategies of singularity theory is to employ coordinate transformations that bring a system to a normal form. The normal forms characterize the extent to which any two systems with the same type of singularity can be transformed into one another by coordinate changes. In dynamical systems theory, the theory of normal forms is more complicated. At the outset, one seeks coordinate changes that simplify the Taylor expansions of a vector field at an equilibrium. A polynomial vector field P with an equilibrium at the origin is described as the ***truncated*** normal form for a class of systems if coordinate changes bring each vector field in the class to a system of the form $P+$ higher order terms.

In the singularity theory of maps, the goal is to transform (germs of) a map exactly to the truncated normal forms. For vector fields, this is seldom possible at degenerate equilibria. Arnold et al. [2] investigated normal forms in the context of slow–fast systems. For simple folds, we are primarily interested in whether the reduced system has trajectories that flow toward or away from the fold. To study this question, it is helpful to rescale or ***desingularize*** the reduced system.

Desingularization of the reduced system of the forced Van der Pol equation is a bit simpler than the general case, so we describe the construction in this specific case. The critical manifold of the forced Van der Pol equation is the graph of the function $y = \frac{x^3}{3} - x$. Differentiate this equation to obtain the relation $\dot{y} = (x^2 - 1)\dot{x}$ that holds on the critical manifold when $\varepsilon = 0$. Since $\dot{y}$ and $\dot{\theta}$ do not depend upon y, substituting the relation into the equation (1.1) yields

$$\begin{aligned} (x^2 - 1)\dot{x} &= -x + a\sin(2\pi\theta) \\ \dot{\theta} &= \omega \end{aligned} \tag{1.5}$$

eliminating y from the equations. However, $\dot{x}$ is infinite at the folds $x = \pm 1$. To desingularize the equation, we rescale time by $(x^2 - 1)$ to obtain the desingularized reduced system

$$\begin{aligned} \dot{x} &= -x + a\sin(2\pi\theta) \\ \dot{\theta} &= \omega(x^2 - 1) \end{aligned} \tag{1.6}$$

The time rescaling comes at a cost: it reverses orientation on the repelling sheet of the critical manifold where $x^2 - 1 < 0$. When relating trajectories of the system (1.6) to those of (1.1), it is crucial to take this orientation reversal into account.

Trajectories of the desingularized reduced system that cross the fold curve are ***not*** limits of trajectories of the full system as $\varepsilon \to 0$. The trajectories of the full system flow away from the fold almost parallel to the x direction of the system (1.2). In the case of the forced Van der Pol equation, these trajectories flow toward the opposite sheet of the critical manifold, where they once again become slow and follow paths close to the trajectories of the reduced system. Thus the limit behavior of the forced Van der Pol equation has ***jumps***: trajectories arriving at a fold are discontinuous, jumping from the point $(\pm 1, \mp 2/3, \theta)$ to the point $(\mp 2, \mp 2/3, \theta)$, which is the intersection of a line parallel to the x axis with the opposite sheet of the critical manifold. From this perspective, the reduced system is a discontinuous or ***hybrid*** dynamical system in which the maps from the fold curves to the critical manifold are discrete time components of the dynamics. For general slow–fast systems, singular perturbation theory describes asymptotic expansions in ε of the flow maps past folds. These expansions are singular, involving fractional powers of ε [18, 40].

Equilibrium points of the desingularized reduced system on its fold curve are called ***folded singularities***. Due to the rescaling factor of desingularization, folded singularities need not be limits of equilibria of the full system of equations (1.2)

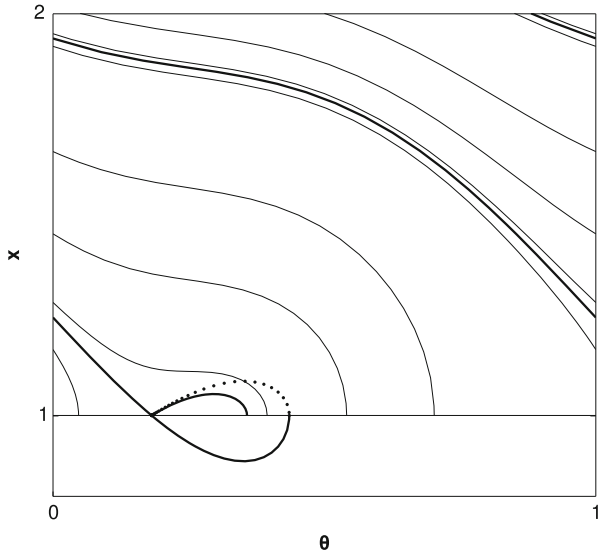

Fig. 1.6 This is a phase portrait of the desingularized reduced flow of the forced Van der Pol equation for parameters $(a, \omega) = (1.1, 1.57)$. All trajectories above $x = 1$ flow to the right. The stable manifold and one branch of the unstable manifold of the folded saddle are drawn as *heavy curves*. Trajectories starting at $x = 2$ to the left of the stable manifold flow directly to the fold curve $x = 1$ where they jump to $x = -2$ (not shown). Trajectories starting at $x = 2$ to the right of the stable manifold follow the unstable manifold of the folded saddle before reaching the fold curve $x = 1$. For the stable manifold itself, singular limits of the full system include trajectories that proceed along the branch of the stable manifold with $x < 1$ and then jump either to the region with $1 < x < 2$ or the region $-2 < x < -1$. The landing points of these jumps in the region $1 < x < 2$ are drawn as a *dotted curve*

as $\varepsilon \to 0$. Indeed, the forced Van der Pol equation has no equilibria at all because $\dot{\theta} = \omega \neq 0$, but it does have folded singularities when $a > 1$. These are located at the points where $x = \pm 1$ and $x = a \sin(2\pi\theta)$. The folded singularities divide the fold into segments, characterized by whether trajectories of the desingularized reduced system flow toward or away from the fold. The usual classification of equilibria of two-dimensional vector fields into nodes, foci, and saddles is applied here. A straightforward calculation shows that the region where folded nodes occur is very small, and that over most of the parameter space with $a > 1$, each fold curve has a folded saddle and a folded focus. Figure 1.6 shows a phase portrait of the desingularized reduced system for parameters $(a, \omega) = (1.1, 1.57)$.

Folded saddles are pivotal to our analysis of chaos in the forced Van der Pol equation. They were first studied by Benoit [5], who analyzed a (truncated) normal form. We describe the salient results from this analysis. We assume that S is the critical manifold of a system with two slow variables and one fast variable, that $L \subset S$ is a fold curve dividing S into an attracting sheet $S_{a,0}$ and a repelling sheet $S_{r,0}$ and that $p \in L$ is a folded saddle. Fenichel theory implies that there are smooth attracting and repelling slow manifolds $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ that are $O(\varepsilon)$ distant from

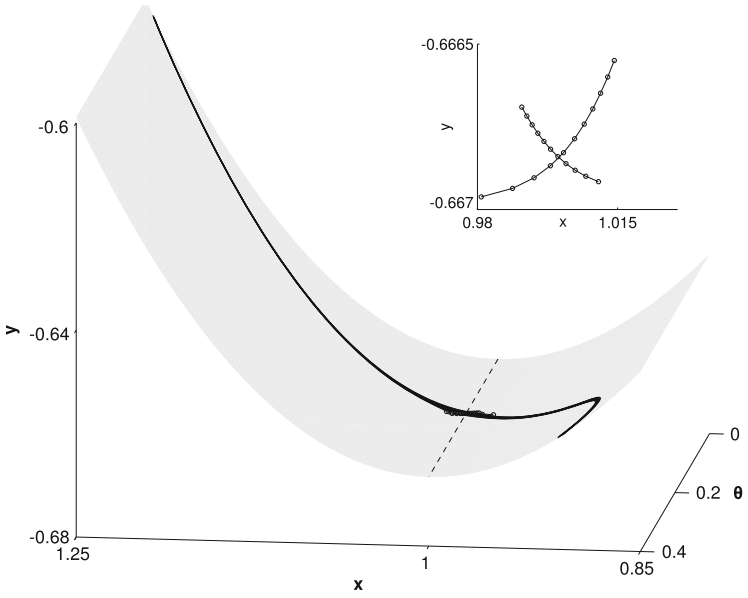

Fig. 1.7 Intersections of the attracting and repelling slow manifolds of the forced Van der Pol equation with parameters $(a, \omega, \varepsilon) = (1.1, 1.57, 0.0001)$. The gray surface is the critical manifold. The *black curves* are bundles of trajectories with initial conditions on the critical manifold and terminal conditions on the cross section $\theta = \theta_s$, the value of θ at the folded saddle of the reduced system. The *inset* shows the terminal points of the two bundles, illustrating that they intersect with an angle that is $O(\varepsilon)$. The intersection point is on the maximal canard of the system

$S_{a,0}$ and $S_{r,0}$ away from L. Extensions of these manifolds along trajectories are still invariant but no longer normally hyperbolic; the extensions show the fate of trajectories as they pass through a neighborhood of L. The first important result is that extensions of $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ intersect near p. The intersection is a trajectory γ, called a ***maximal canard***, that flows from the attracting slow manifold $S_{a,\varepsilon}$ to the repelling slow manifold $S_{r,\varepsilon}$. The second important result is that the extensions of $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$ intersect transversally along γ with an angle of intersection that is $O(\varepsilon)$. Together with the normal hyperbolicity of $S_{a,\varepsilon}$ and $S_{r,\varepsilon}$, this splitting angle is large enough to ensure that the flow along γ is hyperbolic. Trajectories on $S_{a,\varepsilon}$ to one side of γ jump along the fast direction while trajectories on the other side of γ turn away from the fold and flow back along $S_{a,\varepsilon}$ without jumping. Since the solutions of the system (1.2) depend continuously upon their initial conditions for $\varepsilon > 0$, the trajectories on $S_{a,\varepsilon}$ in the immediate vicinity of γ must interpolate between these behaviors. They do so by following γ for varying distances along $S_{r,\varepsilon}$ as canards and then jumping away from the fold ! or back to $S_{a,\varepsilon}$. The distance a trajectory travels before jumping is comparable to (minus) the logarithm of its distance from γ on $S_{a,\varepsilon}$. Consequently, this process stretches the distance between trajectories that start close to each other by a large amount. In the forced Van der Pol equation, the stretching that takes place in the horseshoes that we exhibit comes from the flow of these canards along $S_{r,\varepsilon}$ (Fig. 1.7).

1.5 Return Maps

Poincaré introduced cross sections and return maps as a means of reducing the analysis of an n dimensional vector field to an $(n-1)$ dimensional discrete map. Forced oscillations have global cross sections obtained by fixing a phase of the forcing. For the forced Van der Pol equation, a different choice of cross section gives more insight into the dynamics. Apart from trajectories that follow the repelling slow manifold along canards, all trajectories make fast jumps from a neighborhood of the fold curve defined by $(x, y) = (-1, 2/3)$ to the sheet of the critical manifold with $x > 1$. The half plane defined by $x = 1$ and $y > -2/3$ is transverse to the vector field and serves as a good choice of cross section. Additionally, we observe that the systems (1.1) and (1.6) are symmetric with respect to the transformation $\sigma(x, y, \theta) = (-x, -y, \theta + 1/2)$. Note that $\sigma \circ \sigma$ is the identity. Denoting the flow map from the section $\Sigma = \{(1, y, z) | y > -2/3\}$ to the section $\sigma(\Sigma)$ by F, this motivates us to define the ***half return map*** on Σ to be $H_\varepsilon = \sigma \circ F$. The symmetry of the system implies that $H_\varepsilon \circ H_\varepsilon$ is the return map of Σ. It is easier to analyze H_ε than $H_\varepsilon \circ H_\varepsilon$ since there is less folding. We also want to analyze the singular limit H_0 of H_ε. Apart from the trajectories with canards, all of the returns to Σ occur in a thin strip around the circle $x = 1, y = 2/3$. The image of the singular limit of the return map lies in this circle except for the limits of some trajectories with canards. To compute the singular limit of H_ε, we must take account of the fact that the cross sections are in the middle of jumps. Points with $x = 1, y = 2/3$ jump to points of the critical manifold with $x = 2$, so we use $x = 2$ as our cross section of the reduced system. Thus, we obtain values of H_0 by following a trajectory of the reduced system from $x = 2$ to $x = 1$, jumping to $x = -2$ and then applying σ to obtain a new point with $x = 2$. Apart from the trajectories with canards, H_0 is a one-dimensional map that is the limit of the maps H_ε after accounting for the change of coordinates resulting from the jump from $x = 1$ to $x = -2$.

Analyzing the singular limit of trajectories with canards is a bit more intricate. When $a > 1$, the desingularized reduced system has a folded saddle p on $x = 1$ with stable manifold $W^s(p)$. This stable manifold intersects $x = 2$ at one or more points where there are discontinuities of H_0. Trajectories on one side of $W^s(p)$ flow directly to $x = 1$, while trajectories on the other side of $W^s(p)$ turn and follow the unstable manifold $W^u(p)$ to $x = 1$. These discontinuities can be "patched" by making the map H_0 multivalued. The appropriate patch contains limits of all canard trajectories of the full system as $\varepsilon \to 0$. The singular limits of canards consist of segments on the branch of $W^s(p)$ on the repelling sheet of the critical manifold. The canards can end and jump to one of the attracting sheets of the critical manifold $|x| > 1$ anywhere along this branch of $W^s(p)$. If the jump is "back" to $x > 1$, an explicit calculation establishes that the jump lands on the side of $W^s(p)$ where trajectories follow $W^u(p)$ to $x = 1$, determining the value of H_0. If the jump is "forward" to (x, θ) with $x < -1$, then the forward trajectory of (x, θ) does not encounter the fold curve $x = -1$ before its next jump. To obtain a consistent value of H_0 for this trajectory, we follow the trajectory of (x, θ) back to $x = -2$ and take

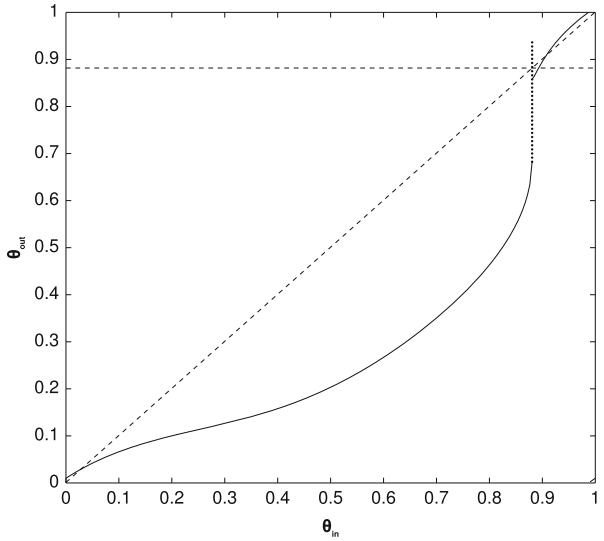

Fig. 1.8 The half return map H_0 of the singular limit of the forced Van der Pol equation for parameter values $(a, \omega) = (1.1, 1.57)$ and cross section $x = 2$ is displayed. The discontinuity of H_0 occurs at the intersection $(\theta_s, 2)$ of $x = 2$ with the stable manifold of the folded saddle. The extension of H_0 to trajectories with singular canards is plotted as a *dotted vertical segment*. The *horizontal dashed line* is $\theta = \theta_s$, showing that the images of the singular canards contains θ_s

the value of θ there to be the value of H_0 for this trajectory. See Fig. 1.8. Thus the patches to the discontinuities of H_0 consist of vertical segments. The extreme value of these segments corresponds to ***maximal canards*** that follow $W^u(p)$ to its first intersection with one of the fold curves. Whether this intersection occurs on $x = 1$ or $x = -1$ depends upon parameters (a, ω); in the parameter region we study it occurs on $x = 1$.

A second aspect of H_0 that is important in its dynamics are critical points where $\frac{dH_0}{d\theta} = 0$. These critical points occur where the desingularized reduced system is tangent to the line $x = 2$. They satisfy the equation $2 + a\sin(2\pi\theta)$, so they occur only in the parameter region where $|a| \geq 2$. This prompts us to focus attention on the parameter region $1 < a < 2$ where $\dot{x} < 0$ on the circle $x = 2$. In this parameter region, H_0 has no critical points and it has a single discontinuity located at the intersection of $W^s(p)$ with $x = 2$. No trajectory on S_r can cross the circle $x = 2$ twice. Bold et al. [6] discuss the dynamics of the system when $a > 2$.

Horseshoes of the forced Van der Pol equation are reflected in the dynamics of H_0 by intervals that are mapped back onto themselves at least twice. In the parameter region $1 < a < 2$ this can only occur via trajectories that contain canards, because outside these canard trajectories (the dotted vertical segment of the return map in Fig. 1.8) H_0 is increasing and injective. The canard trajectories yield a "spike" in the graph of H_0, leading us to look for parameter values where the location of the spike is contained in its image. Numerical computations involving

only the desingularized reduced system can identify parameter regions where this happens.

The proof that the system has horseshoes at these parameter values for sufficiently small ε begins with a computer-assisted step that verifies that the intersection of $W^s(p)$ with $x = 2$ lies in the image of singular canard orbits of H_0. This is hardly a delicate calculation, requiring only a moderately accurate determination of the location of $W^s(p)$ in the strip $-1 < x < 2$. The intersection of the branch of $W^s(p)$ with $x = 1$ determines the location of the maximal canard and the extreme value of the canard spike of H_0. The remainder of the proof is based on the theory that has been described above, together with analytical tools for determining that a mapping has a hyperbolic invariant set.

1.6 Structural Stability, Hyperbolic Invariant Sets, and Axiom A

Dynamical systems whose qualitative properties remain unchanged when the system is perturbed are said to be ***structural stable***. The definition states that a vector field is C^r structurally stable if it has a neighborhood in the space of C^r vector fields that are all topologically equivalent. A ***topological equivalence*** of two vector fields is a homeomorphism that maps trajectories of one system to trajectories of the other, preserving the time orientation of the trajectories. The first investigations of structural stability were carried out in the setting of planar vector fields, where chaotic dynamics is precluded by the Poincaré–Bendixson theorem [25]. The first examples of chaotic systems proved to be structurally stable were discrete time systems, namely Smale's horseshoe and Anosov diffeomorphisms of tori [46]. As recounted above, the horseshoe was inspired by the work of Cartwright and Littlewood on the forced Van der Pol equation. One of the achievements of the work recounted here has been to "close the loop" by proving that the forced Van der Pol equation has parameter values for which it is structurally stable while possessing chaotic invariant sets. This proof relies upon the geometric characterization of structural stability via Smale's Axiom A.

The common theme in the horseshoe and Anosov diffeomorphisms is that the chaotic invariant set has a hyperbolic structure consisting of directions that are stretched by the transformation and directions that are contracted by the transformation. A ***hyperbolic structure*** for an invariant set Λ of a diffeomorphism $f : M \to M$ consists of an invariant splitting of the tangent bundle $T_\Lambda M = E^s_\Lambda \oplus E^u_\Lambda$ that satisfies the inequalities

$$|Df^n v| > c\lambda^{-n}|v|, \qquad v \in E^u_x, x \in \Lambda$$
$$|Df^n v| < c\lambda^{n}|v|, \qquad v \in E^s_x, x \in \Lambda$$

for suitable $0 < c, 0 < \lambda < 1$. Thus, vectors in the stable subspaces E_x^s are contracted at an exponential rate while vectors in the unstable subspaces E_x^u are expanded at an exponential rate. In the case that the set Λ is a fixed point x, the ***stable manifold theorem*** proves that there are invariant stable and unstable manifolds W_x^s and W_x^u tangent to E_x^s and E_x^u. Moreover, W_x^s is characterized geometrically as the set of points y whose trajectories $f^n(y)$ approach x as $n \to \infty$ and W_x^s consists of the points y whose trajectories $f^n(y)$ approach x as $n \to -\infty$. Generalizations of the stable manifold theorem were developed for hyperbolic invariant sets and then under even weaker assumptions about hyperbolicity [44].

Smale's Axiom A for a diffeomorphism f requires that periodic orbits are dense in the non-wandering set Ω of f and that Ω possesses a hyperbolic structure. If the state space of f is a compact manifold, then Axiom A implies that Ω decomposes as a finite union of invariant ***basic sets*** on which f is topologically transitive, i.e., has a dense orbit. The culmination of this circle of ideas was the proof that a diffeomorphism of a compact manifold is structurally stable if and only if it satisfies Axiom A together with the ***no cycle property*** [36]. The no cycle property states that there are no cyclic chains of trajectories connecting basic sets. The extension of this theory to differential equations adds new geometric complexity. Flows cannot expand or contract along the direction of the flow, so hyperbolic structures are defined by splitting tangent spaces into three components—one being the line tangent to the flow. However, forced oscillations have global return maps that are diffeomorphisms, so the relationship between flows and diffeomorphisms is simpler in this setting than for general flows. The forced Van der Pol equation is still one of a very few examples in which this hyperbolic theory of structural stability has been brought to bear upon models defined by differential equations [56].

1.7 Structural Stability of the Forced Van der Pol Equation

We end by sketching how the concepts described above fit together to yield a proof that the forced Van der Pol equation has parameters for which it has a return map containing a horseshoe and it is structurally stable. The starting point is numerical computation of the phase portraits of the reduced system. These are two-dimensional flows with one-dimensional half return maps H_0. When $1 < a < 2$, the half return maps have discontinuities at trajectories that lie in the stable manifold of the folded saddle on the circle $x = 1$. We extend the half return maps to be multivalued at these discontinuity points by computing the possible singular canard trajectories that follow the stable manifold of the folded saddle onto the repelling sheet of the critical manifold and then jumping to one of the two attracting sheets of the critical manifold. By scanning the parameter space, we identify parameter values for which the half return map has a fixed point corresponding to a trajectory that possesses a singular canard. See Fig. 1.8. Haiduc does not give an analytic proof that there is a parameter region with this property.

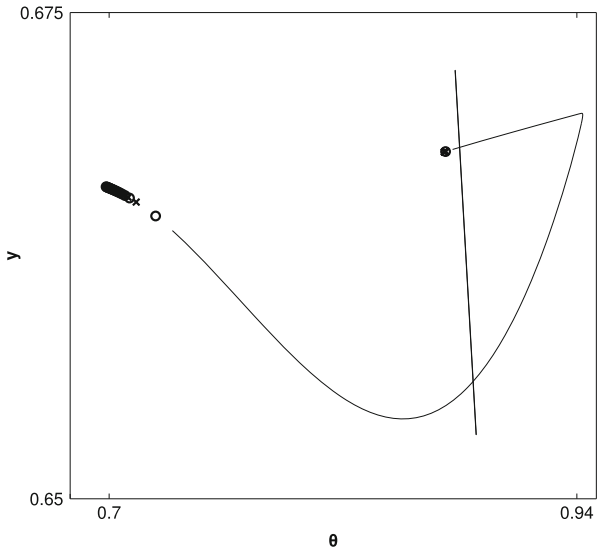

Fig. 1.9 The half return map of the forced Van der Pol equation with parameters $(a, \omega, \varepsilon) = (1.1, 1.57, 0.0001)$ is illustrated. The vertical object is a thin quadrilateral in the cross section $x = 1$. This strip is squeezed vertically and stretched horizontally in an extreme manner. The images of points on the top segment of the strip are plotted with the symbol "o" and the image of points on the bottom segment of the strip are plotted with the symbol "x." The flow of both these segments intersects the repelling slow manifold, but forward integration is unable to track the trajectories that follow the repelling slow manifold. The images of the canard trajectories are plotted as a *solid curve*, computed with a combination of forward and backward integration starting at points on the repelling manifold close to the maximal canard of the system. The returns of these orbits with canards produces a horseshoe

Instead, interval arithmetic is used with an initial value solver to verify the needed estimates. We reemphasize that this is not a delicate calculation.

For small values of $\varepsilon > 0$, the forced Van der Pol equation will have trajectories containing canards whose properties are approximated by the multivalued extension of the reduced system. The canard trajectories flow near the region of the folded saddle and continue along the repelling slow manifold until they jump to one of the attracting slow manifolds and return to the vicinity of the opposite folded saddle. Both directions of jumping occur. The (half) return map of a cross section to these trajectories, displayed in Fig. 1.9, has a hyperbolic invariant set Λ_h that is topologically conjugate to the horseshoe. The expansion of the hyperbolic structure is due to the stretching that occurs as canards jump from the repelling slow manifold at different times, while the contraction of the hyperbolic structure is due to the rapid convergence of trajectories to the attracting slow manifolds. The transversality of the two is a consequence of the $O(\varepsilon)$ splitting between the attracting and repelling slow manifolds along the intersection of their extensions. The set Λ_h consists of intersections with the cross section of trajectories with an infinite number of canards and jumps in both forward and backward time. Labeling each jump with a symbol

that indicates the direction of the jump gives a map ψ from Λ_h to the space of bi-infinite sequences of the two symbols. Using the hyperbolicity of the return map, it is readily proved as with the horseshoe that ψ is a bijection: every bi-infinite sequence of symbols corresponds to exactly one trajectory.

In fact, the horseshoe is embedded in a larger basic set of the system. The half return map has a region containing the canards and extending to the right that maps into itself with a "Z"-shaped shape, and this gives rise to a basic set Λ whose trajectories can be labeled by sequences of three symbols rather than two. The trajectories corresponding to the third symbol return to the folded saddle between their first and second return to the cross section. Their jumps from the repelling manifold occur at a different location than the points in the horseshoe. The symbolic dynamics of the set Λ yield a dense set of periodic orbits in this basic set. The remainder of the non-wandering set of the system consists of an unstable periodic orbit with small amplitudes of x and y and a stable periodic orbit that contains a stable fixed point of the half return map. The system is proved to be structurally stable by demonstrating that wandering trajectories that do not lie in the stable manifold of Λ tend to the stable periodic orbit, and that trajectories that do not lie in the unstable manifold of Λ tend to the unstable periodic orbit in backward time. This precludes the existence of chain recurrent cycles and concludes our overview of the proof of structural stability.

1.8 Afterword

A surprising aspect of the results recounted here is that, unlike the Lorenz attractor [35], they do not yield chaotic trajectories that have been computed with an initial value solver. Because the trajectories contain canards, finding them by integrating forward with an initial value solver is not feasible if ε is sufficiently small. Pragmatically, $\varepsilon = 10^{-3}$ is sufficiently small as we illustrated with the Van der Pol equation without periodic forcing. Even if we did not have to contend with the numerical difficulties of integrating trajectories with canards to exhibit Λ, it would still be hard to find because it is not an attractor. The distinction between chaotic invariant sets and chaotic attractors is important, and the search for chaotic attractors has been a major theme within dynamical systems theory research for decades. Perhaps the earliest example of a hyperbolic chaotic attractor other than Anosov diffeomorphisms and flows is the ***solenoid*** [46], an attractor of a three-dimensional diffeomorphism. A few years after Smale described this example, Plykin [42] constructed a diffeomorphism of the plane that has a hyperbolic chaotic attractor. Nonetheless, chaotic attractors are more common than suggested by these structurally stable examples. Henon [24] discovered that diffeomorphisms of the plane defined by quadratic equations can have chaotic attractors. It is evident that the attractor of the Henon map cannot satisfy Axiom A. An extraordinary set of results beginning with the work of Jakobson [26] on one- dimensional quadratic maps culminated in the demonstration by Benedicks and Carleson [4] that Henon

maps could have chaotic attractors. Their results have been extended to larger classes of "Henon-like" maps, first by Mora and Viana [41] and then further by Wang and Young [55]. In principle, the theory of Henon-like diffeomorphisms should apply to the forced Van der Pol equation. When $a > 2$, the singular half return map H_0 has critical points where its image folds. There are parameter values $a > 2$ where H_0 maps an interval into itself in a unimodal fashion. That is a starting point for applying the theory of Henon-like diffeomorphisms. Guckenheimer et al. [21] took a significant step along this path. (See also Mishchenko et al. [39].) They showed that the singular limits of return maps of smooth slow–fast systems with two slow and one fast variable yield return maps with sufficient smoothness to apply the theory of Wang and Young [54]. To avoid complications of dealing with canards, they studied a modification of the forced Van der Pol equation for which there are no folded singularities or canards but there are critical points of the singular return map. They located parameter regimes where there are chaotic attractors in this modified system of differential equations. Like Haiduc's thesis, this result puts a solid mathematical foundation under heuristic arguments that were used for decades to connect the theory of chaotic dynamical systems to examples defined by actual differential equations. The theory of chaotic attractors is based on the analysis of diffeomorphisms of the plane that perturb rank one maps. This mathematical setting is realized by the singular limits of slow–fast systems.

References

1. Arnol'd, V.I.: Singularity Theory. London Mathematical Society Lecture Note Series, vol. 53. Cambridge University Press, Cambridge (1981). Selected papers, Translated from the Russian, With an introduction by C. T. C. Wall
2. Arnold, V.I., Afrajmovich, V.S., Il'yashenko, Yu.S., Shil'nikov, L.P.: Bifurcation Theory and Catastrophe Theory. Springer, Berlin (1999). Translated from the 1986 Russian original by N. D. Kazarinoff, Reprint of the 1994 English edition from the series Encyclopaedia of Mathematical Sciences [ıt Dynamical systems. V, Encyclopaedia Math. Sci., 5, Springer, Berlin, 1994; MR1287421 (95c:58058)]
3. Barrow-Green, J.: Poincaré and the Three Body Problem. History of Mathematics, vol. 11. American Mathematical Society, Providence (1997)
4. Benedicks, M., Carleson, L.: The dynamics of the Hénon map. Ann. Math. (2) **133**(1), 73–169 (1991)
5. Benoît, É.: Systèmes lents-rapides dans $\mathbf{R}^3$ et leurs canards. In: Third Schnepfenried Geometry Conference, vol. 2 (Schnepfenried, 1982). Astérisque, vol. 109, pp. 159–191. Soc. Math. France, Paris (1983)
6. Bold, K., Edwards, C., Guckenheimer, J., Guharay, S., Hoffman, K., Hubbard, J., Oliva, R., Weckesser, W.: The forced van der Pol equation. II. Canards in the reduced system. SIAM J. Appl. Dyn. Syst. **2**(4), 570–608 (2003, electronic)
7. Cartwright, M.L.: Balthazar van der Pol. J. Lond. Math. Soc. **35**, 367–376 (1960)
8. Cartwright, M.: Some points in the history of the theory of nonlinear oscillations. Bull. Inst. Math. Appl. **10**(9–10), 329–333 (1974)
9. Cartwright, M.L., Littlewood, J.E.: On non-linear differential equations of the second order. I. The equation $\ddot{y} - k(1 - y^2)y + y = b\lambda k \cos(\lambda t + a), k$ large. J. Lond. Math. Soc. **20**, 180–189 (1945)

10. Cartwright, M.L., Littlewood, J.E.: On non-linear differential equations of the second order. II. The equation $\ddot{y} + kf(y)\dot{y} + g(y,k) = p(t) = p_1(t) + kp_2(t)$; $k > 0$, $f(y) \geqq 1$. Ann. Math. (2) **48**, 472–494 (1947)
11. Cartwright, M.L., Littlewood, J.E.: Errata: On non-linear differential equations of the second order. II. Ann. Math. (2) **49**, 1010 (1948)
12. Cartwright, M.L., Littlewood, J.E.: Addendum to 'On non-linear differential equations of the second order. II'. Ann. Math. (2) **50**, 504–505 (1949)
13. Diener, M.: The canard unchained or how fast/slow dynamical systems bifurcate. Math. Intelligencer **6**(3), 38–49 (1984)
14. Dumortier, F., Roussarie, R.: Canard cycles and center manifolds. Mem. Am. Math. Soc. **121**(577), x+100 (1996). With an appendix by Cheng Zhi Li
15. Eckmann, J.-P., Ruelle, D.: Ergodic theory of chaos and strange attractors. Rev. Mod. Phys. **57**(3, part 1), 617–656 (1985)
16. Fenichel, N.: Geometric singular perturbation theory for ordinary differential equations. J. Differ. Equ. **31**(1), 53–98 (1979)
17. Gleick, J.: Making a new science. In: Chaos. Penguin Books, New York (1987).
18. Grasman, J.: Asymptotic Methods for Relaxation Oscillations and Applications. Applied Mathematical Sciences, vol. 63. Springer, New York (1987)
19. Guckenheimer, J., Hoffman, K., Weckesser, W.: Global bifurcations of periodic orbits in the forced van der Pol equation. In: Global Analysis of Dynamical Systems, pp. 261–276. Inst. Phys., Bristol (2001)
20. Guckenheimer, J., Hoffman, K., Weckesser, W.: The forced van der Pol equation. I. The slow flow and its bifurcations. SIAM J. Appl. Dyn. Syst. **2**(1), 1–35 (2003, electronic)
21. Guckenheimer, J., Wechselberger, M., Young, L.-S.: Chaotic attractors of relaxation oscillators. Nonlinearity **19**(3), 701–720 (2006)
22. Haiduc, R: Horseshoes in the forced van der Pol equation. Thesis (Ph.D.)–Cornell University. ProQuest LLC, Ann Arbor (2005)
23. Haiduc, R.: Horseshoes in the forced van der Pol system. Nonlinearity **22**(1), 213–237 (2009)
24. Hénon, M.: A two-dimensional mapping with a strange attractor. Commun. Math. Phys. **50**(1), 69–77 (1976)
25. Hirsch, M.W., Smale, S., Devaney, R.L.: Differential Equations, Dynamical Systems, and an Introduction to Chaos. Pure and Applied Mathematics (Amsterdam), vol. 60, 2nd edn. Elsevier/Academic, Amsterdam (2004)
26. Jakobson, M.V.: Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. Commun. Math. Phys. **81**(1), 39–88 (1981)
27. Jones, C.K.R.T.: Geometric singular perturbation theory. In: Dynamical Systems (Montecatini Terme, 1994). Lecture Notes in Mathematics, vol. 1609, pp. 44–118. Springer, Berlin (1995)
28. Krupa, M., Szmolyan, P.: Extending geometric singular perturbation theory to nonhyperbolic points—fold and canard points in two dimensions. SIAM J. Math. Anal. **33**(2), 286–314 (2001, electronic)
29. Levi, M.: Qualitative analysis of the periodically forced relaxation oscillations. Mem. Am. Math. Soc. **32**(244), vi+147 (1981)
30. Levinson, N.: A second order differential equation with singular solutions. Ann. Math. (2) **50**, 127–153 (1949)
31. Li, T.Y., Yorke, J.A.: Period three implies chaos. Am. Math. Monthly **82**(10), 985–992 (1975)
32. Littlewood, J.E.: Errata: On non-linear differential equations of the second order. III. The equation $\ddot{y} - k\left(1 - y^2\right)\dot{y} + y = b\mu k\cos(\mu t + \alpha)$ for large k, and its generalizations $\cos(\mu t + \alpha)$ for large k, and its generalizations. Acta Math. **98**(1–4), 110 (1957)
33. Littlewood, J.E.: On non-linear differential equations of the second order. III. The equation $\ddot{y} - k(1 - y^2)\dot{y} + y = b\mu k\cos(\mu t + \alpha)$ for large k, and its generalizations. Acta Math. **97**, 267–308 (1957)
34. Littlewood, J.E.: On non-linear differential equations of the second order. IV. The general equation $\ddot{y} + kf(y)\dot{y} + g(y) = bkp(\phi)$, $\phi = t + \alpha$. Acta Math. **98**, 1–110 (1957)
35. Lorenz, E.N.: Deterministic nonperiodic flow. J. Atmos. Sci. **20**, 130–141 (1963)

36. Mañé, R.: Ergodic Theory and Differentiable Dynamics. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], vol. 8. Springer, Berlin (1987). Translated from the Portuguese by Silvio Levy
37. McMurran, S.L., Tattersall, J.J.: The mathematical collaboration of M. L. Cartwright and J. E. Littlewood. Am. Math. Monthly **103**(10), 833–845 (1996)
38. McMurran, S.L., Tattersall, J.J.: Cartwright and Littlewood on van der Pol's equation. In: Harmonic Analysis and Nonlinear Differential Equations (Riverside, CA, 1995). Contemp. Math., vol. 208, pp. 265–276. Amer. Math. Soc., Providence (1997)
39. Mishchenko, E.F., Kolesov, Yu.S., Kolesov, A.Yu., Rozov, N.Kh.: Asymptotic Methods in Singularly Perturbed Systems. Monographs in Contemporary Mathematics. Consultants Bureau, New York (1994). Translated from the Russian by Irene Aleksanova
40. Mishchenko, E.F., Rozov, N.Kh.: Differential Equations with Small Parameters and Relaxation Oscillations. Mathematical Concepts and Methods in Science and Engineering, vol. 13. Plenum Press, New York (1980). Translated from the Russian by F. M. C. Goodspeed
41. Mora, L., Viana, M.: Abundance of strange attractors. Acta Math. **171**(1), 1–71 (1993)
42. Plykin, R.V.: Sources and sinks of A-diffeomorphisms of surfaces. Mat. Sb. (N.S.) **94**(136), 243–264, 336 (1974)
43. Poincaré, H.: Sur les probelème des trois corps et les équations de la dynamique. Acta Math. **13**, 1–270 (1890)
44. Ruelle, D.: Ergodic theory of differentiable dynamical systems. Inst. Hautes Études Sci. Publ. Math. (50), 27–58 (1979)
45. Ruelle, D., Takens, F.: On the nature of turbulence. Commun. Math. Phys. **20**, 167–192 (1971)
46. Smale, S.: Differentiable dynamical systems. Bull. Am. Math. Soc. **73**, 747–817 (1967)
47. Smale, S.: Diffeomorphisms with many periodic points. In: Differential and Combinatorial Topology (A Symposium in Honor of Marston Morse), pp. 63–80. Princeton University Press, Princeton (1965)
48. Smale, S.: Essays on dynamical systems, economic processes, and related topics. In: The Mathematics of Time. Springer, New York (1980).
49. Smale, S.: Finding a horseshoe on the beaches of Rio. Math. Intelligencer **20**(1), 39–44 (1998)
50. Tattersall, J., McMurran, S.: An interview with Dame Mary L. Cartwright, D.B.E., F.R.S. College Math. J. **32**(4), 242–254 (2001)
51. Van der Pol, B.: A theory of the amplitude of free and forced triode vibrations. Radio Rev. **1**, 701–710 (1920)
52. Van der Pol, B.: On "relaxation oscillations". I. Phil. Mag. **2**, 978–992 (1926)
53. Van der Pol, B., Van der Mark, J.: Frequency demultiplication. Nature **120**, 363–364 (1927)
54. Wang, Q., Young, L.-S.: Nonuniformly expanding 1D maps. Commun. Math. Phys. **264**(1), 255–282 (2006)
55. Wang, Q., Young, L.-S.: Toward a theory of rank one attractors. Ann. Math. (2) **167**(2), 349–480 (2008)
56. Wilczak, D.: Uniformly hyperbolic attractor of the Smale-Williams type for a Poincaré map in the Kuznetsov system. SIAM J. Appl. Dyn. Syst. **9**(4), 1263–1283 (2010). With online multimedia enhancements.
57. Yorke, J., Sauer, T.: Chaos. Scholarpedia. http://www.scholarpedia.org/article/Chaos

Chapter 2
Periodic Orbits Close to Grazing for an Impact Oscillator

D.R.J. Chillingworth and A.B. Nordmark

Abstract We show how the geometric *impact surface* approach to the dynamics of an impact oscillator provides an immediate visualization of the criteria that determine the existence of an impacting periodic orbit close to grazing. We recover the criteria set out earlier by A. Nordmark and indicate how the geometric setting and singularity geometry may be exploited to yield appropriate criteria in degenerate situations where the Nordmark criteria would not apply.

2.1 The Impact Oscillator

There are many physical and engineering contexts in which a vibrating or oscillating system encounters impacts—typically accompanied by undesirable noise and wear and thus demanding to be eliminated or at least controlled. See e.g. [1,3,14], as well as [9], for examples and further references.

The simplest model for such a system is the one degree of freedom *impact oscillator* of the form

$$\ddot{x} + f(x, \dot{x}) = g(t) \tag{2.1}$$

where $f : \mathbf{R}^2 \to \mathbf{R}$ and $g : \mathbf{R} \to \mathbf{R}$ are smooth functions, g is T-periodic with $T > 0$, and where x is restricted to the region $x \geq c$. To describe the local

D.R.J. Chillingworth (✉)
Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK
e-mail: drjc@soton.ac.uk

A.B. Nordmark
Department of Mechanics, KTH, 10044 Stockholm, Sweden
e-mail: nordmark@mech.kth.se

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_2,

consequence of impact with the *obstacle* at $x = c$ a *restitution law* is assumed, which usually takes the form of replacing $\dot{x} < 0$ by $-r\dot{x}, r > 0$ at impact.

Impact oscillators can be seen as special cases of N-degree of freedom (autonomous) systems with codimension-1 obstacles, or of general hybrid systems involving piecewise-smooth vector fields in $\mathbf{R}^n$ together with maps defined on the codimension-1 boundaries of the regions of smoothness: see [1] for further discussion.

An impact oscillator (2.1) exhibits certain bifurcation phenomena that arise specifically from the impacts, in addition to those seen in smooth (non-impacting) systems. The most natural example to study first is that in which a hyperbolic (or, at least, isolated and persistent) periodic orbit of period T lying in the region $x > c$ collides with the obstacle at $x = c$ as a parameter μ in the system increases through a critical value μ_0. For $\mu = \mu_0$ the orbit satisfies $\dot{x} = 0$ where $x = c$, that is it *grazes* the obstacle. It is not a priori clear whether such a periodic orbit will persist (with impacts) for $\mu > \mu_0$ or how the result may depend on the nature of the Poincaré map of the grazing orbit.

There have been many studies, both theoretical and numerical, of the dynamical behaviour of an impact oscillator close to grazing: see for example [2, 4, 6–8, 10–12, 15, 16]. A thorough investigation of the persistence or otherwise of a grazing T-periodic orbit was carried out by Nordmark in [13] in the wider context of an n-degree of freedom (autonomous) system with a codimension-1 obstacle, and as an illustration in that paper the general results are applied to the 1-degree of freedom system (2.1). A detailed analysis of criteria for the persistence of orbits of period nT for $n = 1, 2, 3 \ldots$ is carried out, leading to interesting algebra related to the eigenvalues of the Poincaré map. However, simple criteria in terms of the Poincaré map itself are not immediately visible.

The purpose of this paper is, first, to extract the essential data from the general formalism of [13] in order to write down what is needed for the impact oscillator (2.1) in the cases $n = 1$ and $n = 2$. Having done this, we then take a different approach and use the *impact surface* description formulated in [5] in order to reveal in this setting the extremely simple geometric criteria that underlie the phenomenon of grazing bifurcation. Finally we show how this geometry, when converted into algebra, yields the Nordmark criteria.

2.1.1 Nordmark's Criteria

We apply the general formalism from [13] to this setting, and in particular to the cases designated there as (n) where $n = 1$ or $n = 2$, that is cases of impacting orbits close to grazing that have a single impact during a period of T or of $2T$. The aim is to derive simple criteria for the existence of such periodic orbits as a system with a grazing orbit P of period T is perturbed by a parameter μ. We suppose without loss of generality that the grazing occurs when $\mu = 0$.

2.1.1.1 Single Impact T-Periodic Orbits

Following the notation from [13], let

$$A = \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix}$$

denote the Jacobian matrix of the time-T map in the $(x, \dot{x})$-plane for the system (2.1) (in the absence of the obstacle) at the point $p = (c, 0)$ when $\mu = 0$. The T-periodicity of P means that p is a fixed point of the map. It is convenient to introduce also the covector $C = [1, 0]$ and the vector $B = [0, 1]^t$ where ^{t} denotes transpose.[1]

First assume that

$$\Delta_- := \det(I - A) \neq 0$$

so that P (although not necessarily hyperbolic) persists as a T-periodic orbit P_μ in the absence of the obstacle for small $|\mu| > 0$. A central role in [13] is played by the quantity

$$e := C(I - A)^{-1}(\kappa, \lambda)^t \tag{2.2}$$

$$= \Delta_-^{-1}\big((1 - \delta)\kappa + \gamma\lambda\big) \tag{2.3}$$

where $(\kappa, \lambda) = (\frac{\partial x}{\partial \mu}, \frac{\partial \dot{x}}{\partial \mu})$ evaluated at p when $\mu = 0$.

To first order in μ, the location of a T-periodic ($n = 1$) single-impacting orbit (when it exists) relative to the grazing point p is given by $\mu X \in \mathbf{R}^2$ where ([13, eq. (51)])

$$C(I - A)X = 0\,, \quad CX = -1 \tag{2.4}$$

and from [13, eq. (34)] the condition for such an orbit to exist is

$$y_1 e \mu > 0 \tag{2.5}$$

where from [13, eq. (32)] we have (up to a positive multiple)

$$y_1 = B^t(I - A)X. \tag{2.6}$$

[1] The vector B differs from that in [13] by a positive scalar multiple.

From the second equation of (2.4) we see X has the form $X = (-1, x_1)^t$, and then combining (2.4) and (2.6) we may write

$$(I - A)\begin{pmatrix} -1 & 0 \\ x_1 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -\gamma \\ y_1 & 1-\delta \end{pmatrix} \tag{2.7}$$

which on taking determinants gives

$$\gamma\, y_1 = -\Delta_- \,. \tag{2.8}$$

Assuming $\gamma \neq 0$ this gives

$$y_1 e = -\gamma^{-1}\left((1-\delta)\kappa + \gamma\lambda\right)$$

and so the criterion (2.5) gives the following basic result:

Proposition 2.1 (Nordmark). *Suppose* $\Delta_- \neq 0, \gamma \neq 0$ *and also*

$$\psi := (1-\delta)\kappa + \gamma\lambda \neq 0.$$

For $|\mu|$ *sufficiently small, a single-impact orbit of period* T *exists close to the grazing orbit if and only if the parameter* μ *has the opposite sign from* $\gamma\psi$.

In particular this implies that as μ passes through zero an impacting period-T orbit must exist on one side or the other. If it exists for $\mu > 0$ then the impacting orbit is a continuation of the non-impacting orbit, while for the case $\mu < 0$ the free and the impacting orbit exist simultaneously and mutually annihilate at $\mu = 0$. The latter phenomenon is often called a *nonsmooth fold*: see [1].

2.1.1.2 Single Impact $2T$-Periodic Orbits

For a $2T$-periodic orbit ($n = 2$) with single impact the location X is now by analogy with (2.4) given by

$$C(I - A^2)X = 0\,, \quad CX = -1 \tag{2.9}$$

while the conditions for such an orbit to exist are

$$b_{2,1}\, e\, \mu > 0 \tag{2.10}$$

$$y_2\, e\, \mu > 0 \tag{2.11}$$

where

$$b_{2,1} = CAX + 1 \tag{2.12}$$

$$y_2 = B^t(I - A^2)X. \tag{2.13}$$

Let α_2 , γ_2 denote the α, γ-entries in A^2, that is

$$\alpha_2 = \alpha^2 + \beta\gamma = \alpha\eta - \det A \tag{2.14}$$

$$\gamma_2 = \gamma\,\eta \tag{2.15}$$

where $\eta = tr\,A$ is the trace of A. By analogy with (2.7) we have on taking determinants

$$\eta\gamma\, y_2 = -\det(I - A^2) = -\Delta_-\Delta_+ \tag{2.16}$$

with $\Delta_+ = \det(I + A)$, and so

$$\gamma y_2 e = -\eta^{-1}\Delta_+\big((1-\delta)\kappa + \gamma\lambda\big)$$

provided $\eta \neq 0$. If (2.11) holds then the condition (2.10) can simply be expressed by saying that $b_{2,1}$ and y_2 have the same sign. Now if $X = (-1, x_2)^t$ satisfies (2.9) we find

$$b_{2,1} = -\alpha + \gamma x_2 + 1 \tag{2.17}$$

$$= -\alpha + \gamma\gamma_2^{-1}(\alpha_2 - 1) + 1 \tag{2.18}$$

$$= -\eta^{-1}\Delta_- \tag{2.19}$$

using (2.14) and (2.15). From (2.16) the same-sign condition on $b_{2,1}$ and y_2 is therefore precisely the condition that $\gamma\,\Delta_+ > 0$. To summarize:

Proposition 2.2 (Nordmark). *Suppose $\Delta_- \neq 0$, $\eta\psi \neq 0$ and also $\gamma\Delta_+ \neq 0$. For $|\mu|$ sufficiently small, a single-impact orbit of period $2T$ exists close to the grazing orbit if and only if $\gamma\Delta_+ > 0$ and μ has the opposite sign from $\eta\psi$.*

2.1.2 The Impact Surface Approach

We now turn to address the same questions, but from the point of view of the geometry of the *impact surface* as set out in [5]. We briefly describe the formalism, and refer to [5] for further details.

The *impact surface* V_c is defined as

$$V_c = \{(\tau, v, t) \in \mathbf{R}^3 : x_c(\tau, v, t) = c\} \tag{2.20}$$

where $x(c, v, \tau; u)$ denotes the solution of (2.1) with initial data $(x, \dot{x}) = (c, v)$ when $u = \tau$, and we write $x_c(\tau, v, t) = x(c, v, \tau; \tau + t)$.

Clearly V_c contains the $t = 0$ plane Π, and in [5] it is shown that $V_c = \Pi \cup V_c'$ where V_c' is a smooth 2-manifold (except generically for finitely many values of c at which V_c' undergoes a Morse perestroika) that intersects Π transversely along the τ-axis. The impacting system can then be modelled by a discontinuous discrete dynamical system on Π defined as the composition

$$G_c := R \circ \phi_c \circ F_c : \Pi \to \Pi$$

where

$F_c : \Pi \to V_c'$ is the *first hit* map $(\tau, v) \mapsto (\tau, v, t_1)$ where t_1 is the smallest positive value of t for which $(\tau, v, t) \in V_c'$
$\phi_c : V_c' \to \Pi$ is the *reset* map $(\tau, v, t) \mapsto (\tau + t, \dot{x}(\tau + t))$
$R : \Pi \to \Pi$ is the restitution map (such as $(\tau, v) \mapsto (\tau, -rv)$).

In other words, we reconstruct the dynamics by taking each impact at $x = c$ as a new set of initial data and then proceeding (after applying restitution) to the next impact.

Remark 2.1. In general the map F_c may fail to be defined for either of the following two reasons: for given $(\tau, v) \in \Pi$ the subset $\{t : (\tau, v, t) \in V_c'\}$ of $\mathbf{R}$ may be empty, or (when $v = 0$) may accumulate at zero. Neither of these arises in the present discussion, however. It is also the case that infinitely many impacts can occur in a finite time interval (the *chatter* phenomenon), but again for current purposes this does not concern us.

2.1.2.1 Singularity Geometry

The singularity structure of the projection map

$$\pi_c = \pi | V_c' : (\tau, v, t) \mapsto (\tau, v)$$

and of the reset map ϕ_c are crucial in construction of the dynamical map $G_c : \Pi \to \Pi$. For $(\tau, 0, t) \in V_c'$ let $a = a(\tau) := \ddot{x}_c(\tau, 0, 0)$ denote the 'initial' acceleration. Then from [5] we have

Proposition 2.3.

1. *The singular set $S(\pi_c)$ consists of the points on V_c' where $\dot{x}_c = 0$.*
   If $\ddot{x}_c \neq 0$ then the singularity is a fold singularity.
2. *The singular set $S(\phi_c)$ consists of the points on V_c' where $v = 0$.*
   If $a \neq 0$ then the singularity is a fold singularity.

The first part of the proposition is elementary, from the geometry of V_c' itself, although the second part is less obvious. Note that $\pi_c S(\pi_c)$ is the apparent outline (or contour) of V_c' as viewed along the t-axis; in [5] we call $S(\pi_c)$ the *horizon curve* H.

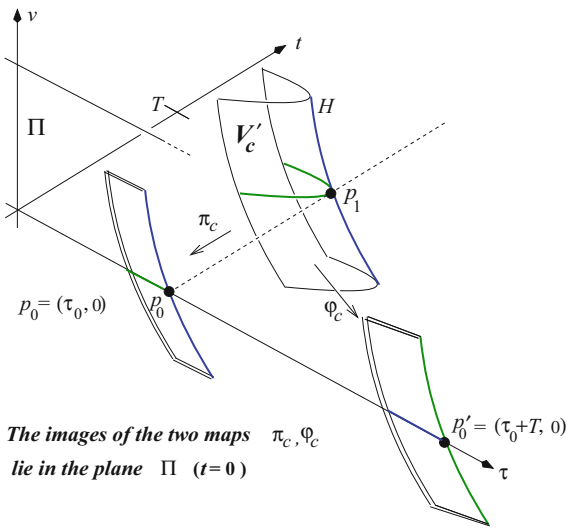

Fig. 2.1 Two fold maps $\pi_c, \phi_c : V_c' \to \Pi$. The point $p_1 \in V_c'$ is a singular (fold) point for both maps

Consider now a grazing and otherwise non-impacting periodic orbit P with period T. In terms of the impact surface the orbit P is represented by a point $p_0 = (\tau_0, 0, 0)$ on the τ-axis such that the line $l(\tau_0)$ through p_0 parallel to the t-axis intersects V_c' precisely at the points $p_n = (\tau_0, 0, nT), n \in \mathbf{Z}$ and is tangent to V_c' there. Suppose the grazing is quadratic, so that $a > 0$ and $\ddot{x}_c > 0$ at all points p_n. From Proposition 2.3 we obtain

Proposition 2.4. *Each of the maps $\pi_c : V_c' \to \Pi$ and $\phi_c : V_c' \to \Pi$ has a fold singularity at each of the points p_n, $n \neq 0$.*

See Fig. 2.1 for an illustration of this geometry for $n = 1$.

The T-periodicity of g means that the impact surface V_c is invariant under translation $\tau \mapsto \tau + T$, so working modulo T on the τ-axis we can regard ϕ_c as taking every $p_n \in V_c'$ to $p_0 \in \Pi$, as does π_c. Therefore at each $p_n, n \neq 0$, we have *two* fold singularities

$$\pi_c, \phi_c : (V_c', p_n) \to (\Pi, p_0).$$

We now use this geometry to derive a criterion for the existence of impacting period-T orbits after perturbation based on the following elementary properties of fold maps. By *smooth curve* we mean a smooth arc or 1-manifold.

Lemma 2.1. *Let $f : (\mathbf{R}^2, q) \to (\mathbf{R}^2, p)$ be a fold singularity with singular set $S(f)$ (a smooth curve through q) in $\mathbf{R}^2$; then $U(f) := fS(f)$ is a smooth curve (fold image curve) through p in $\mathbf{R}^2$. Let λ be a smooth curve in $\mathbf{R}^2$ passing through q, and suppose the tangent to λ at q does not coincide with the kernel of the derivative $Df(q) : \mathbf{R}^2 \to \mathbf{R}^2$. Then $f(\lambda)$ is a smooth curve through p and is tangent to $U(f)$.*

These facts can easily be deduced from the normal form $(x, y) \mapsto (x, y^2)$ for a fold map $(\mathbf{R}^2, 0) \to (\mathbf{R}^2, 0)$.

2.1.3 Single Impact Period T Orbits

Suppose now that $p \in \Pi$ represents a single-impact T-periodic orbit, so that p is a fixed point of the dynamical map $G_c : \Pi \to \Pi$, that is

$$p = G_c(p) = R \circ \phi_c \circ F_c(p). \tag{2.21}$$

This can be re-expressed as

$$\pi_c(q) = R \circ \phi_c(q) \tag{2.22}$$

where $q = F_c(p) \in V_c'$. Thus, rather than seek a fixed point p of G_c directly, we instead look for a point q on V_c' that is taken to the *same* point under the two maps π_c and $R \circ \phi_c$; the advantage here is that both of these maps are smooth.

Taking local coordinates (s, v) on V_c' at p_1 (here $t = T + s$) we find that the maps π_c and $\phi_c : V_c' \to \Pi$ can be expressed as

$$\pi_c : (s, v) \mapsto (a^{-1}v, v) + O(2) \tag{2.23}$$

$$\phi_c : (s, v) \mapsto (s + a^{-1}v, as + v) + O(2) \tag{2.24}$$

for constant $a = \ddot{x}(p_1) = \ddot{x}(p_0) > 0$. The result (2.23) is easily obtained by noting that the fold image curve is orthogonal to $(\frac{\partial x_c}{\partial \tau}, \frac{\partial x_c}{\partial v})$ evaluated at p_1 and using the identity (2.32) below, while (2.24) follows from expressing V_c' locally near p_1 as a graph $\tau = \tau(t, v)$, checking that $(\frac{\partial \tau}{\partial t}, \frac{\partial \tau}{\partial v}) = (0, a^{-1})$ at p_1 and again using (2.32).

Consequently, in order for (2.22) to be satisfied by a point $q \in V_c'$ near p_1 the coordinates (s, v) of q must satisfy

$$a^{-1}v = s + a^{-1}v + O(2) \tag{2.25}$$

$$v = -ras - rv + O(2) \tag{2.26}$$

which for small (s, v) holds only at the origin. Thus we already have the following simple but otherwise not immediately obvious result:

Proposition 2.5. *Single-impact period-T orbits cannot exist arbitrarily close to a period-T grazing orbit with nonzero acceleration at the graze.*

A geometric view of this fact is as follows. Equation (2.26) represents (near p_1) a smooth curve Γ through p_1 , namely the locus of those points $q \in V_c'$ such that the v-coordinates of $\pi(q)$ and $R \circ \phi(q)$ coincide. Let $\Gamma' = \pi_c(\Gamma)$ and $\Gamma'' = R \circ \phi_c(\Gamma)$

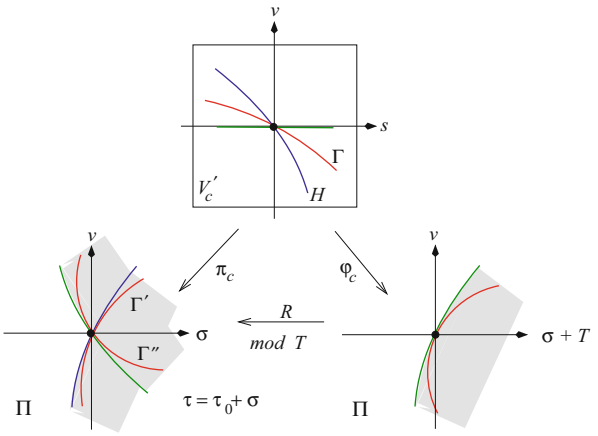

Fig. 2.2 The images Γ' and Γ'' of Γ under the two fold maps π_c and $R \circ \phi_c$ modulo T

be the respective images of Γ: then $\pi_c(q) = R \circ \phi_c(q)$ precisely when the curves Γ' and Γ'' intersect at that point. Now since Γ is tangent neither to the kernel of $D\pi_c(p_1)$ (the s-axis) nor the kernel of $D\phi_c$ at p_1 (which is spanned by $(a, -1)$) it follows from Lemma 2.1 that Γ' and Γ'' are tangent to the fold image curves of π_c and $R \circ \phi_c$, respectively. However, (2.23) and (2.24) show that these fold curves are tangent to the directions $(1, a)$ and $(1, -ra)$, respectively, and so with τ reduced modulo T these curves clearly intersect only at $(\tau, v) = (\tau_0, 0)$. See Fig. 2.2.

From this picture it is now easy to visualize the effect of a smooth perturbation to the system, parametrized by a scalar parameter μ. We suppose that f and g in (2.1) are replaced by f_μ and g_μ (with c also possibly replaced by c_μ) taking their original values at $\mu = 0$. The stability of fold singularities ensures that for sufficiently small $|\mu|$ there exists a unique point $p_1(\mu)$ on $V_c'(\mu)$, varying smoothly with μ, which is a fold singularity point for both $\pi_c, \phi_c : V_c'(\mu) \to \Pi$ with $c = c_\mu$. Let $k = k(\mu)$ and $l = l(\mu)$ denote the τ-coordinate of the image of $p_1(\mu)$ under π_c, ϕ_c respectively, so that the equations of the tangent lines to the fold image curves of π_c, ϕ_c through $(k, 0), (l, 0)$ are

$$v = m(\tau - k) \tag{2.27}$$

$$v = n(\tau - l) \tag{2.28}$$

with slopes $m = m(\mu)$ and $n = n(\mu)$ that are close to a and $-ra$, respectively, for small $|\mu|$. It is elementary to check that the v coordinate of the point of intersection of the two lines (2.27), (2.28) is positive (so that a single impact T-periodic orbit occurs locally) if and only if $k - l$ has the same sign as $(\frac{1}{n} - \frac{1}{m})$. Since for $\mu = 0$ we have $(\frac{1}{n} - \frac{1}{m}) = -a^{-1}(r^{-1} + 1) < 0$, the criterion becomes simply that $l - k > 0$ for sufficiently small $|\mu|$. In the present context by construction $k = \tau$ and $l = \tau + t - T$ and so we arrive at the following result.

Proposition 2.6. *A single-impact T-periodic orbit occurs for sufficiently small $|\mu| \neq 0$ if and only if the t-coordinate $t_1(\mu)$ of $p_1(\mu)$ is greater than T.*

In other words, there is a T-periodic impacting orbit provided that the unique nearby initially grazing orbit that grazes a second time takes longer than time T to do so.

From the definition of $p_1(\mu)$ as the unique point close to p_1 which satisfies $x_c = \dot{x}_c = 0$ as well as $v = 0$ we find from implicit differentiation that

$$t_1'(0) = -a^{-1}\left(\tfrac{\partial x_c}{\partial \tau}\right)^{-1} \det W \tag{2.29}$$

where W denotes the matrix

$$\begin{pmatrix} \frac{\partial x_c}{\partial \tau} & \frac{\partial x_c}{\partial \mu} \\ \frac{\partial \dot{x}_c}{\partial \tau} & \frac{\partial \dot{x}_c}{\partial \mu} \end{pmatrix}$$

evaluated at $p_1(0)$. If we assume $t_1'(0) \neq 0$ then since $a > 0$ the necessary and sufficient condition for the existence of a single-impact T-periodic orbit close to grazing becomes explicitly:

$$(\tfrac{\partial x_c}{\partial \tau})^{-1} \det W \mu < 0. \tag{2.30}$$

At first sight this appears distinct from Nordmark's criterion in Proposition 2.1 which involves derivatives with respect to v (the terms γ, δ) rather than with respect to τ. The two are reconciled, however, using the fact that, regarding the system (without the obstacle) as generating a flow $\{\Psi_t\}$ in $(x, \dot{x}, \tau)$-space, the Jacobian matrix $D\Psi_t(c, v, \tau)$ takes the initial tangent vector $(v, a, 1)^t$ to the time-t tangent vector $(\dot{x}, \ddot{x}, 1)^t$. For a grazing periodic orbit of period T this means that $(v, a, 1)^t$ is an eigenvector of $D\Psi_T(c, 0, \tau)$ with eigenvalue 1, giving the identities

$$\tfrac{\partial x_c}{\partial v}\, a + \tfrac{\partial x_c}{\partial \tau} = 0 \tag{2.31}$$

$$\tfrac{\partial \dot{x}_c}{\partial v}\, a + \tfrac{\partial \dot{x}_c}{\partial \tau} = a \tag{2.32}$$

with derivatives evaluated at p_0 (or p_1). Using these we find that

$$\det W = -a((1-\delta)\kappa + \gamma\lambda) = -a\psi$$

in the notation of Sect. 2.1.1. Thus the criterion (2.30) becomes $\gamma\psi < 0$ just as in Proposition 2.1.

Observe that here we have not required the condition $\det(I - A) \neq 0$, and so (in contrast to [13] where this is heavily used) we have no information about the persistence of a non-impacting T-periodic orbit for $|\mu| \neq 0$.

2.1.4 Single Impact $2T$-Periodic Orbits

We next apply the impact surface geometry to the criterion for persistence of a $2T$-periodic single-impact orbit. Clearly there are two criteria that together are necessary and sufficient for persistence of a $2T$-periodic single-impact orbit, namely:

1. The criterion analogous to (2.30) at $p_2(0)$ rather than $p_1(0)$
2. The condition that V_c' does not intervene between $p_0(\mu)$ and $p_2(\mu)$.

To formulate the second criterion, we need information not only about the positions of fold image curves in Π but also about which side of such a curve is the one that lies in the image of the fold map. From the normal form $(x, y) \mapsto (x, y^2)$ for a fold map we deduce that a coordinate-independent description is as follows.

Lemma 2.2. *Let $f : (\mathbf{R}^2, q) \to (\mathbf{R}^2, p)$ be a fold singularity as in Lemma 2.1. Let $u \in K = \ker DF(q)$ be a nonzero vector. The restriction $f|(q + K)$ has the form*

$$f : q + \xi u \mapsto p + \xi^2 w + O(\xi^3)$$

for a nonzero vector $w \in \mathbf{R}^2$ that does not lie in the range of $Df(q)$. Thus $p + \eta w$ lies in the range of f for all sufficiently small $\eta > 0$.

In our case we have that at every point p_n $(n > 0)$ the kernel K_π of $D\pi_c(p_n)$ is the s-axis while the kernel K_ϕ of $D\phi_c(p_n)$ is spanned by the vector $(1, -a)$, using local coordinates (s, v) on V_c' near p_n as in (2.24). Moreover, representing V_c' locally as the graph of a smooth function $\tau = \tau(s, v)$ we find by implicit differentiation of (2.20) that at each such p_n

$$\tfrac{\partial^2 \tau}{\partial t^2} = -a\big(\tfrac{\partial x_c}{\partial \tau}\big)^{-1} = \big(\tfrac{\partial x_c}{\partial v}\big)^{-1} =: \gamma_n^{-1}$$

using (2.31), and also (with a little more effort) that

$$(a, -1) \cdot D^2\phi_c|K_\phi \,:\, \xi(-1, a) \mapsto -\xi^2 a\gamma_n^{-1} \det A^n = -k\gamma_n^{-1}$$

for $k > 0$, where the dot denotes scalar product. Using this information, sketches of the images of the fold maps π_c, ϕ_c close to p_2 and the map π_c close to p_1 quickly indicate that the condition for V_c' *not* to intervene between $p_0(\mu)$ and $p_2(\mu)$ is

$$\big(t_1'(\mu) - t_2'(\mu)\big)\gamma > 0 \tag{2.33}$$

where $\gamma = \gamma_1$ by definition.

The point $p_n(\mu) = (\tau_n(\mu), 0; t_n(\mu))$ is obtained by solving

$$x_c(\tau, 0; t, \mu) = \dot{x}_c(\tau, 0; t, \mu) = 0$$

close to $p_n = p_n(0)$, from which we find

$$\begin{pmatrix} \frac{\partial x_c}{\partial \tau} & 0 \\ \frac{\partial \dot{x}_c}{\partial \tau} & a \end{pmatrix} \begin{pmatrix} \tau_n'(0) \\ t_n'(0) \end{pmatrix} = - \begin{pmatrix} \frac{\partial x_c}{\partial \mu} \\ \frac{\partial \dot{x}_c}{\partial \mu} \end{pmatrix}$$

with partial derivatives evaluated at p_n, giving

$$\tau_n'(0) = -\left(\tfrac{\partial x_c}{\partial \tau}\right)_n^{-1} \kappa_n = a^{-1}\gamma_n^{-1}\kappa_n \tag{2.34}$$

with the obvious notation and assuming as before that γ and η (hence also γ_2 by (2.15)) are nonzero. The criterion (2.33) therefore becomes

$$\kappa\mu > \gamma_2^{-1}\gamma\kappa_2\mu. \tag{2.35}$$

From the identity

$$\kappa_2 = (\alpha + 1)\kappa + \gamma\lambda$$

this becomes

$$\kappa\mu > \eta^{-1}((\alpha + 1)\kappa + \gamma\lambda)\mu, \tag{2.36}$$

that is (since a number has the same sign as its inverse)

$$\eta((1 - \delta)\kappa + \gamma\lambda)\mu < 0. \tag{2.37}$$

For the criterion (1) the counterpart of (2.30) at p_2 is

$$\left(\tfrac{\partial x_c}{\partial \tau}\right)_2^{-1} \det W_2 \, \mu < 0,$$

where W_2 is the matrix (2.1.3) evaluated at p_2. Using the easily verified fact that $W_2 = (I + A)W$, we may rewrite this as

$$\eta^{-1}\gamma^{-1}\Delta_+((1 - \delta)\kappa + \gamma\lambda)\mu < 0.$$

From (2.37) we see that the additional condition for criterion (1) to hold is that $\gamma\Delta_+ > 0$, and so we recover Proposition 2.2.

2.1.5 Conclusion

We have shown how the impact surface approach to the geometry of grazing bifurcation allows an easy derivation of criteria for the existence of single-impact

T-periodic orbits or $2T$-periodic orbits close to grazing for a T-periodically forced impact oscillator. As in the study by Nordmark [13], the methods can naturally be extended to nT-periodic orbits with m impacts. However, a further advantage of the impact surface description is that the tools of singularity theory can be applied in this context to study multiparameter bifurcation of periodic orbits close to degenerate grazing ($a = 0$) where the geometry of the relevant singularities (described in [5]) is more complicated than the fold geometry that arises here. We aim to pursue this study in further work.

References

1. di Bernardo, M., Budd, C.J., Champneys, A.R., Kowalczyk, P.: Piecewise-Smooth Dynamical Systems: Theory and Applications. Springer, Berlin (2008)
2. di Bernardo, M., Budd, C.J., Champneys, A.R., Kowalczyk, P., Nordmark, A.B., Tost, G.O., Piiroinen, P.T.: Bifurcations in nonsmooth dynamical systems. SIAM Rev. **50**, 629–701 (2008)
3. Brogliato, B.: Nonsmooth Mechanics. Springer, London (1999)
4. Budd, C.J., Dux, F.J., Cliffe, A.: The effect of frequency and clearance variations on single-degree-of-freedom impact oscillators, J. Sound Vib. **184**, 475–502 (1995)
5. Chillingworth, D.R.J.: Discontinuity geometry for an impact oscillator. Dyn. Syst. **17**, 389–420 (2002)
6. Chin, W., Ott, E., Nusse, H.E., Grebogi, C.: Grazing bifurcation in impact oscillators. Phys. Rev. E **50**, 4427–4444 (1994)
7. Dankowicz, H., Jerrelind, J.: Control of near-grazing dynamics in impact oscillators. Proc. R. Soc. Lond. A **461**, 3365–3380 (2005)
8. Ivanov, A.P.: Impact oscillations: linear theory of stability and bifurcations. J. Sound Vib. **178**, 361–378 (1994)
9. Küpper, T., Hosham, H.A., Weiss, D.: Non-smooth dynamical systems via reduction methods. In: Johann, A. (ed.) Recent Trends in Dynamical Systems. Springer Proceedings in Mathematics & Statistics 2013
10. Molenaar, J., de Weger, J.G., van de Water, W.: Mappings of grazing-impact oscillators. Nonlinearity **14**, 301–321 (2001)
11. Nordmark, A.B.: Non-periodic motion caused by grazing incidence in an impact oscillator. J. Sound Vib. **145**, 279–297 (1991)
12. Nordmark, A.B.: Universal limit mapping in grazing bifurcations. Phys. Rev. E **55**, 266–270 (1997)
13. Nordmark, A.B.: Existence of periodic orbits in grazing bifurcations of impacting mechanical oscillators. Nonlinearity **14**, 1517–1542 (2001)
14. Stewart, D.: Rigid-body dynamics with friction and impact. SIAM Rev. **42**, 3–39 (2000)
15. Whiston, G.W.: Singularities in vibro-impact dynamics. J. Sound Vib. **152**, 427–460 (1992)
16. Zhao, X., Dankowicz, H.: Unfolding degenerate grazing dynamics in impact actuators. Nonlinearity **19**, 399–418 (2006)

Chapter 3
Branches of Periodic Orbits in Reversible Systems

André Vanderbauwhede

Abstract In the typical reversible systems which appear in many applications (symmetric) periodic solutions appear in one-parameter families. In this short survey we describe how these branches of periodic orbits originate from equilibria, terminate at homoclinic orbits, and branch from each other in period-doubling bifurcations or higher order subharmonic bifurcations. Adding external parameters allows to study degenerate cases and the transition from degenerate to non-degenerate situations.

3.1 Introduction

It is a kind of popular statement—initiated by Henri Poincaré himself—that in Hamiltonian systems the subset of periodic orbits forms a sort of backbone for the full dynamics of the system. What may be less well known is that in reversible systems the subset of symmetric periodic orbits shows a similarly rich behavior as its analogue in the Hamiltonian case. Of course, in many classical applications reversible systems are also Hamiltonian (and vice-versa), but reversible systems can be studied in their own right; actually, in many cases results for reversible systems are somewhat easier to obtain, only relying on symmetry properties and avoiding symplectic structures.

In this brief note we survey some of the basic branching behavior of symmetric periodic orbits in reversible systems. For simplicity we restrict to the simplest case of reversibility, and our statements will be rather descriptive instead of technically complete. Even then, keeping in mind that a simple picture is worth more than a thousand words, we would like to urge the interested reader to download the

A. Vanderbauwhede (✉)
Department of Pure Mathematics, Ghent University, Krijgslaan 281, 9000 Gent, Belgium
e-mail: avdb@cage.ugent.be

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_3,

slides of our RTDS talk on this topic [17] and keep them at hand as we move forward.

The results which we describe are either very classical, or based on joint work we did during the last decade (or longer) in collaboration with Jürgen Knobloch, Bernold Fiedler, Maria-Cristina Ciocci, Francisco Javier Muñoz Almaraz, Jorge Galán, Emilio Freire, and Sebius Doedel. More details can be found in the papers [1–3, 5, 6, 10–14] for the older results, [7, 8] for more recent work, and [9, 15, 16] for more general results on the Lyapunov–Schmidt reduction which is in the background of our approach.

3.2 Reversible Systems

Consider a smooth finite-dimensional system

$$\dot{x} = F(x), \qquad (x \in \mathbb{R}^n,\ F : \mathbb{R}^n \to \mathbb{R}^n \text{ smooth}) \tag{3.1}$$

and the corresponding flow $\tilde{x} = \tilde{x}(t; x)$. We say that such system is ***reversible*** if there exist a closed subgroup $\Gamma \subset GL(n; \mathbb{R})$ and a nontrivial group homomorphism $\chi : \Gamma \to \{1, -1\}$ such that

$$gF(x) = \chi(g)F(gx), \qquad \forall g \in \Gamma,\ \forall x \in \mathbb{R}^n. \tag{3.2}$$

It follows immediately that

$$g\tilde{x}(t; x) = \tilde{x}(\chi(g)t; gx), \qquad \forall g \in \Gamma;\ \forall t \in \mathbb{R},\ \forall x \in \mathbb{R}^n. \tag{3.3}$$

Here we will restrict to the simplest possible case where $\Gamma = \{I, R\}$, with $R \in \mathscr{L}(\mathbb{R}^n)$ a linear involution (i.e., $R^2 = I$) and $\chi(R) = -1$; the reversibility condition (3.2) then reduces to

$$RF(x) = -F(Rx), \qquad \forall x \in \mathbb{R}^n; \tag{3.4}$$

the flow satisfies

$$R\tilde{x}(t; x) = \tilde{x}(-t; Rx), \qquad \forall t \in \mathbb{R},\ \forall x \in \mathbb{R}^n. \tag{3.5}$$

We also assume that

$$n = 2N \quad \text{and} \quad \dim(\text{Fix}(R)) = N. \tag{3.6}$$

The prototype example of such reversible system is given by the scalar second order equation

$$\ddot{y} + f(y) = 0, \qquad (y \in \mathbb{R},\ f : \mathbb{R} \to \mathbb{R} \text{ smooth}). \tag{3.7}$$

For this equation we have $N = 1$, $x = (y, \dot{y})$ and $Rx = R(y, \dot{y}) = (y, -\dot{y})$. Depending on f the phase plane typically consists of a succession of centers and saddle points along the y-axis. For the particular case $f(y) = y(1 - y)$ we get a center at $y = 0$ and a saddle point at $y = 1$; there is a homoclinic orbit attached to the saddle, and the region inside this homoclinic is filled with period orbits. The full picture is symmetric under R.

We are particularly interested in ***symmetric*** orbits, i.e., orbits $\gamma = \{\tilde{x}(t; x_0) \mid t \in \mathbb{R}\}$ such that $R(\gamma) = \gamma$. It is easy to show that an orbit γ is symmetric if and only if $\gamma \cap \text{Fix}(R) \neq \emptyset$. Choosing x_0 to be one of the intersection points we have then that $R\tilde{x}(t; x_0) = \tilde{x}(-t; x_0)$. Excluding symmetric equilibria an orbit γ is ***symmetric*** and ***periodic*** if and only if $\gamma \cap \text{Fix}(R) = \{x_0, x_1\}$ for two distinct points $x_0 \neq x_1$; the minimal period of such symmetric periodic orbit equals two times the time needed to travel from x_0 to x_1: if $T > 0$ is such that $x_1 = \tilde{x}(T; x_0)$ and $\tilde{x}(t; x_0) \neq x_1$ for $0 \leq t < T$ then the minimal period is $2T$. It follows that symmetric periodic orbits are generated by the intersection points of the N-dimensional subspace $\text{Fix}(R)$ with the $(N + 1)$-dimensional manifold $\{\tilde{x}(t; x) \mid t \in \mathbb{R},\ x \in \text{Fix}(R)\}$; therefore symmetric periodic orbits typically appear in ***one-parameter families***.

It is then a natural question to ask how these one-parameter families of symmetric periodic orbits start, finish, and/or branch from each other. The simple example above already shows a partial answer. The center with the surrounding periodic orbits is an example of the ***Lyapunov Center Theorem*** for reversible systems: under appropriate technical conditions each equilibrium whose linearization has a pair of simple purely imaginary eigenvalues $\pm i\omega_0$ ($\omega_0 > 0$) is contained in a two-dimensional invariant manifold filled with symmetric periodic orbits whose minimal period converges to $2\pi/\omega_0$ as one approaches the equilibrium. At the other end the period tends to infinity, and the periodic orbits tend to the symmetric homoclinic orbit. This is an example of a ***period blow-up***: again under appropriate technical conditions, each non-degenerate symmetric homoclinic orbit to a hyperbolic fixed point is the limit of a one-parameter family of symmetric periodic orbits whose minimal period tends to infinity as one approaches the homoclinic.

These are essentially the only phenomena one can see for $N = 1$; of course many more possibilities arise for $N > 1$. As a guiding example we can consider the following system (with f as before):

$$\ddot{y} + f(y) = z, \qquad \ddot{z} + g(z) = 0, \qquad (g(0) = 0,\ g'(0) = 1). \tag{3.8}$$

Here $N = 2$, $x = (y, \dot{y}, z, \dot{z})$ and $Rx = R(y, \dot{y}, z, \dot{z}) = (y, -\dot{y}, z, -\dot{z})$. A particular question we can ask about this example (and which we will address further on) is: are there, close to the symmetric periodic orbits for $z = 0$, any other periodic orbits with $z \neq 0$ but small? Also, adding external parameters may help to see certain transitions from more degenerate to less degenerate situations.

3.3 Reversible Hopf Bifurcation

One of the main bifurcations appearing in general systems is the Hopf bifurcation, where an equilibrium looses stability under the change of an external parameter because a pair of simple eigenvalues of the linearization crosses the imaginary axis; as a result a small periodic orbit bifurcates from the equilibrium. In reversible systems such crossing of the imaginary axis is not possible. Indeed, consider a reversible system depending on a scalar parameter $\lambda \in \mathbb{R}$:

$$\dot{x} = F(x, \lambda), \qquad (RF(x, \lambda) = -F(Rx, \lambda), \ \forall \lambda); \tag{3.9}$$

assume also that $x = 0$ is an equilibrium for all λ ($F(0, \lambda) = 0$), and let $A_\lambda := D_x F(0, \lambda)$. It is easy to see that if $\mu \in \mathbb{C}$ is an eigenvalue of A_λ, then so is $-\mu$. As a consequence, if A_λ has a pair of simple eigenvalues on the imaginary axis, then this pair of eigenvalues can not be moved off the imaginary axis under a change of the parameter λ. The only way eigenvalues can be moved off the imaginary axis is by two pairs of purely imaginary eigenvalues meeting each other and then splitting off into the complex plane, two to the left, two to the right. This is precisely the situation described by the ***reversible Hopf bifurcation***. A similar situation appears in Hamiltonian systems at a so-called ***Hamiltonian Hopf bifurcation***.

So assume that for $\lambda < 0$ the linearization A_λ has two pairs of simple purely imaginary eigenvalues which coalesce for $\lambda = 0$ in a pair of (non-semisimple) purely imaginary eigenvalues, and such that for $\lambda > 0$ we have a quadruple of eigenvalues $\pm\alpha \pm i\beta$ ($\alpha > 0$, $\beta > 0$). For $\lambda < 0$ but small the two pairs of purely imaginary eigenvalues satisfy the non-resonance condition required by the reversible Lyapunov Center Theorem, and therefore we have for such λ two branches of symmetric periodic orbits originating from the equilibrium at $x = 0$. Detailed analysis shows that as λ goes to zero and becomes positive two scenarios are possible. In the first scenario the two families which exist for $\lambda < 0$ are locally connected into a single family which shrinks to the equilibrium as λ approaches zero and then disappears for $\lambda > 0$. In the second scenario the two families (for $\lambda < 0$) are not locally connected but become tangent to each other as λ goes to zero, and then connect to each other into a single family detached from the equilibrium for $\lambda > 0$. A detailed analysis of the associated dynamics can be found for example in Section 4.3.3 of [4]; the first scenario mentioned above corresponds to the "defocusing case," while the second scenario corresponds to the "focusing case,"

3.4 Generic Subharmonic Branching

Next we turn our attention to the example system (3.8). We assume that the unperturbed system (for $z = 0$) has a family of symmetric periodic orbits (as explained in Sect. 3.2), parametrized by some parameter $\alpha \in \mathbb{R}$; we call this the

primary family. Denote the minimal period along this family by $T(\alpha)$. Now suppose that at $\alpha = \alpha_0$ there is a bifurcation of another branch of periodic orbits, with $z \neq 0$, and parametrized by $\rho \in \mathbb{R}$; denote the minimal period along this branch by $\tilde{T}(\rho)$. Then the limiting period along the bifurcating branch must be an integer multiple of the period at the branching point along the primary family, i.e., there must be some $q \in \mathbb{N}$ such that $\lim_{\rho\to 0} \tilde{T}(\rho) = qT(\alpha_0)$. In a more general situation this ***resonance condition*** translates in the fact that a symmetric period orbit can only be at a branching point of periodic solutions if it has a pair of multipliers which are roots of unity, i.e., of the form

$$\exp(\pm i\theta_0), \qquad \text{with } \theta_0 = 2\pi p/q, \;\; p, q \in \mathbb{N}, \; \gcd(p, q) = 1. \tag{3.10}$$

One of the possible approaches to study the branching of periodic orbits at such resonant periodic orbit is by using the Poincaré map. More precisely, one considers a Poincaré map $P : \Sigma \to \Sigma$ where the section Σ is chosen to be R-invariant; clearly $\dim(\Sigma) = 2N - 1$. Putting the origin on Σ at the intersection point with the resonant periodic orbit we have then $P(0) = 0$ and (because of the reversibility) $RPR = P^{-1}$; as a consequence 1 is always an eigenvalue of $DP(0)$, with odd (algebraic) multiplicity ≥ 1. Further on we will assume that 1 is a ***simple*** eigenvalue of $DP(0)$ (generically this will be the case). The resonance condition then implies that $DP(0)$ has a pair of eigenvalues of the form (3.10), with $0 < p < q$ and $q \geq 2$. The problem is then to study the bifurcation of q-periodic points of P from the fixed point at the origin. This can be done by an appropriate Lyapunov–Schmidt reduction.

We should also mention the important property that the reversibility implies that if $\mu \in \mathbb{C}$ is a multiplier of a symmetric periodic orbit then so is μ^{-1}; in particular, if a symmetric periodic orbit (along a branch of such symmetric periodic orbits) has a pair of simple multipliers on the unit circle, then by moving along the branch these multipliers have to stay on the unit circle.

3.4.1 Period Doubling

For the case $q = 2$ we assume that -1 is an eigenvalue of $DP(0)$ with geometric multiplicity ***one*** and algebraic multiplicity ***two***. The problem of finding period-doubled solutions then reduces to a three-dimensional bifurcation equation with $\mathbb{D}_2$-symmetry: the dimension equals the dimension of the generalized kernel of $DP^2(0)$, the $\mathbb{D}_2$-symmetry from a combination of the reversibility with the $\mathbb{Z}_2$-symmetry which originates from the fact that if x is a solution of $P^2(x) = x$, then so is $P(x)$.

Denoting the coordinates on the three-dimensional "bifurcation space" by (α, ξ, η) and using the symmetries the bifurcation equations reduce to

$$\xi\varphi(\alpha, \xi^2) = 0 \qquad \text{and} \qquad \eta = 0, \tag{3.11}$$

where φ is a scalar function with $\varphi(0,0) = 0$. This gives us immediately the solution branch $\{(\alpha, 0, 0) \mid \alpha \in \mathbb{R}\}$, corresponding to the primary branch. If we assume the ***transversality condition*** $\partial\varphi/\partial\alpha(0,0) \neq 0$ then we get a branch of ***period-doubled solutions*** $\{(\alpha^*(\xi^2), \xi, 0) \mid \xi > 0\}$ (we only have to consider $\xi > 0$ since ξ and $-\xi$ correspond to the same 2-periodic orbit of P). The transversality condition means that as we move along the primary branch a pair of simple multipliers moves toward -1 on the unit circle, coalesces at -1, and then splits on the real axis in a generic way.

3.4.2 Subharmonic Branching

For $q \geq 3$ we assume that the resonant multipliers (3.10) are ***simple multipliers***. The Lyapunov–Schmidt reduction for the bifurcation of q-periodic points of P results in a three-dimensional problem with a $\mathbb{D}_q$-symmetry. The three-dimensional bifurcation space corresponds to the three-dimensional kernel of $DP^q(0)$, which itself corresponds to the three simple eigenvalues 1 and $\exp(\pm i\theta_0)$ of $DP(0)$; it is convenient to denote the coordinates in this space by $u = (\alpha, z) = (\alpha, \rho\exp(\pm i\theta)) \in \mathbb{R} \times \mathbb{C}$. The symmetry is generated by the reversibility and by the implicit $\mathbb{Z}_q$-symmetry of the problem: if $x \in \Sigma$ generates a q-periodic point of P, then so do $P(x)$, $P^2(x)$, …, $P^{q-1}(x)$ and $P^q(x) = x$. On the bifurcation space the $\mathbb{D}_q$-symmetry is generated by $S_0 u = S_0(\alpha, z) = (\alpha, \exp(i\theta_0)z)$ and $Ru = R(\alpha, z) = (\alpha, \bar{z})$.

Using the symmetry one finds that the bifurcation equations take the form

$$\mathscr{B}(u) = \left(h_0(u)\mathrm{im}(z^q), ih_1(u)z + ih_2(u)\bar{z}^{q-1}\right) = 0, \tag{3.12}$$

where the functions $h_i : \mathbb{R} \times \mathbb{Z} \to \mathbb{R}$ $(i = 0, 1, 2)$ are smooth, with $h_1(0) = 0$ and $h_i(u) = h_i(S_0 u) = h_i(Ru)$ for all $u \in \mathbb{R} \times \mathbb{C}$ and $i = 0, 1, 2$. There are two immediate consequences. First, we have $\mathscr{B}(\alpha, 0) = 0$ for all $\alpha \in \mathbb{R}$, giving the solution branch $\{(\alpha, 0) \mid \alpha \in \mathbb{R}\}$ corresponding to the primary branch. Second, if either $h_0(0,0) \neq 0$ or $h_2(0,0) \neq 0$, then nontrivial solutions (with $z \neq 0$) can only exist if (for $z = \rho\exp(i\theta)$)

$$\mathrm{im}(z^q) = \rho^q \sin(q\theta) = 0. \tag{3.13}$$

Since all $S_0^j u = S_0(\alpha, z) = (\alpha, \exp(ij\theta_0)z)$ $(j \in \mathbb{Z})$ generate the same q-periodic orbit of P we have to consider only two cases: $\theta = 0$ and $\theta = \pi/q$, both with $\rho > 0$. For these the bifurcation equations reduce to a single scalar equation, respectively

$$b_+(\alpha, \rho) = h_1(\alpha, \rho) + \rho^{q-2}h_2(\alpha, \rho) = 0 \tag{3.14}$$

and

$$b_-(\alpha,\rho) = h_1(\alpha,\rho\exp(i\pi/q)) - \rho^{q-2}h_2(\alpha,\rho\exp(i\pi/q)) = 0. \qquad (3.15)$$

We have already mentioned that $h_1(0,0) = 0$; therefore, assuming the ***transversality condition*** $\partial h_1/\partial\alpha(0,0) \neq 0$ we obtain from (3.14) and (3.15) two solution branches $\{(\alpha_+^*(\rho),\rho) \mid \rho \geq 0\}$ and $\{(\alpha_-^*(\rho),\rho\exp(i\pi/q)) \mid \rho \geq 0\}$, respectively. For $q \geq 5$ these two branches are very close to each other, as the two sides of an Arnol'd tongue: if $\alpha_+^*(\rho_+) = \alpha_-^*(\rho_-) = d$ for some small $d > 0$, then $|\rho_+ - \rho_-| = \mathrm{O}\left(d^{\frac{d-2}{2}}\right)$. Also, the solutions along the two branches of subharmonics are symmetric.

A further calculation shows that for $q \geq 5$ and if $h_2(0) \neq 0$), then the solutions along one of the two branches are ***elliptic*** and those along the other branch ***hyperbolic***, in the following sense. The solutions along both branches have 4 multipliers close to 1; two of these (counting multiplicities) are exactly equal to 1; along the elliptic branch the two other multipliers are on the unit circle, while along the hyperbolic branch they are on the real axis—one inside and one outside of the unit circle.

For $q = 3$ and $q = 4$ (the so-called weak resonant cases) the second term in $b_+(\alpha,\rho)$ and $b_-(\alpha,\rho)$ is of order ρ, respectively ρ^2, and hence interferes with the order ρ^2 term in h_1; therefore these cases require a separate analysis which we do not comment on any further in this note.

The attentive reader may notice here a certain resemblance with the analysis of subharmonic bifurcations in dissipative systems, in particular the appearance of Arnol'd tongues and the distinction between weak and strong resonances. We should warn that this resemblance is purely formal, since the reversible and the dissipative cases are widely different. In dissipative systems subharmonic bifurcation is a codimension two phenomenon, while in reversible (and Hamiltonian) systems it is codimension zero, i.e., it appears robustly in fixed systems. In dissipative systems the Arnol'd tongues are regions in ***parameter space***; points inside or at the border of the tongues correspond to systems having subharmonic solutions. In the reversible case which we discuss here the "tongues" are just two curves in a Poincaré section of ***phase space*** the points of which correspond to subharmonic solutions (in particular, the points "inside" the tongues do not correspond to periodic orbits).

In order to get the generic picture sketched above (two bifurcating branches of subharmonics, one elliptic and one hyperbolic, and close to each other as the two sides of an Arnol'd tongue) we need next to $h_2(0) \neq 0$ (which is essentially a condition on some higher order terms in the normal form) basically two conditions: (a) a pair of ***simple*** resonant multipliers of the form (3.10), and (b) the transversality condition $\partial h_1/\partial\alpha(0,0) \neq 0$. This transversality condition means that as we move along the primary branch (using α as a parameter) a pair of simple multipliers crosses the critical pair (3.10) with nonzero speed. This brings us to the question what happens when one of these conditions is not satisfied; the answer forms the subject of the next section.

3.5 Degenerate Subharmonic Branching

First we briefly consider the case where we have a symmetric periodic orbit with a resonant pair of multipliers which are not simple. More precisely, we assume that we have a symmetric periodic orbit with a resonant pair of multipliers (3.10) (for simplicity we will restrict to $q \geq 5$) with geometric multiplicity one but algebraic multiplicity two (this requires $N \geq 3$). Typically this means that as we move along the primary branch to which the critical periodic orbit belongs two pair of simple multipliers move along the unit circle and meet each other at the critical pair (3.10); then they split off the unit circle, with one pair inside and one pair outside the unit circle. Because of the requirement that the two pairs meet exactly at the resonant pair (3.10) this is a codimension one situation. Adding an appropriate external parameter $\lambda \in \mathbb{R}$ we get a situation where for $\lambda < 0$ the meeting point of the two pairs of multipliers is at one side of the critical pair (3.10), while for $\lambda > 0$ it is at the other side. This means that both for $\lambda < 0$ and $\lambda > 0$ we have a non-degenerate situation, with a pair of multipliers crossing the resonant value with nonzero speed. So, both for $\lambda < 0$ and for $\lambda > 0$ we have two branches of subharmonics bifurcating from the primary branch; obviously, one then expects the same to happen for $\lambda = 0$, and this is precisely what comes out of a (rather lengthy) detailed analysis.

A more interesting situation arises when the resonant pair of multipliers (3.10) is simple, but the transversality condition is not satisfied. This is a situation which can already arise in our example system (3.8) when along the primary family the (minimal) period reaches a local maximum or minimum. This maximum or minimum has to happen precisely at a periodic orbit for which the resonance condition is satisfied, so again this is a codimension one situation. To give a full description we add an external parameter $\lambda \in \mathbb{R}$; for example, in (3.8) we replace $f(y)$ by $f(y, \lambda)$ where $f : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ satisfies appropriate technical conditions such that for the primary family given by the unperturbed equation $\ddot{y} + f(y, \lambda) = 0$ the following situation arises. For fixed $\lambda < 0$ a pair of simple multipliers moves along the unit circle, passes the resonant pair (3.10), turns around a little bit further on, and passes again the resonant pair. As λ increases toward zero the turning point comes closer to the resonant pair, for $\lambda = 0$ the turning point is exactly at the resonant pair, while for $\lambda > 0$ the turning happens before the resonant pair is reached. This means that for $\lambda < 0$ we have two generic subharmonic bifurcations close to each other along the primary family, while for $\lambda > 0$ there is no such subharmonic bifurcation.

To study this degenerate case we can use the same Lyapunov–Schmidt reduction as before; we get a three-dimensional reduced problem of the form (3.12) which reduces via (3.13) to two scalar bifurcation equations of the form

$$b_{\pm}(\alpha, \rho, \lambda) = A\lambda + B\rho^2 + C\alpha^2 + \text{h.o.t.} = 0; \tag{3.16}$$

the two bifurcation functions $b_+(\alpha, \rho, \lambda)$ and $b_-(\alpha, \rho, \lambda)$ only differ in the higher terms (again we assume $q \geq 5$). Assuming $ABC \neq 0$ we obtain two possible

bifurcation scenarios, depending on the sign of BC; we assume $AC > 0$. For $BC > 0$ we get the ***banana scenario***, where the two branches of subharmonics (one elliptic, one hyperbolic) which for $\lambda < 0$ start at each of the two bifurcation points are locally connected, shrink as λ approaches zero, and disappear for $\lambda \geq 0$. For $BC < 0$ we get the ***banana split scenario***: the two times two subharmonic branches which exist for $\lambda < 0$ are not locally connected, they become tangent to each other for $\lambda = 0$, and for $\lambda > 0$ connect to each other and detach from the primary branch, giving two branches of subharmonics, one elliptic and the other hyperbolic.

To illustrate the result one can look at the four-dimensional system

$$\ddot{y} + \left(y + y^2 + \gamma y^3\right) = z, \qquad \ddot{z} + \omega^2 z = 0; \tag{3.17}$$

this system has two parameters γ and ω. For $\gamma = 0.3$ the period map for the unperturbed system shows a maximum. For $\omega = 0.4003$ one finds two local branches of 5-subharmonics, very close to each other and forming the boundary of a very thin banana. However, and contrary to the theoretical results we have just explained, it is not so that one of the two branches is elliptic and the other hyperbolic. Instead, along each of the two branches we see a transition from elliptic to hyperbolic. The explanation of this deviation from the theory brings us to our last section.

3.6 Change of Stability Without Bifurcation

Our example (3.17) is not fully generic, in the sense that not all the conditions in order to get the banana scenario as described before are satisfied. The reason for this nongenericity stems from the fact that the system (3.17) has a first integral, namely the (square of the) amplitude of the forcing equation: $H = z^2 + \omega^2 \dot{z}^2$. It is also clear that this first integral reaches a maximum along each of the two branches of subharmonics. And for such situation one can prove the following general result: when along a branch of symmetric periodic orbits in a reversible system with a first integral this integral reaches either a maximum or a minimum, then at the critical periodic orbit one will typically see a change of stability, from elliptic to hyperbolic or vice versa. This result forms a counterexample to the "folk theorem" in bifurcation theory which says that a change of stability leads to bifurcation. Here we get a single branch of periodic orbits along which we see a change of stability, from elliptic to hyperbolic, without any further bifurcation.

It is easy to change our earlier banana and banana split scenarios for this nongeneric situation with a first integral. Finally, the example system

$$\ddot{y} + \left(y + y^2 + \gamma y^3\right) = z, \qquad \ddot{z} + \left(\omega^2 + \epsilon y\right) z = 0 \tag{3.18}$$

allows to visualize (numerically) the transition from the nongeneric situation (for $\epsilon = 0$) to the generic one. Increasing ϵ from zero one sees how along the two sides of the banana the transition points between elliptic and hyperbolic subharmonics start to shift and finally disappear (for sufficiently small bananas). The theoretical study of this transition is still underway...

Acknowledgements The more recent work reported in this note has been supported by the University of Sevilla and the MICIIN/FEDER grant number MTM2009-07849.

References

1. Ciocci, M.-C., Vanderbauwhede, A.: On the bifurcation and stability of periodic orbits in reversible and symplectic diffeomorphisms. In: Degasperis, A., Gaeta, G. (eds.) Symmetry and Perturbation Theory, pp. 159–166. World Scientific, Singapore (1999)
2. Ciocci, M.-C.: Bifurcation of periodic orbits and persistence of quasi-periodic orbits in families of reversible systems. PhD. Thesis, University of Ghent (2003)
3. Ciocci, M.-C., Vanderbauwhede, A.: Bifurcation of periodic points in reversible diffeomorphisms. In: Elaydi, S., Ladas, G., Aulbach, B. (eds.) New Progress in Difference Equations—Proceedings ICDEA2001, Augsburg, pp. 75–93. Chapman & Hall/CRC, Boca Raton (2004)
4. Haragus, M., Iooss, G.: Local Bifurcations, Center Manifolds, and Normal Forms in Infinite-Dimensional Dynamical Systems. Universitext Series. Springer, Berlin (2010)
5. Knobloch, J., Vanderbauwhede, A.: Hopf bifurcation at k-fold resonances in reversible systems. Unpublished Preprint. Available at http://cage.ugent.be/~avdb/articles/k-fold-Res-Rev.pdf (1995)
6. Knobloch, J., Vanderbauwhede, A.: A general reduction method for periodic solutions in conservative and reversible systems. J. Dyn. Differ. Equ. **8**, 71–102 (1996)
7. Muñoz Almaraz, F.J., Freire, E., Galán, J., Vanderbauwhede, A.: Change of stability without bifurcation: an example. In: Bohner, M., Došlá, Z., Ladas, G., Ünal, M., Zafer, A. (eds.) Difference Equations and Applications—Proceedings of the ICDEA 2008 Conference, Istanbul, 2008, pp. 3–18. Bahçeşehir University Publications, Istanbul (2009)
8. Muñoz Almaraz, F.J., Freire, E., Galán, J., Vanderbauwhede, A.: Non-degenerate and degenerate subharmonic bifurcation in reversible systems (2012, personal notes)
9. Takens, F., Vanderbauwhede, A.: Local invariant manifolds and normal forms. In: Broer, H.W., Hasselblatt, B., Takens, F. (eds) Handbook of Dynamical Systems, vol. 3, pp. 89–124. Elsevier, Amsterdam (2010)
10. Vanderbauwhede, A.: Subharmonic branching in reversible systems. SIAM J. Math. Anal. **21**, 954–979 (1990)
11. Vanderbauwhede, A., Fiedler, B.: Homoclinic period blow-up in reversible and conservative systems. Z. Angew. Math. Phys. **43**, 292–318 (1992)
12. Vanderbauwhede, A.: Branching of periodic solutions in time-reversible systems. In: Broer, H., Takens, F. (eds.) Geometry and Analysis in Non-linear Dynamics. Pitman Research Notes in Math., vol. 222, pp. 97–113. Longman Scientific & Technical, Harlow (1992)
13. Vanderbauwhede, A.: Branching of periodic orbits in Hamiltonian and reversible systems. In: Agarwal, R.P., Neuman, F., Vosmansky, J. (eds.) Proceedings Equadiff 9 (Brno 1997), pp. 169–181. Masaryk University, Brno (1998)
14. Vanderbauwhede, A.: Subharmonic bifurcation at multiple resonances. In: Elaydi, S., Allen, F., Elkhader, A., Mughrabi, T., Saleh, M. (eds.) Proceedings of the Mathematics Conference (Birzeit, August 1998), pp. 254–276. World Scientific, Singapore (2000)

15. Vanderbauwhede, A.: Lyapunov–Schmidt method for dynamical systems. In: Meyers, R.A. (ed.) Encyclopedia of Complexity and System Science, pp. 5299–5315. Springer, New York (2009)
16. Vanderbauwhede, A.: Lyapunov–Schmidt—a discrete revision. In: Elaydi, S., Oliveira, H., Ferreira, J.M., Alves, J.F. (eds.) Discrete Dynamics and Difference Equations—Proceedings of the ICDEA 2007 Conference, Lisbon, pp. 120–143. World Scientific, Singapore (2010)
17. Vanderbauwhede, A.: Branches of Periodic Orbits in Reversible Systems. Slides of the RTDS 2012 Lecture. Download available at http://cage.ugent.be/~avdb/english/talks/RTDS2012.pdf (2012)

Chapter 4
Canard Explosion and Position Curves

Freddy Dumortier

Dedicated to the 60th birthday of Juergen Scheurle

Abstract The paper deals with smooth two-dimensional singular perturbation problems. Attention goes to the canard explosion for a generic Hopf or jump breaking mechanism.

We introduce the notion of position curve and study the typical shape of a position curve. We also study catastrophes of limit cycles, obtaining results in ε-uniform neighbourhoods, both in phase space and in parameter space.

4.1 Introduction

The "canard explosion" has first been detected and studied using non-standard analysis (see e.g. [1, 2]). A description by means of standard techniques, more precisely by "geometric singular perturbation theory", has been provided for the Van der Pol case in [11]. A treatment of a slightly more general case using the techniques introduced in [11] can be found in [16].

Essentially, in the papers mentioned above, for small values of the small parameter ε and up to a time reversal, a limit cycle is born in a Hopf bifurcation and grows monotonically until getting the shape of a traditional relaxation oscillation with two fast and two slow movements. All this happens, when changing the "breaking parameter" (see [4]) in a regular way. The expansion of the limit cycle starts slowly, and then suddenly "explodes", while changing its shape from a small amplitude "round" form to its final form. This is what happens in the "most generic" case of canard explosion. Both in [11] and in [16] this is well described, based on

F. Dumortier (✉)
Hasselt University, Campus Diepenbeek, Agoralaan - Gebouw D, 3590 Diepenbeek, Belgium
e-mail: freddy.dumortier@uhasselt.be

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_4,

"blow up" and the use of center manifolds. During the complete process the limit cycle stays unique. For the Van der Pol case (see [11]) this has been proven in [3] (see also [10]). This fact is taken for granted in [16] but is not proven there. In the meantime a proof has been given in [8], based on the theory developed in [14].

Krupa and Szmolyan [16] also have a case in which a saddle node bifurcation of limit cycles occurs during the explosion. The figure that is made of the position of the limit cycles is qualitatively correct, but is quantitatively in contradiction with [1], where it is proven that, during the explosion, one can only have one so-called flying (i.e. rapidly changing in size) canard, the other remaining more or less unchanged in shape, or so-called sitting. The terminology comes from [1]. The occurrence of elementary catastrophes of any order, during some canard explosions, has been studied in [12]. It has not yet been analyzed precisely how this can match with the observation on the unicity of the "flying" canard. Moreover in [4] the difference is studied between the order of the catastrophes when changing the breaking parameter, versus a situation in which the canard passage remains unbroken. In both cases all bifurcations occur near a zero of the "slow divergence integral". In the latter case the order of the catastrophe near a zero is exactly given by the multiplicity of this zero. In the former case, and under a regular change of the "breaking parameter", the order of the catastrophe is one unit higher, involving the occurrence of one more limit cycle than the number predicted by the multiplicity of the zero of the slow divergence integral.

We will now show how this can be in agreement with [1], as well as in agreement with the upper bound on the number of limit cycles, as proven in [12].

At this point we provide a quite stronger result than the one given in [12]. In [12] the occurrence of the catastrophe of higher codimension is proven inside a "domain" that shrinks when $\varepsilon \downarrow 0$. The results of [12] do not apply to a full tubular neighbourhood of the slow fast cycle at which the slow divergence integral is zero, neither to a full neighbourhood of the parameter value at which the situation occurs. In this paper we will obtain results that are valid in a full neighbourhood in the product of the phase space and the parameter space. For a precise statement we refer to Theorem 4.1 that we will formulate at the end of Sect. 4.2.

Exactly like in [4], the study we make also applies to the so-called jump breaking mechanism that has first been studied in [13].

We intend to rely heavily on [4], in order not to have to repeat the notions that have been introduced in full detail there. In Sect. 4.2.1 we will only recall some essential elements in order to be able to state the results and start the proofs. We will also not repeat the proofs that have been given in [4] when using the results from [4].

Section 4.2.2 of this paper can be considered as a natural continuation of [4], stressing some facts about the canard explosion that might lead to misinterpretation if not dealt with carefully.

In Sect. 4.2.3 we introduce the notion of "position curve" and show how a generic position curve has a "typical shape", as represented in Fig. 4.5. For precise numerical calculation of such position curves we refer to the forthcoming paper [9], a paper that also contains other interesting numerical observations. We end Sect. 4.2 by formulating Theorem 4.1.

In Sect. 4.3 we study the typical shape of position curves.

Section 4.4 deals with the proof of Theorem 4.1 and can be considered as the most important contribution of this paper, since we obtain results on limit cycles and their bifurcations that hold in a full neighbourhood in the product of the phase space and the parameter space. The results hold near any slow–fast cycle of type FSTS for the Hopf breaking mechanism and of type FSJS for the jump breaking mechanism. The precise definition of these slow–fast cycles will be recalled in Sect. 4.2. The proof relies heavily on the structure theorems that have been proven in the papers [6, 13]. A survey of these results can also be found in [4]. The proof requires a direct treatment of differences of exponentially small functions.

In Sect. 4.5 we discuss the consequences of Theorem 4.1 as well as the remaining problems for a complete understanding of the bifurcations of limit cycles near slow–fast cycles of the type described above. Special attention is given to the role of the "flying" canard in these bifurcations.

4.2 Setting of the Problem and Statement of Results

4.2.1 Generic Breaking Mechanisms and Nearby Transition Maps

As announced in the introduction we recall the description of the generic breaking mechanisms as presented in [4]. There are two such generic breaking mechanisms, one called Hopf breaking mechanism, that has been studied in [12], and the other called jump breaking mechanism that has been studied in [13]. We consider smooth equations that can locally be given an expression

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

with f and g smooth functions, $\lambda = (\lambda_1, \ldots, \lambda_n)$, $\varepsilon > 0$ and $\varepsilon \sim 0$. In case we study planar systems, we can often use an expression as in (4.1) globally, but as we will see the results we want to present also apply to smooth systems defined on arbitrary smooth surfaces. For $\varepsilon = 0$ we call (4.1) the (λ-family of) *layer equation(s)*; it is subject to an invariant foliation (locally) given by $\{y = \text{ constant}\}$.

The set $\{f(x, y, \lambda) = 0\}$ is called the (λ-family of) *critical curve(s)*, also called *slow curve(s)*. On the slow curves can be defined a so-called slow dynamics by considering $\dot{y} = g(x, y, 0, \lambda)$ along the slow curve. In this paper we use the name *slow curve* and will restrict to connected curves; a *slow–fast orbit* will be a connected succession of (connected) slow curves and fast orbits of a layer equation, whose orientations fit. If such slow–fast orbit is closed, i.e. homeomorphic to a circle S^1, we call it a *slow–fast cycle*.

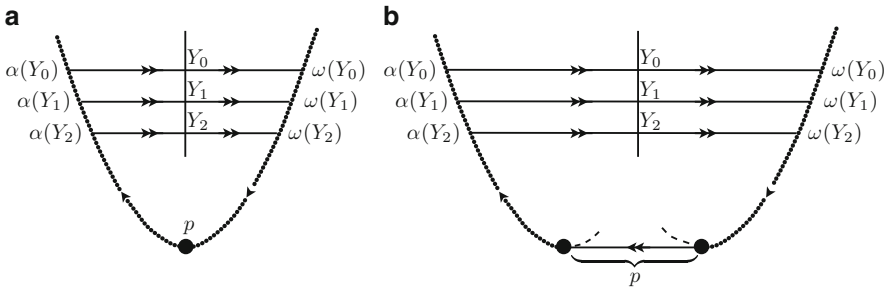

Fig. 4.1 Breaking mechanisms and layer variables

At most points of a slow curve we require a layer equation $(2.1)_{\varepsilon=0}$ to be normally hyperbolic, either attracting or repelling.

We accept this property not to hold at isolated points, that we call contact points. At such points the critical curve has a contact with a leaf of the foliation $\{y = \text{constant}\}$.

In this paper we only admit two kinds of contact points, namely the *generic jump point* and the *generic turning point*.

Definition 4.1. We say that (4.1) has at $(x, y, \lambda) = (x_0, y_0, \lambda_0)$ a generic jump point, respectively a generic turning point, if, after smooth rescaling of the variables (x, y) and rescaling of time, system (4.1) can, locally near (x_0, y_0, λ_0), be written as

$$\begin{cases} \dot{x} = y - x^2 + x^3 h_1(x, y, \varepsilon, \lambda) \\ \dot{y} = -\varepsilon(1 + x h_2(x, \lambda) + O(||(y, \varepsilon)||), \end{cases} \tag{4.2}$$

for some smooth functions h_1 and h_2, in case of a generic jump point, and as

$$\begin{cases} \dot{x} = y - x^2 + x^3 h_1(x, y, \varepsilon, \lambda) \\ \dot{y} = -\varepsilon(a(\lambda) + x + x^2 h_2(x, \lambda) + O(||(y, \varepsilon)||), \end{cases} \tag{4.3}$$

for some smooth functions h_1, h_2 and with $a(\lambda_0) = 0$ and $\lambda \mapsto a(\lambda)$ a smooth submersion at λ_0, in case of a generic turning point.

We speak of a *generic jump breaking mechanism* when in a layer equation we encounter a slow–fast cycle as represented in Fig. 4.1b, consisting of an attracting slow curve ending in a generic jump point (as in (4.2)), a fast orbit connecting this jump point to a second generic jump point, and a repelling slow curve starting at the second jump point. We denote by p the union of the two jump points and the fast orbit between them. Let us denote by γ^- (resp. γ^+) the attracting (resp. repelling) slow curve. We suppose that the slow dynamics is nonzero on $\gamma^- \cup \gamma^+$, directed towards p on γ^- and away from p on γ^+. Like in Fig. 4.1 we suppose that there exists a layer of fast orbits that is adherent to both γ^- and γ^+. We characterize the

different orbits of this fast layer by a regular parameter Y on some transverse section R (see Fig. 4.1a, b). We call Y a *layer variable* and for each orbit, characterized by $Y \in R$, we denote its ω-limit on γ^- by $\omega(Y)$ and its α-limit on γ^+ by $\alpha(Y)$. A slow–fast cycle can then be characterized by the Y-value where it cuts R. We say that such a slow–fast cycle is of type FSJS, standing for F(ast) S(low) J(ump) S(low).

Transverse to the fast orbit contained in p we also choose a segment that we denote by T and on T we choose any regular parameter U. For $\varepsilon = 0$ and for $\lambda \sim \lambda_0$ there is a fast orbit γ_1 having the first jump point as α-limit and a fast orbit γ_2 having the second jump point as ω-limit. Both γ_i cut T at a single point $h_i(\lambda)$. We suppose that

$$\lambda \mapsto h_1(\lambda) - h_2(\lambda)$$

is a submersion at $\lambda = \lambda_0$. This condition is part of the definition of a generic jump mechanism. By a smooth reparametrization we can suppose that

$$\lambda = (a, \mu),$$

where a, defined as $a = h_1(\lambda) - h_2(\lambda)$, is called a *breaking parameter*. We will from now on denote the slow–fast system as $X_{(\varepsilon,a,\mu)}$.

For $\varepsilon > 0$ we can define two transition maps from R to T by following the $X_{(\varepsilon,a,\mu)}$-orbits in respectively forward time and backward time. Let us denote the first map by $(U, \varepsilon) = (P_1(Y), \varepsilon)$ and the second map by $(U, \varepsilon) = (P_2(Y), \varepsilon)$.

Before describing the essential properties of P_1 and P_2, let us recall some definition from [6, 14].

Definition 4.2 (ε-Regularly Smoothness). We say that a function $f(z, \varepsilon)$, with $z \in I\!R^p$, for some p, is *ε-regularly smooth in z* (or ε-regularly C^∞ in z) if f is continuous and all partial derivatives of f with respect to z exist and are continuous in (z, ε).

In R we choose a smooth reference curve $\Sigma_{Y_r} = \{Y = Y_r(\varepsilon)\}$ like, e.g. $\{Y = Y_r\}$, with Y_r constant. We orientate R and T in a way that both P_1 and P_2 are orientation preserving and that we work on domains of Y-values with $Y > Y_r$.

We denote, for $\varepsilon > 0$, by $W^f_{Y_r}$ (resp. $W^b_{Y_r}$) the saturation, i.e. union of forward (resp. backward) orbits through points of Σ_{Y_r}; $W^f_{Y_r}$ and $W^b_{Y_r}$ are clearly C^∞. We also consider their continuous extension at $\varepsilon = 0$. It has been proven in [13], based on results of [7], that these continuous extensions, that we still denote as $W^f_{Y_r}$ and $W^b_{Y_r}$, cut T in curves that can be described as graphs of functions that are ε-regularly smooth in (Y, a, μ).

In [13] has also been proven that for Y with $Y > Y_r$ and $\omega(Y) \in \gamma^-$ we have following expression for P_1.

$$P_1(Y, \varepsilon, a, \mu) = f_1(\varepsilon, a, \mu) + \exp\left(\frac{1}{\varepsilon} A_1(Y, \varepsilon, a, \mu)\right), \tag{4.4}$$

where both f_1 and A_1 are ε-regularly smooth in respectively (a, μ) and (Y, a, μ), $\{U = f_1(\varepsilon, a, \mu)\} = W^f_{Y_r} \cap T$, and

$$A_1(Y, 0, a, \mu) = I_1(Y, a, \mu), \tag{4.5}$$

with I_1 the slow divergence integral calculated along $[\omega(Y), p[\subset \gamma_-$. To precisely define this slow divergence integral, we choose any regular parameter r on $\overline{\gamma^-}$, denote by $r(Y)$ the r-value of $\omega(Y)$ and by r^* the r-value of p; if we let $\dot{r} = \varphi(r, a, \mu)$ denote the slow dynamics on γ^- then I_1 is defined as

$$I_1(Y, a, \mu) = \int_{r(Y)}^{r^*} div X_{(0,a,\mu)}(r) \cdot \frac{1}{\varphi(r, a, \mu)} dr, \tag{4.6}$$

where $div X_{(0,a,\mu)}(r)$ denotes the divergence of $X_{(0,a,\mu)}$, calculated at the point of γ^- having the parameter value r.

Remark 4.1. In a plane the divergence is meant to be a divergence with respect to the standard volume form. In using these results on surfaces one can proceed like in [6].

The structure of P_2 is completely similar as in (4.4), for functions f_2 and A_2 that are ε-regularly smooth in respectively (a, μ) and (Y, a, μ), $\{U = f_2(\varepsilon, a, \mu)\} = W^b_{Y_r} \cap T$ and $A_2(Y, 0, a, \mu) = I_2(Y, a, \mu)$ as in (4.4), where $I_2(Y, a, \mu)$ is defined as in (4.6), and with r a regular parameter on $\overline{\gamma^+}$ such that $\dot{r} = \varphi(r, a, \mu)$ represents the slow dynamics on γ^+.

We speak of a *generic Hopf breaking mechanism* when in a layer equation we encounter a slow–fast orbit as represented in Fig. 4.1a, consisting of a succession of an attracting slow orbit γ^-, a generic turning point p, as defined in (4.3), and a repelling slow orbit. We suppose that the slow dynamics is nonzero on $\gamma^- \cup \gamma^+$, directed towards p on γ^- and away from p on γ^+, and that there exists a fast layer that is adherent to both γ^- and γ^+. The study of the generic Hopf breaking mechanism is similar to the study of the generic jump mechanism. We again introduce a transverse section R and a layer variable Y. We also introduce the related notions $\omega(Y)$ and $\alpha(Y)$ as before. Again we characterize the slow–fast cycles under consideration by the Y-value at which they cut R and we call them FSTS cycles, where FSTS stands for F(ast) S(low) T(urning point) S(low).

For the definition of the section T and the variable U we use, near p, the expression (4.3). However, as explained in [12], we first reparametrize λ, changing it into (a, μ), and then write

$$(\varepsilon, a) = (u^2, uA), \tag{4.7}$$

for some $u \in \mathbb{R}$ and $A \in \mathbb{R}$. We call A a *breaking parameter of the Hopf breaking mechanism*. In expression (4.3) we define $T = \{x = 0\}$ and on T we do not define a regular parameter but use U, defined as

$$y = u^2 U. \tag{4.8}$$

From (4.7) and (4.8) it is clear that both A and U can be considered as regular parameters as long as $\varepsilon > 0$, but no longer at $\varepsilon = 0$. Nevertheless, working with these "blown-up" U and A, instead of regular U and a, as we did in the jump case, is the right way to work. It permits to make exactly the same statements on the transition maps $P_i : R_i \to T$ in the Hopf case as we did for the jump case, i.e. the same results hold on $W^f_{Y_r}$, $W^b_{Y_r}$ and the related expression of the transition maps, changing ε-regular smoothness by u-regular smoothness.

Of course a function is u-regularly smooth in variables z if and only if it is ε-regularly smooth in z, but from now on we intend to use, where appropriate, the u-notions to remind the blowing up (4.7).

We will hence use

$$P_1(Y, u, A, \mu) = f_1(u, A, \mu) + \exp\left(\frac{1}{u^2} A_1(Y, u, A, \mu)\right), \tag{4.9}$$

where both f_1 and A_1 are u-regularly smooth in respectively (A, μ) and (Y, A, μ) as well as a similar expression for P_2.

Also the relation of A_1 and A_2 with the slow divergence integral as presented in (4.5) and (4.6) holds for the Hopf breaking mechanism.

The proof of the results, both in the jump and in the Hopf case, heavily relies on a blow-up of the contact points (see [12, 13]). In this paper there is no need to recall the blow-up technique, since we will only use the properties we stated; the understanding of the property does not require acquaintance with the blow-up technique.

4.2.2 Control Curves and Manifold of Closed Orbits

Continuing with the notations of Sect. 4.2.1, we restrict the layer variable Y to some segment $[Y_1, Y_2] \subset R$, that lies to the right of Y_r.

Using (4.4)–(4.6) and (4.9), we know that limit cycles of $X_{\varepsilon,a,\mu}$ are implicitly given by

$$\exp\left(\frac{1}{u^k}(A_1(Y, u, a, \mu))\right) - \exp\left(\frac{1}{u^k}(A_2(Y, u, a, \mu))\right) = f(u, a, \mu), \tag{4.10}$$

with $\varepsilon = u^k$, for some $k \in \mathbb{N}_1$ with $f(u, a, \mu) = f_2(u, a, \mu) - f_1(u, a, \mu)$ and $f(0, a, \mu) = a$. In this equation a stands for the traditional breaking parameter in the

jump breaking mechanism, but can also stand for a breaking parameter in the Hopf breaking mechanism. In the latter case a does not need to be equal to A but might be equal to $A.K(u, A, \mu)$ for some strictly positive function K that is u-regularly smooth in (A, μ). In any case we continue working with (a, μ) as new parameters.

In (4.10), in agreement with Sect. 4.2.1, we use $\varepsilon = u^k$, with $k = 3$ for the jump mechanism and $k = 2$ for the Hopf mechanism. All functions appearing in (4.10) are u-regularly smooth in their respective values (Y, a, μ) or (a, μ).

Applying the implicit function theorem to f we know that there exists a curve $a = c_0(u, \mu)$, that is u-regularly smooth in μ, with the property that $W^f_{Y_r} \cap T = W^b_{Y_r} \cap T$. We call such curve a *control curve*:

$$f(u, c_0(u, \mu), \mu) = 0.$$

Remark 4.2. We recall that $f(u, a, \mu)$ and hence also $c_0(u, \mu)$ depend on the particular choice of initial conditions Σ_{Y_r} (see [13]).

We can also apply the implicit function theorem to the full equation (4.10) inducing the existence of a function $a = c(Y, u, \mu)$, that describes all solutions of (4.10) in the region under consideration. The function c is u-regularly smooth in (Y, μ). For $u > 0$ it describes the closed orbits that cut the section R at a value $Y \in [Y_1, Y_2]$. We call it the *manifold of closed orbits*. It satisfies the equation:

$$\exp\left(\frac{1}{u^k}(A_1(Y, u, c(Y, u, \mu), \mu))\right) - \exp\left(\frac{1}{u^k}(A_2(Y, u, c(Y, u, \mu), \mu))\right) = f(u, c(Y, u, \mu), \mu). \tag{4.11}$$

A control curve is generally represented in a (u, a)-diagram. It depends on the choice of (Y_r, μ). If we fix μ, then the appropriate Y_r-family of control curves consists of curves that are exponentially close to each other. If we fix Y_r, then the appropriate μ-family of control curves can present differences that are of finite order in u.

4.2.3 Position Curves and Statement of Results

Besides representing the manifold of closed orbits (see Sect. 4.2.2) as a (Y, μ)-family of control curves in a (u, a)-diagram, it can also be interesting to represent it as a (u, μ)-family of curves in a (Y, a)-diagram. We call these curves *position curves*. For $u = 0$ a position curve is given by the Y-axis, but for $u > 0$ it expresses the value of the breaking parameter a at which the system $X_{\varepsilon,a,\mu}$ can have a closed orbit cutting R at some value Y. More interesting however is that the position curve also shows the number of limit cycles that we can encounter for a certain value a

of the breaking parameter. Of course the representation is (u, μ)-dependent with $u > 0$. Since a is the parameter controlling the breaking in the breaking mechanism we prefer to represent the position curve in a (a, Y)-diagram. We however recall that it can be represented by a graph $a(Y)$ but mostly not as a graph $Y(a)$.

In the Sects. 4.3 and 4.4 we will provide interesting results on position curves. We will see that essential information is contained in the slow-divergence integral along Γ_Y, where Γ_Y is the slow–fast cycle cutting R at the value Y. We denote this slow divergence integral by $I(Y)$, or better by $I(Y, \mu)$. The slow divergence integral $I(Y, \mu)$ is defined as

$$I(Y, u) = I_1(Y, 0, \mu) - I_2(Y, 0, \mu),$$

with I_1 and I_2 as defined in Sect. 4.2.1 (see (4.6) for I_1 and the paragraph below Remark 4.1 for I_2). When, for a fixed $(\mu, Y) = (\mu_0, Y_0)$, $I(Y_0, \mu_0) < 0$, then limit cycles γ_Y with $Y \sim Y_0$, $u \sim 0$ and $\mu \sim \mu_0$, are hyperbolically attracting in the sense that the derivative of the Poincaré map at this limit cycle is strictly negative for $u > 0$ and tends to $-\infty$ if $u \downarrow 0$. Such a "canard-type" limit cycle, in the terminology of [1], is called a *flying canard*.

In Sect. 4.3 we will explain why this terminology is well chosen and we will calculate the speed at which it flies, depending on a. We will also prove that for every value a there is at most one flying canard. This fact had first been observed by Benoit for the Hopf breaking mechanism and was studied by means of non-standard analysis.

The study of the case where $I(Y_0, \mu_0) > 0$ can be reduced to the former one, leading to similar results.

In Sect. 4.3 we will also start considering the case where $I(Y_0, \mu_0) = 0$. As we will see this is more delicate. For a number of results we simply refer to Sect. 4.3. There is however one result that we prefer to formulate as a theorem (Theorem 4.1). For the statement of Theorem 4.1 we not only rely on the notations as introduced in the previous paragraphs, but we also introduce some extra notations.

Knowing A_1 and A_2 from (4.10) we introduce

$$\tilde{A} = \tilde{A}_1 - \tilde{A}_2, \tag{4.12}$$

with

$$\tilde{A}_i(Y, u, a, \mu) = A_i(Y, u, a, \mu) + u^k \log\left(\frac{\partial A_i}{\partial Y}(Y, u, a, \mu)\right), \tag{4.13}$$

for $i = 1, 2$.

In Sect. 4.4 the function $\tilde{A}$ will be denoted as $\tilde{A}^1$, since we will also have to introduce $\tilde{A}^i$, with $i \geq 1$.

We know that

$$\tilde{A}(Y,u,a,\mu) = I(Y,\mu) + O(u) + O(a)$$

and that a similar development also holds for all partial derivatives w.r.t. (Y,μ).

We now suppose that, at some (Y_0,μ_0), the slow divergence integral I represents an elementary catastrophe of codimension n, in the sense that:

$$I(Y_0,\mu_0) = \frac{\partial I}{\partial Y}(Y_0,\mu_0) = \ldots = \frac{\partial^n I}{\partial Y^n}(Y_0,\mu_0) = 0, \frac{\partial^{n+1} I}{\partial Y^{n+1}}(Y_0,\mu_0) > 0,$$

$$\det\left(\partial(I, \frac{\partial I}{\partial Y}, \ldots, \frac{\partial^{n-1} I}{\partial Y^{n-1}})/\partial(\mu_1,\ldots,\mu_n)\right)(Y_0,\mu_0) \neq 0, \tag{4.14}$$

where we take $\mu \sim \mu_0$.

We define the sets Σ_i in (Y,u,μ)- space as:

$$\Sigma_i = \{(Y,u,\mu) : \tilde{A}(Y,u,c(Y,u,\mu),\mu) = \frac{\partial \tilde{A}}{\partial Y}(Y,u,c(Y,u,\mu),\mu) = \ldots = \frac{\partial^i \tilde{A}}{\partial Y^i}(Y,u,c(Y,u,\mu),\mu) = 0\}, \tag{4.15}$$

for $0 \leq i \leq n$.

Seen the stability of the properties (4.14) we can prove (see e.g. [17]) that, inside sufficiently small neighbourhoods of $(Y_0,0,0,\mu_0)$ in (Y,u,a,μ)-space, the Σ_i are manifolds of respective dimension $n-i+1$.

We have $\Sigma_n = \{(Y_0(u),u,c(Y_0(u),u,\mu_0(u)),\mu_0(u))\}$ for some u-regularly smooth functions $Y_0(u)$ and $\mu_0(u)$ with $(Y_0(0),\mu_0(0)) = (Y_0,\mu_0)$.

Theorem 4.1. *Consider, in a smooth two-dimensional singular perturbation problem $X_{\varepsilon,a,\mu}$, a FSTS cycle in a generic Hopf breaking mechanism or a FSJS cycle in a generic jump breaking mechanism (see Fig. 4.1). Let $a = c(Y,u,\mu)$ be the equation of the manifold of closed orbits (manifold M), with $\varepsilon = u^k$, for $k = 2$ or $k = 3$ in respectively the Hopf and the jump breaking mechanism. We suppose that the cycle occurs for some $(Y,\mu) = (Y_0,\mu_0)$. Then there exists a neighbourhood V of $(Y_0,0,\mu_0)$ in (Y,u,μ)-space and some $a_0 > 0$ such that inside $(V \times [-a_0,a_0]) \cap \{u > 0\}$ the sets*

$$M_i = \{(Y,u,\mu) : a = c(Y,u,\mu), \frac{\partial c}{\partial Y}(Y,u,\mu) = \ldots = \frac{\partial^{i+1} c}{\partial Y^{i+1}}(Y,u,\mu) = 0\}, \tag{4.16}$$

for $0 \leq i \leq n$, *are given by*

$$\{(Y, u, c(Y, u, \mu), \mu) : (Y, u, \mu) \in \Sigma_i\},$$

and hence are manifolds of respective dimension $n - i + 1$.

For $(Y, u, \mu) \in M_i \setminus M_{i+1}$, *with* $0 \leq i \leq n - 1$ *we have*

$$\frac{\partial^{i+2} c}{\partial Y^{i+2}}(Y, u, \mu).\frac{\partial^{i+1} \tilde{A}}{\partial Y^{i+1}}(Y, u, c(Y, u, \mu), \mu) > 0,$$

and for $(Y, u, \mu) \in M_n$ *we have*

$$\frac{\partial^{n+1} c}{\partial Y^{n+1}}(Y, u, \mu).\frac{\partial^{n+1} \tilde{A}}{\partial Y^{n+1}}(Y, u, c(Y, u, \mu), \mu) > 0.$$

For $(Y, u, \mu) \in M_n$ *we also have*

$$\det\left(\partial(\frac{\partial c}{\partial Y}, \ldots, \frac{\partial^n c}{\partial Y^n})/\partial(\mu_1, \ldots, \mu_n)\right)(Y_0(u), u, c(Y_0(u), u, \mu_0(u)), \mu_0(u)) \neq 0,$$

having the same sign as the determinant in (4.14).

The proof of Theorem 4.1 can be found in Sect. 4.4. In Sect. 4.5 we will discuss consequences of Theorem 4.1 as well as some remaining problems. Special attention will go to the role of the flying canard in the bifurcation.

4.3 Typical Shape of Generic Position Curves

4.3.1 Flying Canards

We will start by studying the canard cycles near some Γ_{Y_0}, with $Y_0 \in [Y_1, Y_2]$, under the condition that $I(Y_0, \mu_0) < 0$. As we know from Sect. 4.2 all closed orbits, for $u > 0$, are given by $a = c(Y, u, \mu)$.

If we take $Y \sim Y_0$, $u \sim 0$, $\mu \sim \mu_0$, then, by supposition:

$$A_1(Y, u, c(Y, u, \mu), \mu) < A_2(Y, u, c(Y, u, \mu), \mu). \tag{4.17}$$

Recall that both $A_i(Y, u, c(Y, u, \mu), \mu)$ are strictly negative, if we keep $Y \sim Y_0$, $u \sim 0$, $\mu \sim \mu_0$. In fact

$$A_i(Y, u, a, \mu) = I_i(Y, a, \mu) + O(u), \tag{4.18}$$

and we supposed that

$$I_1(Y_0, 0, \mu_0) < I_2(Y_0, 0, \mu_0), \tag{4.19}$$

both quantities being strictly negative.

If we derive (4.11) w.r.t. Y we get

$$\exp(\frac{1}{u^k}(\tilde{A}_1(Y, u, c(Y, u, \mu), \mu))) - \exp(\frac{1}{u^k}(\tilde{A}_2(Y, u, c(Y, u, \mu), \mu))) \\ = u^k \cdot \frac{\partial c}{\partial Y}(Y, u, \mu) \cdot F(Y, u, c(Y, u, \mu), \mu), \tag{4.20}$$

where

$$\tilde{A}_i(Y, u, a, \mu) = A_i(Y, u, a, \mu) + u^k \log\left(\frac{\partial A_i}{\partial Y}(Y, u, a, \mu)\right) \tag{4.21}$$

and

$$F(Y, u, a, \mu) = \tfrac{\partial f}{\partial a}(u, a, \mu) + \tfrac{1}{u^k}\left(\sum_{i=1}^{2}(-1)^i \tfrac{\partial A_i}{\partial a}(Y, u, a, \mu)\right) \\ \cdot \exp\left(\tfrac{1}{u^k}(A_i(Y, u, a, \mu))\right) \\ = 1 + O(u).$$

For both breaking mechanisms (4.21) implies that

$$\tilde{A}_i(Y, u, c(Y, u, \mu), \mu) = I_i(Y, 0, \mu) + O(u).$$

Let us, from now on, write $I_i(Y, \mu)$ instead of $I_i(Y, 0, u)$. From (4.20) and (4.21) we see that

$$u^k \cdot \frac{\partial c}{\partial Y}(Y, u, \mu) = -\exp\left(\frac{1}{u^k}(I_2(Y, \mu) + O(u))\right). \tag{4.22}$$

It implies that $\frac{\partial c}{\partial Y}(Y, u, \mu)$, for $Y \sim Y_0$, $\mu \sim \mu_0$ and $u > 0$, $u \sim 0$, is strictly negative, inducing the existence of an inverse function

$$Y = p(a, u, \mu),$$

expressing that for $(a, u, \mu) \sim (c(Y_0, u, \mu_0), 0, \mu_0)$, $u > 0$, we find a unique limit cycle, cutting R at some $Y \sim Y_0$. From (4.22) we see that

$$\frac{\partial p}{\partial a}(c(Y, u, \mu), u, \mu) = -u^k \exp\left(-\frac{1}{u^k}(I_2(Y, \mu) + O(u))\right). \tag{4.23}$$

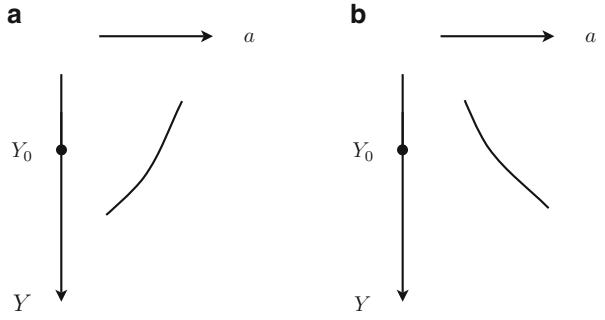

Fig. 4.2 Graphs of $p(a, u, \mu)$, for respectively an attracting and a repelling flying canard

For $u \sim 0$, $Y \sim Y_0$, $\mu \sim \mu_0$, and $u > 0$, we see that $\frac{\partial p}{\partial a}(c(Y, u, \mu), u, \mu)$ is a large negative value, showing that, for changing a, the function $p(a, u, \mu)$ changes very fastly. We could say that the canard "flies" with a fast speed, following the terminology of [1].

For decreasing a the value $p(a, u, \mu)$ increases and the speed $\left|\frac{\partial p}{\partial a}(c(Y, u, \mu), u, \mu)\right|$ decreases since $-I_2(Y, \mu)$ decreases. This behaviour is represented in a qualitative way in Fig. 4.2a, depending on a fixed (u, μ), in a (a, Y)-diagram In case $I(Y_0, \mu_0) > 0$ we can apply the former analysis, but (4.22) and (4.23) change into respectively

$$u^k \frac{\partial c}{\partial Y}(Y, u, \mu) = \exp\left(\frac{1}{u^k}(I_1(Y, \mu) + O(u))\right) \tag{4.24}$$

and

$$\frac{\partial p}{\partial a}(c(Y, u, \mu), u, \mu) = u^k \exp\left(-\frac{1}{u^k}(I_1(Y, \mu) + O(u))\right). \tag{4.25}$$

As long as $I(Y, \mu_0)$ remains strictly negative (resp. strictly positive) the analysis made for $Y \sim Y_0$ remains valid and the graph of $p(a, u, \mu)$ remains qualitatively like in Fig. 4.2a (resp. Fig. 4.2b). The value of $\left|\frac{\partial p}{\partial a}(c(Y, u, \mu), u, \mu)\right|$ will continue decreasing when Y increases.

Let us now continue with the case $I(Y_0, \mu_0) < 0$ and fix $\mu = \mu_0$. As explained in different papers, among which [4], there is no possibility to get a limit cycle γ_Y cutting R at some $Y < Y_0$ if we fix (u, μ_0) and take $a = c(Y_0, u, \mu_0)$. In fact all orbits cutting R at $Y < Y_0$ asymptotically tend to the limit cycle γ_{Y_0}. The same happens for all $Y > Y_0$ at which $I(Y, \mu_0)$ remains strictly negative.

The basin of attraction of γ_{Y_0} can only stop when we approach some $\tilde{Y}_0$ at which $I(\tilde{Y}_0, \mu_0) = 0$.

Let us now study what happens near the zeros of $I(Y,\mu_0)$, under the condition that we have a limit cycle γ_{Y_0}, i.e. a limit cycle cutting R at $Y=Y_0$ for some fixed $(u,\mu)\sim(0,\mu_0)$, and with $a=c(Y_0,u,\mu)$.

We know that

$$\begin{aligned}&\exp\left(\tfrac{1}{u^k}(A_1(Y_0,u,c(Y_0,u,\mu),\mu))\right)-\exp\left(\tfrac{1}{u^k}(A_2(Y_0,u,c(Y_0,u,\mu),\mu),\mu))\right)\\&=f(u,c(Y_0,u,\mu),\mu).\end{aligned}\tag{4.26}$$

To get other limit cycles γ_Y for the same parameter values $(u,c(Y_0,u,\mu),\mu)$ we need to look for values Y that are solution of

$$\begin{aligned}&\exp\left(\tfrac{1}{u^k}(A_1(Y,u,c(Y_0,u,\mu),\mu))\right)-\exp\left(\tfrac{1}{u^k}(A_2(Y,u,c(Y_0,u,\mu),\mu))\right)\\&=f(u,c(Y_0,u,\mu),\mu).\end{aligned}\tag{4.27}$$

Combining (4.26) and (4.27), this implies that we look for solutions of

$$\begin{aligned}&\exp\left(\tfrac{1}{u^k}(A_1^0(Y,u,\mu))\right)-\exp\left(\tfrac{1}{u^k}(A_1^0(Y_0,u,\mu))\right)\\&=\exp\left(\tfrac{1}{u^k}(A_2^0(Y,u,\mu))\right)-\exp\left(\tfrac{1}{u^k}(A_2^0(Y_0,u,\mu))\right)\end{aligned}\tag{4.28}$$

where $A_i^0(Y,u,\mu)=A_i(Y,u,c(Y_0,u,\mu),\mu)$, and $Y>Y_0$.
We see that

$$\begin{aligned}&\exp\left(\tfrac{1}{u^k}A_i^0(Y,u,\mu))\right)-\exp\left(\tfrac{1}{u^k}(A_i^0(Y_0,u,\mu))\right)\\&=\tfrac{1}{u^k}\int_{Y_0}^{Y}\exp\left(\tfrac{1}{u^k}\tilde{A}_i^0(y,u,\mu)\right)dy,\end{aligned}$$

where

$$\begin{aligned}\tilde{A}_i^0(y,u,\mu)&=A_i^0(y,u,\mu)+u^k\log\left(\frac{\partial A_i^0}{\partial Y}(y,u,\mu)\right)\\&=I_i(y,\mu)+O(u).\end{aligned}$$

Equation (4.28) can be rewritten as:

$$\begin{aligned}&\psi(Y)=\psi_{(u,\mu)}(Y)\\&=\int_{Y_0}^{Y}\left[\exp\left(\tfrac{1}{u^k}(\tilde{A}_1^0(y,u,\mu))\right)-\exp\left(\tfrac{1}{u^k}(\tilde{A}_2^0(y,u,\mu))\right)\right]dy=0.\end{aligned}\tag{4.29}$$

As long as $\tilde{A}_1^0(y,u,\mu)<\tilde{A}_2^0(y,u,\mu)$, it is hence not possible to get a second limit cycle γ_Y with $Y>Y_0$.

We have $\psi(Y_0)=0$ and

$$\frac{\partial\psi}{\partial Y}(Y)=\exp\left(\frac{1}{u^k}(\tilde{A}_1^0(Y,u,\mu))\right)-\exp\left(\frac{1}{u^k}(\tilde{A}_2^0(Y,y,\mu))\right).\tag{4.30}$$

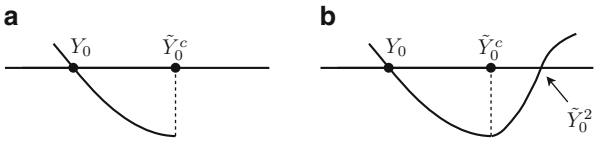

Fig. 4.3 Position curve near a simple zero of the slow divergence integral

Hence $\frac{\partial \psi}{\partial Y}(Y_0) < 0$ and $\frac{\partial \psi}{\partial Y}(Y)$ remains negative as long as $\tilde{A}_1^0(y,u,\mu) < \tilde{A}_2^0(y,u,\mu)$ for $y \in [Y_0, Y]$.

The function ψ will have a critical point at Y iff $\tilde{A}_1^0(Y,u,\mu) = \tilde{A}_2^0(Y,u,\mu)$, and such a value Y has to lay close to a zero of $I(Y,\mu_0)$.

4.3.2 Simple Zeros of $I(Y,\mu_0)$

If we suppose that $I(\tilde{Y}_0,\mu_0) = 0$, and that $\tilde{Y}_0$ is a simple zero of $I(Y,\mu_0)$, then the Implicit Function Theorem implies that, for $(u,\mu) \sim (0,\mu_0)$ there is a unique zero $\tilde{Y}_0^c$ of $(\tilde{A}_1^0 - \tilde{A}_2^0)(Y,u,\mu)$, $\tilde{Y}_0^c \sim \tilde{Y}_0$ and this zero is simple. Of course the zero $\tilde{Y}_0^c$ depends on (u,μ) and tends to $\tilde{Y}_0$ for $(0,\mu) \to (0,\mu_0)$. $\tilde{Y}_0^c$ is the first critical point of ψ for $Y \geq Y_0$ (see Fig. 4.3a). From Theorem 4.3(2) of [4] we know that ψ has a simple zero at some $\tilde{Y}_0^z = \tilde{Y}_0^z(u,\mu)$, for $(u,\mu) \sim (0,\mu_0)$ with $\tilde{Y}_0^z(0,\mu_0) = \tilde{Y}_0$. Of course we need to have $\tilde{Y}_0^z(u,\mu) > \tilde{Y}_0^c(u,\mu)$ and we know that $\frac{\partial \psi}{\partial Y}$ is strictly positive from $\tilde{Y}_0^c$ on, including at $\tilde{Y}_0^z$. The graph of ψ will be like in Fig. 4.3b. We see that $(\tilde{Y}_0^z - \tilde{Y}_0^c)(u,\mu) \to 0$ for $(u,\mu) \to (0,\mu_0)$.

For $Y > \tilde{Y}_0^z$ we can continue in the same way and we know that ψ will have no extra zero, unless we are close to a new zero of $I(Y,\mu_0)$. Near such a zero $\hat{Y}_0$, we will encounter a unique critical point $\hat{Y}_0^c(u,\mu)$ and a unique (simple) zero $\hat{Y}_0^z(u,\mu)$, with $\hat{Y}_0^c(u,\mu) < \hat{Y}_0^z(u,\mu)$ and such that both $\hat{Y}_0^c(u,\mu)$ and $\hat{Y}_0^z(u,\mu)$ tend to $\hat{Y}_0$ for $(u,\mu) \to (0,\mu_0)$.

In Fig. 4.4 we represent the shape of the position curve $p(a,u,\mu)$ below a piece of curve representing a flying canard. Near each simple zero of $I(Y,\mu_0)$ the position curve contains an almost horizontal level. This level is alternatively increasing and decreasing for respectively an attracting or a repelling limit cycle. To see how these pieces of the position curve connect to each other, we will successively work at the different levels $\{Y = \tilde{Y}_0\}$ where $I(Y,\mu_0)$ has a simple zero. We start with a first one. As long as we keep $a = c(Y,u,\mu)$, with $Y_0 \geq Y > \tilde{Y}_0$ we have a situation as represented in Fig. 4.4. We hence need to make a better choice of control curve. After all we know from [12, 13] (see also Theorem 4.3(3) of [4]) that for $Y \sim \tilde{Y}_0$ it is possible to find a generic saddle node bifurcation of limit cycles. To see this it suffices to go back to (4.11), given the equation of the closed orbits:

$$\exp\left(\int \frac{1}{u^k}(A_1(Y,u,a,\mu))\right) - \exp\left(\frac{1}{u^k}(A_2(Y,u,a,\mu))\right) - f(u,a,\mu) = 0. \tag{4.31}$$

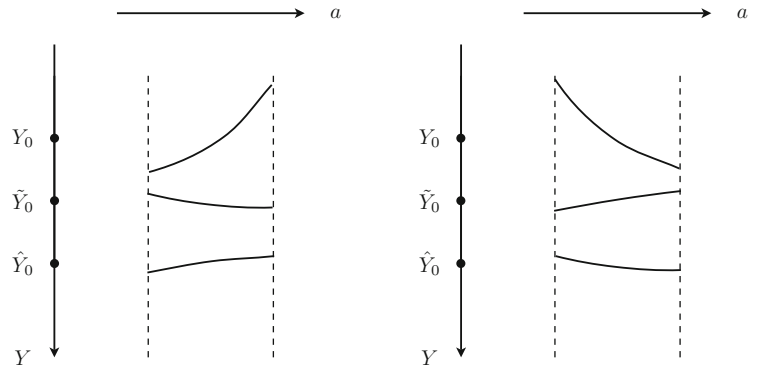

Fig. 4.4 Position curve under a flying canard

Recall that f is calculated w.r.t. some reference Y_r with $Y_r < Y_0$. In order to find a limit cycle of multiplicity two, for $Y \sim \tilde{Y}_0$, we look for a solution of (4.31) that is also solution of

$$\tilde{A}_1(Y, u, a, \mu) - \tilde{A}_2(Y, u, a, \mu) = a \tag{4.32}$$

with $\tilde{A}_i$ as defined in (4.21).

We consider (4.31)–(4.32) as a mapping from $\mathbb{R}^2 \times \mathbb{R}^{p+1} \to \mathbb{R}^2$, depending on the variables (Y, a) and the parameters (u, μ) and we calculate, at $(\tilde{Y}_0, 0, 0, \mu_0)$, the differential with respect to (Y, a). We get:

$$\begin{pmatrix} O(u) & -(1+O(u)) \\ \frac{\partial I}{\partial Y}(\tilde{Y}_0, \mu_0) + O(u) & O(u) \end{pmatrix}.$$

This matrix is invertible, implying that for $(Y, u, a, \mu) \sim (\tilde{Y}_0, 0, 0, \mu_0)$ there is a unique solution

$$(Y, a) = (\tilde{Y}_0^s(u, \mu), \tilde{c}(u, \mu)), \tag{4.33}$$

for which both (4.31) and (4.32) hold. Of course we have that

$$\tilde{c}(u, \mu) = c(\tilde{Y}_0^s(0, \mu), 0, \mu). \tag{4.34}$$

We will now use $\tilde{c}(u, \mu)$ as control curve. A first observation is that

$$\frac{\partial c}{\partial Y}(\tilde{Y}_0^s(u, \mu), u, \mu) = 0.$$

Indeed from (4.31) we get:

$$\begin{aligned} &u^2(1+O(u))\tfrac{\partial c}{\partial Y}(Y,u,\mu) = \\ &\exp\left(\tfrac{1}{u^k}(\tilde{A}_1(Y,u,c(Y,u,\mu),\mu))\right) - \exp\left(\tfrac{1}{u^k}(\tilde{A}_2(Y,u,c(Y,u,\mu),\mu))\right). \end{aligned} \tag{4.35}$$

Evaluating (4.35) at $Y = \tilde{Y}_0^s(u,\mu)$, the claim now follows from (4.34), knowing that (4.33) is a solution of (4.32). From (4.35) we also see that $(\tilde{Y}_0^s(u,\mu),u,\mu)$ are the only values at which $\frac{\partial c}{\partial Y} = 0$, at least for $(Y,u,\mu) \sim (\tilde{Y}_0,0,\mu_0)$. If we derive (4.35) w.r.t. Y we get:

$$\begin{aligned} &u^{2k}(1+O(u))\tfrac{\partial^2 c}{\partial Y^2}(Y,u,\mu) + O(u^{2k+1})\tfrac{\partial c}{\partial Y}(Y,u,\mu) \\ &= \exp\left(\tfrac{1}{u^k}(\tilde{\tilde{A}}_1(Y,u,c(Y,u,\mu),\mu))\right) - \exp\left(\tfrac{1}{u^k}(\tilde{\tilde{A}}_2(Y,u,c(Y,u,\mu),\mu))\right), \end{aligned} \tag{4.36}$$

with

$$\begin{aligned} \tilde{\tilde{A}}_i(Y,u,a,\mu) &= \tilde{A}_i(Y,u,a,\mu) + u^k \log\left(\frac{\partial \tilde{A}_i}{\partial Y}(Y,u,a,\mu)\right) \\ &= I_i(Y,\mu) + O(u). \end{aligned}$$

If we evaluate (4.36) at $Y = \tilde{Y}_0^s(u,\mu)$ we get

$$\begin{aligned} &u^{2k}(1+O(u))\tfrac{\partial^2 c}{\partial Y^2}(\tilde{Y}_0^s(u,\mu),u,\mu) = \\ &\left[\left(\tfrac{\partial \tilde{A}_1}{\partial Y} - \tfrac{\partial \tilde{A}_2}{\partial Y}\right)(\tilde{Y}_0^s(u,\mu),u,c(\tilde{Y}_0^s(u,\mu),u,\mu),\mu)\right]. \\ &\exp\left(\tfrac{1}{u^k}(\tilde{A}_1(\tilde{Y}_0^s(u,\mu),u,c(\tilde{Y}_0^s(u,\mu),u,\mu),\mu))\right). \end{aligned} \tag{4.37}$$

Since the quantity in between brackets is equal to

$$\frac{\partial I}{\partial Y}(\tilde{Y}_0,\mu) + O(u),$$

which is nonzero, we see that $\frac{\partial^2 c}{\partial Y^2}(\tilde{Y}_0^s(u,\mu),u,\mu)$, for $(u,\mu) \sim (0,\mu_0)$ is nonzero and has the same sign as $\frac{\partial I}{\partial Y}(\tilde{Y}_0,\mu_0)$.

It implies that $c(Y,u,\mu)$, for $(u,\mu) \sim (0,\mu_0)$ and $u > 0$ has a Morse type maximum if $\frac{\partial I}{\partial Y}(\tilde{Y}_0,\mu_0) > 0$ and a Morse type minimum if $\frac{\partial I}{\partial Y}(\tilde{Y}_0,\mu_0) < 0$.

We can continue the analysis of $p(a,u,\mu)$ and $c(Y,u,\mu)$ by working with (4.29), however changing the control curve $c(Y_0,u,\mu)$ to $\tilde{c}(u,\mu) = c(\tilde{Y}_0^s(u,\mu),u,\mu)$.

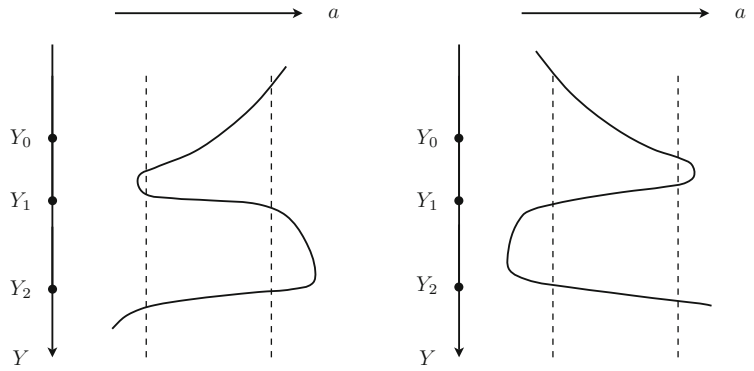

Fig. 4.5 Typical shape of a position curve

We easily find that no limit cycles are possible for $Y < \tilde{Y}_0^s(u,\mu)$, since there are no zeros of $I(Y,\mu_0)$ at such values. However at zeros of $I(Y,\mu_0)$ with $Y > \tilde{Y}_0^s(u,\mu)$, we can make an analysis as we just made.

As a conclusion we see that a typical position curve on some segment of R has a shape as represented in Fig. 4.5.

The "levels" that are encountered beneath any value of the position curve are not only typical for a position curve, but necessary.

The position curves that are drawn in Fig. 4.5 are related to a slow divergence integral $I(Y,\mu_0)$ that only has simple zeros. In Sect. 4.4 we will study what happens at a multiple zero of $I(Y,\mu_0)$.

4.4 Catastrophes of Canard Type Limit Cycles

We continue the study that we started in Sect. 4.3, using the same notations as introduced there. In this section we will look what happens at a multiple zero Y_0 of $I(Y,\mu_0)$. We continue relying on Eq. (4.10), of which $a = c(Y,u,\mu)$ is a solution (see (4.11)), representing the "manifold of closed orbits".

The jth derivative of (4.10), for $j \geq 1$, can be written as:

$$\frac{1}{u^{jk}}\left[\exp\left(\frac{1}{u^k}(\tilde{A}_1^j(Y,u,a,\mu))\right) - \exp\left(\frac{1}{u^k}(\tilde{A}_2^j(Y,u,a,\mu))\right)\right],$$

where $\tilde{A}_i^j$ is defined recursively as

$$\tilde{A}_i^j(Y,u,a,\mu) = \tilde{A}_i^{j-1}(Y,u,a,\mu) + u^k \log\left(\frac{\partial \tilde{A}_i^{j-1}}{\partial Y}(Y,u,a,\mu)\right),$$

and $\tilde{A}_i^0 = A_i$.

To find limit cycles of codimension $m+2$ we solve (4.10) in combination with

$$\tilde{A}_1^j(Y,u,a,\mu) - \tilde{A}_2^j(Y,a,a,\mu) = 0, \text{ for } j = 1,\ldots,m+1. \tag{4.38}$$

The common solutions to all equations of (4.38) coincide with the common solutions of the equations

$$\frac{\partial^j(\tilde{A}_1^1 - \tilde{A}_2^1)}{\partial Y^j}(Y,u,a,\mu), \text{ for } j = 0,\ldots,m. \tag{4.39}$$

We also know that

$$\frac{\partial^j(\tilde{A}_1^1 - \tilde{A}_2^1)}{\partial Y^j}(Y,u,a,\mu) = \frac{\partial^j I}{\partial Y^j}(Y,\mu) + O(u) + O(a). \tag{4.40}$$

We now suppose that, at some $Y = Y_0$—and for simplicity in notation we choose $Y_0 = 0$ and $\mu_0 = 0$—we have:

$$I(0,0) = \frac{\partial I}{\partial Y}(0,0) = \ldots = \frac{\partial^n I}{\partial Y^n}(0,0) = 0, \frac{\partial^{n+1} I}{\partial Y^{n+1}}(0,0) > 0,$$

$$\det\left(\partial(I, \frac{\partial I}{\partial Y}, \ldots, \frac{\partial^{n-1} I}{\partial Y^{n-1}})/\partial(\mu_1,\ldots,\mu_n)\right)(0,0) \neq 0,$$

where we take $\mu = (\mu_1,\ldots,\mu_n) \sim (0,\ldots,0)$.

Similar conditions hold for $(\tilde{A}_1^1 - \tilde{A}_2^1)$ at $(Y,u,a,\mu) = (0,0,0,0)$, see (4.40), and we apply the Malgrange preparation Theorem on the u-regularly smooth $\tilde{A}_1^1 - \tilde{A}_2^1$ to write it as:

$$(\tilde{A}_1^1 - \tilde{A}_2^1)(Y,u,a,\mu) = P(Y,u,a,\mu)\cdot(Y^{n+1} + \sum_{i=0}^{n-1}\lambda_i Y^i), \tag{4.41}$$

with P u-regularly smooth in (Y,a,μ) and strictly positive; the λ_i are new parameters that are u-regularly smooth combinations of the old ones. From now we also suppose that $\mu = (\lambda_0,\ldots,\lambda_{n-1})$. Also Y is a new variable obtained in an u-regularly smooth way.

Remark 4.3. The whole expression (4.41) can depend on extra parameters in an u-regularly smooth way, but we do not add this to the notation.

Deriving (4.11) w.r.t. Y we get:

$$\begin{aligned} &u^k(1+O(u))\tfrac{\partial c}{\partial Y}(Y,u,\mu) = \\ &\exp\left(\tfrac{1}{u^k}(\tilde{A}_1^1(Y,u,c(Y,u,\mu),\mu))\right) - exp\left(\tfrac{1}{u^k}(\tilde{A}_2^1(Y,u,c(Y,u,\mu),\mu))\right). \end{aligned} \tag{4.42}$$

Seen expression (4.41) the derivative $\frac{\partial c}{\partial Y}(Y,u,\mu)$, for $u>0$, is nonzero whenever

$$Y^{n+1}+\sum_{i=0}^{n-1}\lambda_i Y^i\neq 0,$$

and its sign is given by the sign of this expression.

Remains to look what happens at the zeros of $\tilde{A}_1^1-\tilde{A}_2^1$, namely along

$$Y^{n+1}+\sum_{i=0}^{n-1}\lambda_i Y^i=0.$$

To make a study along this manifold of critical closed orbits we introduce a new parameter $\tilde{Y}\sim 0$, keeping $\lambda_1,\ldots,\lambda_{n-1}$ and changing λ_0 by

$$\lambda_0=-\left(\tilde{Y}^{n+1}+\sum_{i=1}^{n-1}\lambda_i\tilde{Y}^i\right). \tag{4.43}$$

Referring to [5] we say that we "clone" Y. The parameters $(\tilde{Y},\lambda_1,\ldots,\lambda_{n-1})$ parametrize, by means of (4.43), the stratum Σ_0 in parameter space to which we can restrict our study.

We introduce the notation

$$\mu_c(\tilde{Y})=\left(-(\tilde{Y}^{n+1}+\sum_{i=1}^{n-1}\lambda_i\tilde{Y}^i),\lambda_1,\ldots,\lambda_{n-1}\right).$$

Let us now prove that

$$\frac{\partial^2 c}{\partial Y^2}(\tilde{Y},u,\mu_c(\tilde{Y}))=(1+O(u))\left[\frac{\partial}{\partial Y}(\tilde{A}_1^1-\tilde{A}_2^1)(\tilde{Y},u,a_c(\tilde{Y}),\mu_c(\tilde{Y}))\right]\cdot \frac{1}{u^{2k}}\exp\left(\frac{1}{u^k}(\tilde{A}_1^1(\tilde{Y},u,a_c(\tilde{Y}),\mu_c(\tilde{Y})\right), \tag{4.44}$$

where $a_c(\tilde{Y})=c(\tilde{Y},u,\mu_c(\tilde{Y}))$.

For this we derive (4.42) w.r.t. Y and get:

$$u^{2k}\left[(1+O(u))\tfrac{\partial^2 c}{\partial Y^2}(Y,u,\mu)+O(u)\tfrac{\partial c}{\partial Y}(Y,u,\mu)\right] = \tfrac{\partial \tilde{A}_1^1}{\partial Y}(Y,u,a,\mu)\cdot\exp\left(\tfrac{1}{u^k}\tilde{A}_1^1(Y,u,a,\mu)\right)-\tfrac{\partial \tilde{A}_2^1}{\partial Y}(Y,u,a,\mu)\cdot\exp\left(\tfrac{1}{u^k}\tilde{A}_2^1(Y,u,a,\mu)\right),$$

since $\frac{\partial}{\partial a}(\exp(\tilde{A}_r^1))$, with $r=1,2$, are u-regularly smooth functions in (Y,a,μ) that are flat in u. Restricting to $(a,\mu)=(a_c(\tilde{Y}),\mu_c(\tilde{Y}))$ we see that $\frac{\partial c}{\partial Y}=0$ as well as

$\tilde{A}_1^1 = \tilde{A}_2^1$, so that we get (4.44). Expression (4.44) implies that, along $\frac{\partial c}{\partial Y} = 0$ and for $u > 0$, $\frac{\partial^2 c}{\partial Y^2}$ will be nonzero whenever

$$(n+1)Y^n + \sum_{i=1}^{n-1} i\lambda_i Y^{i-1} \neq 0;$$

its sign is given by the sign of this expression.

We can now further restrict to a stratum $\Sigma_1 \subset \Sigma_0$ in parameter space that is parametrized by $(\tilde{Y}, \lambda_2, \ldots, \lambda_{n-1})$, and along which $\frac{\partial c}{\partial Y} = \frac{\partial^2 c}{\partial Y^2} = 0$.

We introduce the notation:

$$\mu_c^1(\tilde{Y}) = (\tilde{\lambda}_0(\tilde{Y}), \tilde{\lambda}_1(\tilde{Y}), \lambda_2, \ldots, \lambda_{n-1}),$$

with

$$\begin{aligned} &\tilde{\lambda}_1(\tilde{Y}) = -\left(\sum_{i=2}^{n-1} i\lambda_i \tilde{Y}^{i-1} + (n+1)\tilde{Y}^n\right) \\ &\text{and} \\ &\tilde{\lambda}_0(\tilde{Y}) = -\left(\tilde{Y}^{n+1} + \sum_{i=1}^{n-1} \lambda_i \tilde{Y}^i + \tilde{\lambda}_1(\tilde{Y})\right). \end{aligned} \tag{4.45}$$

We define $a_c^1(\tilde{Y})$ accordingly as:

$$a_c^1(\tilde{Y}) = c(\tilde{Y}, u, \mu_c^1(\tilde{Y})).$$

This procedure can now be continued recursively. The conditions $\frac{\partial^j}{\partial Y^j}(Y^{n+1} + \sum_{i=1}^{n-1} \lambda_i Y^i) = 0$, for $j = 0, \ldots, l$, with $1 \leq l \leq n-1$, define a stratum Σ_l in parameter space, that can be parametrized by $(\tilde{Y}, \lambda_{l+1}, \cdots, \lambda_{n-1})$ in case $l \leq n-2$ and by $\tilde{Y}$ in case $l = n-1$. The parametrization $\mu = \mu_c^l(\tilde{Y})$ is a straightforward generalization of (4.45) and we define $a_c^l(\tilde{Y}) = c(\tilde{Y}, u, \mu_c^l(\tilde{Y}))$. We define Σ_n to be the point $(\tilde{Y}, \mu) = (0, 0)$. Along Σ_l, with $l = 1, \ldots, n$ we will prove that

$$\begin{aligned} \frac{\partial^{l+2}}{\partial Y^{l+2}}(\tilde{Y}, u, \mu_c^l(\tilde{Y})) = (1 + O(u)) \left[\frac{\partial^{l+1}}{\partial Y^{l+1}}(\tilde{A}_1^1 - \tilde{A}_2^1)(\tilde{Y}, u, a_c^l(\tilde{Y}), \mu_c^l(\tilde{Y}))\right] \frac{1}{u^{2k}} \cdot \\ \exp\left(\frac{1}{u^k}(\tilde{A}_1^1(\tilde{Y}, u, a_c^l(\tilde{Y}), \mu_c^l(\tilde{Y})\right), \end{aligned} \tag{4.46}$$

where $a_c^l(\tilde{Y}) = c(\tilde{Y}, u, \mu_c^l(\tilde{Y}))$.

To see this, we recursively derive (4.42) w.r.t.Y, giving:

$$u^{k(j+1)}\left[(1+O(u))\tfrac{\partial^{j+1}}{\partial Y^{j+1}}(Y,u,\mu)+\sum_{s=1}^{j}O_s(u)\tfrac{\partial^s c}{\partial Y^s}(Y,u,\mu)\right]$$
$$=\exp\left(\tfrac{1}{u^k}(\tilde{A}_1^{j+1}(Y,u,c(Y,u,\mu),\mu))\right)-\exp\left(\tfrac{1}{u^k}(\tilde{A}_2^{j+1}(Y,u,c(Y,u,\mu),\mu))\right), \tag{4.47}$$

for $j=1,\ldots,n$, where $O(u)$ and $O_s(u)$, $s=1,\ldots,j$, stand for u-regularly smooth functions in $(Y,u,c(Y,u,\mu),\mu)$ that are $O(u)$.

Also for each $j=1,\ldots,n$, and $r=1,2$ we recall that

$$\exp\left(\frac{1}{u^k}\tilde{A}_r^{j+1}\right)=\frac{\partial\tilde{A}_r^j}{\partial Y}\cdot\exp\left(\frac{1}{u^k}\tilde{A}_r^j\right)=u^k\frac{\partial}{\partial Y}\left(\exp\left(\frac{1}{u^k}\tilde{A}_r^j\right)\right). \tag{4.48}$$

From (4.48) follows that:

$$\exp\left(\tfrac{1}{u^k}\tilde{A}_r^3\right)=u^k\cdot\tfrac{\partial}{\partial Y}\left(\exp\left(\tfrac{1}{u^k}\tilde{A}_r^2\right)\right)=u^k\tfrac{\partial}{\partial Y}\left(\tfrac{\partial\tilde{A}_r^1}{\partial Y}\cdot\exp\left(\tfrac{1}{u^k}\tilde{A}_r^1\right)\right)$$
$$=\left(u^k\tfrac{\partial^2\tilde{A}_r^1}{\partial Y^2}+\left(\tfrac{\partial\tilde{A}_r^1}{\partial Y}\right)^2\right)\cdot\exp\left(\tfrac{1}{u^k}\tilde{A}_r^1\right).$$

Recursively we see that for $2\le j\le n$:

$$\exp\left(\frac{1}{u^k}\tilde{A}_r^{j+1}\right)=\left(u^{k(j-1)}\frac{\partial^j\tilde{A}_r^1}{\partial Y^j}+P_j\left(\frac{\partial\tilde{A}_r^1}{\partial Y},\ldots,\frac{\partial^{j-1}\tilde{A}_r^1}{\partial Y^{j-1}}\right)\right)\exp\left(\frac{1}{u^k}\tilde{A}_r^1\right) \tag{4.49}$$

where P_j is some universal polynomial in its $(j-1)$ variables. The coefficients of the polynomial can contain powers u^{kq}, with $0\le q\le j-2$.

We can now change the right-hand side of (4.46) by using (4.49), for parameter values along Σ_0, and get:

$$u^{k(j+1)}\left((1+O(u))\tfrac{\partial^{j+1}c}{\partial Y^{j+1}}(Y,u,\mu)+\sum_{s=1}^{j}O_s(u)\tfrac{\partial^s c}{\partial Y^s}(Y,u,\mu)\right)$$
$$=\left(u^{k(j-1)}\tfrac{\partial^j}{\partial Y^j}(\tilde{A}_1^1-\tilde{A}_2^2)+P_j\left(\tfrac{\partial\tilde{A}_1^1}{\partial Y},\ldots,\tfrac{\partial^{j-1}\tilde{A}_1^1}{\partial Y^{j-1}}\right)-P_j\left(\tfrac{\partial\tilde{A}_2^1}{\partial Y},\ldots,\tfrac{\partial^{j-1}\tilde{A}_2^1}{\partial Y^{j-1}}\right)\right)\cdot$$
$$\exp\left(\tfrac{1}{u^k}\tilde{A}_1^1\right). \tag{4.50}$$

Along Σ_l, for $2\le l\le n$, we clearly see that (4.46) is a consequence of (4.50).

Along Σ_0 we can also prove that for $0 \leq i \leq n-1$ we have:

$$(1+O(u))\frac{\partial^2 c}{\partial\lambda_i \partial Y}(\tilde{Y},u,\mu_c^0(\tilde{Y})) = \left[\frac{\partial}{\partial\lambda_i}(\tilde{A}_1^1-\tilde{A}_2^1)(\tilde{Y},u,a_c^0(\tilde{Y}),\mu_c^0(\tilde{Y}))\right]\cdot\frac{1}{u^{2k}}\exp\left(\frac{1}{u^k}(\tilde{A}_1^1(\tilde{Y},u,a_c^0(\tilde{Y}),\mu_c^0(\tilde{Y}))\right). \tag{4.51}$$

We therefore derive expression (4.42) w.r.t. λ_i and get:

$$\begin{aligned} &u^{2k}(1+O(u))\frac{\partial^2 c}{\partial\lambda_i\partial Y}(Y,u,\mu)+O(u^{2k+1})\frac{\partial c}{\partial Y}(Y,u,\mu) = \\ &\left(\frac{\partial\tilde{A}_1^1}{\partial\lambda_i}(Y,u,c(Y,u,\mu),\mu)+\frac{\partial\tilde{A}_1^1}{\partial a}(Y,u,c(Y,u,\mu),\mu)\frac{\partial c}{\partial\lambda_i}(Y,u,\mu)\right)\cdot \\ &\exp\left(\frac{1}{u^k}(\tilde{A}_1^1(Y,u,c(Y,u,\mu),\mu))\right) \\ &-\left(\frac{\partial\tilde{A}_2^1}{\partial\lambda_i}(Y,u,c(Y,u,\mu),\mu)+\frac{\partial\tilde{A}_1^2}{\partial a}(Y,u,c(Y,u,\mu),\mu)\cdot\frac{\partial c}{\partial\lambda_i}(Y,u,\mu))\right)\cdot \\ &\exp\left(\frac{1}{u^k}(\tilde{A}_2^1(Y,u,c(Y,u,\mu),\mu)\right). \end{aligned} \tag{4.52}$$

Along Σ_0, where $\frac{\partial c}{\partial Y}=0$, $\tilde{A}_1^1=\tilde{A}_2^1$ and $\frac{\partial\tilde{A}_1^1}{\partial a}=\frac{\partial\tilde{A}_2^1}{\partial a}$, we see that (4.51) is a consequence of (4.52).

We would now like to study $\frac{\partial^{l+2}c}{\partial\lambda_i\partial Y^{l+1}}$ with $l\geq 1$ and $l\leq i\leq n-1$

We start with $l=1$ and derive (4.44) w.r.t. λ_i. We get:

$$\begin{aligned} &u^{2k}\left[(1+O(u))\frac{\partial^3 c}{\partial\lambda_i\partial Y^2}(Y,u,\mu)+O_1(u)\frac{\partial^2 c}{\partial\lambda_i\partial Y}(Y,u,\mu)+\sum_{v=1}^{2}O_v(u)\frac{\partial^v c}{\partial Y^v}(Y,u,\mu)\right] \\ &=\frac{\partial}{\partial\lambda_i}\left[\exp\left(\frac{1}{u^k}(\tilde{A}_1^2(Y,u,c(Y,u,\mu),\mu))\right)\right. \\ &\left.-\exp\left(\frac{1}{u^k}(\tilde{A}_2^2(Y,u,c(Y,u,\mu),\mu))\right)\right], \end{aligned} \tag{4.53}$$

where the functions $O_1(u)$, $O_v(u)$ and $O(u)$ are u-regularly smooth in (Y,μ) and $O(u)$.

From the first equality of (4.48) we get, for $r=1,2$, and $\tilde{A}_r^2=\tilde{A}_r^2(Y,u,c(Y,u,\mu),\mu)$:

$$\begin{aligned} &\frac{\partial}{\partial\lambda_i}\left[\exp\left(\frac{1}{u^k}(\tilde{A}_r^2)\right)\right]=\frac{\partial^2\tilde{A}_r^1}{\partial\lambda_i\partial Y}\cdot\exp\left(\frac{1}{u^k}\tilde{A}_r^1\right)+\frac{1}{u^k}\cdot\frac{\partial\tilde{A}_r^1}{\partial Y}\frac{\partial\tilde{A}_r^1}{\partial\lambda_i}\cdot\exp\left(\frac{1}{u^k}\tilde{A}_r^1\right) \\ &+\frac{\partial c}{\partial\lambda_i}\left(\frac{\partial^2\tilde{A}_r^1}{\partial a\partial Y}+\frac{1}{u^k}\cdot\frac{\partial\tilde{A}_r^1}{\partial a}\right)\exp\left(\frac{1}{u^k}\tilde{A}_r^1\right). \end{aligned} \tag{4.54}$$

Along Σ_1, and because of (4.41) we know that

$$\frac{\partial\tilde{A}_1^1}{\partial a}=\frac{\partial\tilde{A}_2^1}{\partial a},\text{ as well as }\frac{\partial^2\tilde{A}_1^1}{\partial a\partial Y}=\frac{\partial^2\tilde{A}_2^1}{\partial a\partial Y}.$$

Using (4.54) to adapt the right-hand side of (4.52) and restricting to Σ_1, we transform (4.53) into:

$$u^{2k}\left[(1+O(u))\frac{\partial^3 c}{\partial\lambda_i\partial Y^2}(Y,u,\mu)+O_1(u)\frac{\partial^2 c}{\partial\lambda_i\partial Y}(Y,u,\mu)\right]$$
$$=\left(\frac{\partial^2}{\partial\lambda_i\partial Y}(\tilde{A}_1^1-\tilde{A}_2^1)\right)\exp\left(\frac{1}{u^k}\tilde{A}_1^1\right)+\frac{1}{u^k}\left(\frac{\partial}{\partial\lambda_i}\left(\tilde{A}_1^1-\tilde{A}_2^1\right)\right)\frac{\partial\tilde{A}_1^1}{\partial Y}\cdot\exp\left(\frac{1}{u^k}\tilde{A}_1^1\right). \tag{4.55}$$

Because of (4.50), expression (4.55) leads to:

$$u^{2k}(1+O(u))\frac{\partial^3 c}{\partial\lambda_i\partial Y^2}(Y,u,\mu)=$$
$$\left(\frac{\partial^2}{\partial\lambda_i\partial Y}(\tilde{A}_1^1-\tilde{A}_2^1)(\tilde{Y},u,a_c^1(\tilde{Y}),\mu_c^1(\tilde{Y}))+\right. \tag{4.56}$$
$$\left.\frac{1}{u^k}\left(\frac{\partial\tilde{A}_1^1}{\partial Y}(\tilde{Y},u,a_c^1(\tilde{Y}),\mu_c^1(\tilde{Y}))+O(u)\right)\cdot\frac{\partial}{\partial\lambda_i}\left(\tilde{A}_1^1-\tilde{A}_2^1\right)\right)\cdot\exp\left(\frac{1}{u^k}\tilde{A}_1^1\right).$$

Along Σ_1 we have

$$\frac{\partial}{\partial\lambda_i}\left(\tilde{A}_1^1-\tilde{A}_2^1\right)\left(\tilde{Y},u,a_c^1(\tilde{Y}),\mu_c^1(\tilde{Y})\right)=P\left(\tilde{Y},u,a_c^1(\tilde{Y}),\mu_c^1(\tilde{Y})\right)\cdot\tilde{Y}^i.$$

In any case, at $(\tilde{Y},\mu)=(0,0)$, we see that for $i\geq 1$:

$$(1+O(u))\frac{\partial^3 c}{\partial\lambda_i Y^2}(0,u,0)=\frac{\partial^2}{\partial\lambda_i\partial Y}(\tilde{A}_1^1-\tilde{A}_2^1)(0,u,0,0)\cdot\frac{1}{u^{2k}}\cdot\exp\left(\frac{1}{u^k}\tilde{A}_1^1\right) \tag{4.57}$$

We can now continue by induction on l, for $l\geq 2$, by deriving (4.47) w.r.t. λ_i. We get:

$$u^{k(l+1)}\left[(1+O(u))\frac{\partial^{l+2}c}{\partial\lambda_i\partial Y^{l+1}}(Y,u,\mu)+\sum_{s=1}^{l}O_s(u)\frac{\partial^{s+1}c}{\partial\lambda_i\partial Y^s}(Y,u,\mu)+\right.$$
$$\left.\sum_{v=1}^{l+1}O_v(u)\frac{\partial^v c}{\partial Y^v}(Y,u,\mu)\right] \tag{4.58}$$
$$=\frac{\partial}{\partial\lambda_i}\left[\exp\left(\frac{1}{u^k}\left(\tilde{A}_1^{l+1}(Y,u,c(Y,u,\mu),\mu)\right)\right)-\exp\left(\frac{1}{u^k}(\tilde{A}_2^{l+1}(Y,u,c(Y,u,\mu),\mu)\right)\right].$$

To adapt the right-hand side of (4.58) we derive (4.48) w.r.t. λ_i and get, changing j by l, and for $r=1,2$:

$$\frac{\partial}{\partial\lambda_i}\left(\exp\left(\tfrac{1}{u^k}\tilde{A}_r^{l+1}\right)\right) =$$

$$\left(u^{k(l-1)}\tfrac{\partial^{l+1}\tilde{A}_r^1}{\partial\lambda_i\partial Y^l} + \sum_{v=1}^{l-1} Q_v\left(\tfrac{\partial\tilde{A}_r^1}{\partial Y},\ldots,\tfrac{\partial^{l-1}\tilde{A}_r^1}{\partial Y^{l-1}}\right)\tfrac{\partial^{v+1}\tilde{A}_r^1}{\partial\lambda_n\partial Y^v}\right)\cdot\exp\left(\tfrac{1}{u^k}\tilde{A}_r^1\right)$$

$$+\left(u^{k(l-1)}\tfrac{\partial^l\tilde{A}_r^1}{\partial Y^l} + P_l\left(\tfrac{\partial\tilde{A}_r^1}{\partial Y},\ldots,\tfrac{\partial^{l-1}\tilde{A}_r^1}{\partial Y^{l-1}}\right)\right)\cdot\tfrac{1}{u^k}\cdot\tfrac{\partial\tilde{A}_r^1}{\partial\lambda_i}\exp\left(\tfrac{1}{u^k}\tilde{A}_r^1\right). \tag{4.59}$$

$$+\tfrac{\partial c}{\partial\lambda_i}\cdot Q_l\left(\tfrac{\partial\tilde{A}_r^1}{\partial a},\tfrac{\partial^2\tilde{A}_r^1}{\partial a\partial Y},\ldots,\tfrac{\partial^{l+1}\tilde{A}_r^1}{\partial a\partial Y^l}\right)\cdot\exp\left(\tfrac{1}{u^k}\tilde{A}_r^1\right)$$

The functions Q_v, $v = 1,\ldots,l$ are universal polynomials in their respective variables. We will now continue working along Σ_l, permitting to use that $\frac{\partial^v c}{\partial Y^v} = 0$ for $v = 1,\ldots,l+1$, that $\frac{\partial^{j+1}}{\partial a\partial Y}(\tilde{A}_1^1 - \tilde{A}_2^1) = 0$ for $j = 0,\ldots,l$, as well as $\frac{\partial^j}{\partial Y^j}(\tilde{A}_1^1 - \tilde{A}_2^1) = 0$ for $j = 0,\ldots,l$.

Combining (4.58) and (4.59) along Σ_l we now get:

$$u^{k(l+1)}\left[(1+O(u))\tfrac{\partial^{l+2}c}{\partial\lambda_i\partial Y^{l+1}} + \sum_{s=1}^{l} O_s(u)\tfrac{\partial^{s+1}c}{\partial\lambda_i\partial Y^s}\right] =$$

$$\left[u^{k(l-1)}\tfrac{\partial^{l+1}}{\partial\lambda_i\partial Y^l}(\tilde{A}_1^1 - \tilde{A}_2^1) + \sum_{v=1}^{l-1} Q_v\left(\tfrac{\partial\tilde{A}_1^1}{\partial Y},\ldots,\tfrac{\partial^{l-1}\tilde{A}_1^1}{\partial Y^{l-1}}\right)\tfrac{\partial^{v+1}}{\partial\lambda_i\partial Y^v}(\tilde{A}_1^1 - \tilde{A}_2^1)\right.$$

$$\left.+\left(u^{k(l-1)}\tfrac{\partial^l\tilde{A}_1^1}{\partial Y^l} + P_l\left(\tfrac{\partial\tilde{A}_1^1}{\partial Y},\ldots,\tfrac{\partial^{l-1}\tilde{A}_1^1}{\partial Y^{l-1}}\right)\right)\cdot\tfrac{1}{u^k}\cdot\tfrac{\partial}{\partial\lambda_i}(\tilde{A}_1^1 - \tilde{A}_2^1)\right]\exp\left(\tfrac{1}{u^k}\tilde{A}_1^1\right). \tag{4.60}$$

Recursively on l, starting with (4.51) for $l = 0$ and (4.56) for $l = 1$, we can prove that:

$$u^{k(l+1)}(1+O(u))\tfrac{\partial^{l+2}c}{\partial\lambda_i\partial Y^{l+1}} =$$
$$\left[u^{k(l-1)}\tfrac{\partial^{l+1}}{\partial\lambda_i\partial Y^l}(\tilde{A}_1^1 - \tilde{A}_2^1) + \sum_{v=0}^{l} R_v\tfrac{\partial^{v+1}}{\partial\lambda_i\partial Y^v}(\tilde{A}_1^1 - \tilde{A}_2^1)\right]\cdot\exp\left(\tfrac{1}{u^k}\tilde{A}_1^1\right), \tag{4.61}$$

where for N_l sufficiently large, the functions $u^{N_l}\cdot R_v$ are u-regularly smooth functions in $(\tilde{Y}, a_c^l(\tilde{Y}), \mu_c^l(\tilde{Y}))$.

4.5 Consequences of Theorem 4.1 and Remaining Problems

Under the notations introduced in Sect. 4.2 and under the conditions expressed in Theorem 4.1, we know that limit cycles are situated on the manifold of closed orbits, given by:

$$F(Y,u,a,\mu) = c(Y,u,\mu) - a = 0. \tag{4.62}$$

The properties of c along $(Y,\mu)=(Y_0(u),\mu_0(u))$, as stated in Theorem 4.1, imply that following properties hold on F:

$$F(Y,u,a,\mu)=\frac{\partial F}{\partial Y}(Y,u,a,\mu)=\ldots=\frac{\partial^{n+1}F}{\partial Y^{n+1}}(Y,u,a,\mu)=0,$$

$$\frac{\partial^{n+2}F}{\partial Y^{n+2}}(Y,u,a,\mu)>0,$$

$$\det\left(\partial(F,\frac{\partial F}{\partial Y},\ldots,\frac{\partial^n F}{\partial Y^n})/\partial(a,\mu_1,\ldots,\mu_n)\right)(Y,u,a,\mu)\neq 0,$$

if we restrict to $(Y,u,a,\mu)=(Y_0(u),u,c(Y_0(u),u,\mu_0(u)),\mu_0(u))$ and keep $u\sim 0$.

The results of Theorem 4.1 are valid in a full neighbourhood $[Y_0-\delta,Y_0+\delta]\times B_\delta(\mu_0)\times]0,u_0[$, for some $\delta>0$, $u_0>0$.

It is hence clear that for $u>0$, there is an elementary catastrophe of codimension $(n+1)$ of limit cycles near respectively the FSTS-cycle (in the Hopf breaking mechanism) or the FSJS-cycle (in the jump breaking mechanism) under consideration, if we take δ sufficiently small.

This does however not necessarily imply that, for each $u\in]0,u_0]$, the local catastrophe near $(Y_0(u),\mu_0(u))$ extends in a trivial way and that all bifurcations on the limit cycles are expressed by this catastrophe.

From Sect. 4.3 we already know that this can even not be the case for a catastrophe of codimension 1, i.e. for a generic saddle-node bifurcation of limit cycles. Indeed, if we choose some u with $u\in]0,u_0]$, then there exist appropriate $\delta>0$ and $\delta'>0$ such that—up to the sign of a—the position curve looks like in Fig. 4.6a.

It implies that, for each value of a, we have 0 or 2 limit cycles, multiplicity taken into account.

If we now fix $[Y_0-\delta',Y_0+\delta]$ and let $u\downarrow 0$, then we know from Sect. 4.3 that the position curve will not only tend to $\{u=0\}$ but also take a typical shape like in Fig. 4.6b.

This movement only expresses that the flying canard moves much faster under the change of a than a sitting one does. For some values of a we now have the possibility to encounter a single limit cycle of multiplicity one. It implies that the bifurcations expressed by the elementary catastrophe are not the only bifurcations that we encounter in a u-uniform neighbourhood of Y_0. We have to add a boundary bifurcation, permitting the flying canard to escape from the fixed neighbourhood $[Y_0-\delta',Y_0+\delta]$.

Similar boundary bifurcations happen of course also for the elementary catastrophes of higher codimension. It explains how an elementary catastrophe of codimension n on sitting canards can become an elementary catastrophe of one codimension higher when it gets hit by a flying canard. The resulting catastrophe of codimension $n+1$ does not extend in a uniformly trivial way w.r.t. to u.

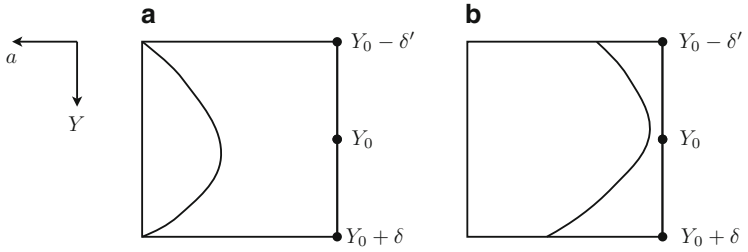

Fig. 4.6 Evolution of position curve when changing u

This boundary bifurcation is comparable to a similar one that has been proven to exist in the unfolding of a codimension 3 nilpotent singularity with an elliptic sector (see [15]).

Besides this boundary bifurcation there is also another remaining problem concerning the bifurcation diagram itself. The problem starts from codimension 3 on. In this (swallowtail) catastrophe there is the possibility of encountering the simultaneous occurrence of two saddle-node bifurcations. In a standard swallowtail catastrophe this happens along a single curve in parameter space. With the information provided in the statement of Theorem 4.1 it is not possible to be sure of this result in a fixed neighbourhood of Y_0, uniformly in u.

Even the extra information on the derivatives $\frac{\partial^{j+1} c}{\partial \lambda_i \partial Y^j}$ that we obtained in Sect. 4.4, during the proof of Theorem 4.1, does not yet seem to suffice. So further analysis is required concerning the so-called Riemann–Hugoniot bifurcation set of the catastrophe.

References

1. Benoit, E.: Relation entrée-sortie. C R Acad. Sci. Paris. Ser. I. Math. **293**, 293–296 (1981)
2. Benoit, E., Callot, J.F., Diener, F., Diener, M.: Chasse au canard. Collect. Math. **31–32**, 37–119 (1981)
3. Copell, W.A.: Some quadratic systems with at most one limit cycle. Dyn. Rep. **2**, 61–68 (1988)
4. Dumortier, F.: Slow divergence integral and balanced canard solutions. Qual. Theory Dyn. Syst. **10** 65–85 (2011)
5. Dumortier, F.: Sharp upperbounds for the number of large amplitude limit cycles in polynomial Liénard equations. Discrete Contin. Dyn. Syst. Ser. A **32**, 1465–1479 (2012)
6. De Maesschalck, P., Dumortier, F.: Time analysis and entry-exit relation near planar turning points. J. Differ. Equ. **215**, 225–267 (2005)
7. De Maesschalck, P., Dumortier, F.: Canard solutions at non-generic turning points. Trans. Am. Math. Soc. **358**, 2291–2334 (2006)
8. De Maesschalck, P., Dumortier, F.: Slow–fast Bogdanov–Takens bifurcations. J. Differ. Equ. **250**, 1000–1025 (2011)
9. De Maesschalck, P., Desroches, M.: Time analysis and entry-exit relation near planar turning points. Preprint

10. Dumortier, F., Li, C.: Quadratic Liénard equations with quadratic damping. J. Differ. Equ. **139**, 41–59 (1997)
11. Dumortier, F., Roussarie, R.: Canard cycles and center manifolds. Mem. Am. Math. Soc. **577**, 1–96 (1996)
12. Dumortier, F., Roussarie, R.: Multiple canard cycles in generalized Liénard equations. J. Differ. Equ. **174**, 1–29 (2001)
13. Dumortier, F., Roussarie, R.: Bifurcation of relaxation oscillations in dimension two. Discrete Contin. Dyn. Syst. Ser. A **19**, 631–674 (2007)
14. Dumortier, F., Roussarie, R.: Birth of canard cycles. Discrete Contin. Dyn. Syst. Ser. S **2**, 723–781 (2009)
15. Dumortier, F., Roussarie, R., Sotomayor, J.: Generic 3-parameter families of planar vector fields, unfoldings of saddle, focus and elliptic singularities with nilpotent linear parts. In: Bifurcations of planar vector fields: nilpotent singularities and Abelian integrals. Lecture Notes in Mathematics, vol. 1480, pp. 1–164. Springer, Berlin (1991)
16. Krupa, M., Szmolyan, P.: Relaxation oscillation and canard explosion. J. Differ. Equ. **174**, 312–368 (2001)
17. Golubitsky, M., Guillemin, V.: Stable Mappings and Their Singularities. Graduate Texts in Mathematics, vol. 14. Springer, Berlin (1973). ISBN 0-387-90072-1

Chapter 5
Bifurcation for Non-smooth Dynamical Systems via Reduction Methods

T. Küpper, H.A. Hosham, and D. Weiss

Dedicated to Jürgen Scheurle in celebration of his 60th birthday and his many extraordinary contributions to dynamical systems

Abstract Due to the presence of discontinuities, non-smooth dynamical systems (PWS) present a wide variety of bifurcations, which cannot be explained by the classical theory, for instance, transition from sticking to sliding due to friction and sudden loss of stability as typically observed in mechanics. These phenomena are due to interactions between the boundaries and the phase trajectories that cross them from one region to another. In the present work, we review the concept of invariant sets given as cone-like objects which has turned out as an appropriate generalization of the notion of center manifolds. The existence of invariant cones containing a segment of sliding orbits and stability properties of those cones are also investigated. Based on these results we present new bifurcation phenomena in a class of 3D-PWS concerning sliding modes. Further we show that the dynamics within the sliding motion area is described by a simple one-dimensional equation. We illustrate various forms of bifurcation, stick-slip motion, and the reduction procedure by a six-dimensional brake system given by three coupled oscillators.

T. Küpper (✉) · H.A. Hosham
Mathematical Institute, University of Cologne, Weyertal 86-90, 50931 Köln, Germany
e-mail: kuepper@math.uni-koeln.de; hbakit@math.uni-koeln.de

D. Weiss
Mathematical Institute, University of Tübingen, Auf der Morgenstelle 10, 72076 Tübingen, Germany
e-mail: weiss@na.uni-tuebingen.de

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_5,

5.1 Introduction

Non-smooth dynamical systems occur as models in many applications related to science, engineering, economics, and control theory. Typically non-smooth effects are due to dry friction or impacts in mechanics, switches in electrical systems, control of pacemakers through external state-dependent impulses, etc.

A large class of such situations can be modelled by systems of ordinary differential equations defined on adjacent components of the phase space together with additional rules for the transition from one component to another.

A typical situation can be written in the form

$$\dot{\xi} = f_i(\xi), (\xi \in \mathscr{M}_i)(i = 1, \ldots, n),$$

where the phase space $\mathbb{R}^n$ is separated into disjoint and open sets $\mathscr{M}_i$ such that $\mathbb{R}^n = \bigcup_i^n \bar{\mathscr{M}}_i$.

The transition rules can be formulated as

$$\xi(t_+^*) = R(\xi(t_-^*)),$$

when the trajectory $\xi(t) \in \mathscr{M}_j (t < t^*)$ reaches the boundary $\partial\mathscr{M}_j$ of $\mathscr{M}_j$ at the time t^*.

Typical constellations for transition are

- Direct crossing from $\mathscr{M}_j$ to some $\mathscr{M}_i$
- Sliding in $\partial\mathscr{M}_j$ for some time
- Jumps due to impacts

Of course, the class of non-smooth systems allows more general types of equations such as algebraic components (related to DAE), PDE, or even mixed forms usually called hybrid systems.

Here we will restrict our attention to systems described by ODE.

The fundamental theory for such systems as far as an appropriate notion of the term "solution" as well as fundamental properties like existence and uniqueness are concerned have already been laid by Filippov [8,9]. Qualitative properties have been less studied. More than half a century ago investigations concerning the dynamics and bifurcations for non-smooth systems raised a rather new topic developed during the past decades. An excellent recent review is given in [5].

First investigations had been stimulated by experiments exploring phenomena related to dry friction or impacts. As an idealized experimental set up to analyze such phenomena friction and impact oscillators [1,5] in various forms have been designed which can easily be modeled by differential equations. Since experimental results show similar effects concerning for example bifurcation as were known for smooth systems, it was suggested to analyze the corresponding mathematical systems systematically and to develop appropriate tools known from classical bifurcation theory.

Questions of particular interest were related to a characterization of solutions with regard to stability, to various mechanisms of bifurcation especially to the generation of periodic orbits, to qualify systems by characteristic numbers such as Lyapunov exponents, or to establish procedures like the center manifold approach to reduce higher dimensional systems to an equivalent lower dimensional system carrying the essential dynamics.

As all these methods crucially depend on an approximation by linearization, differentiability is needed which by definition does not hold for non-smooth systems.

Hence, new approaches had to be developed. According to the degree of nonsmoothness the systems can be divided into various classes. The simplest situation is given by continuous but not differentiable systems. In that case solutions are (absolutely) continuous going along with a direct crossing from one component to another.

Discontinuous systems allow a great variety of phenomena due to an abrupt change of the corresponding vector fields, for example sliding motion, if all vector fields of the adjacent components are directed toward the boundary. The solution is then forced to remain within the boundary leading to a sliding motion which is governed by the Filippov extension [8, 9].

A particular case is given by impact systems; due to impacts the trajectory in phase space is no longer continuous but involves jumps. For mechanical systems jumps typically occur in the velocity component. As an illustrative example we refer to the motion of bells where the interaction of the coupled system of bell and clapper and the influence of impacts can be analyzed [12, 15].

Lyapunov exponents are frequently used to characterize stable resp. chaotic motion. Since the classical definition depends on properties of the linearized flow, existence for non-smooth system is not obvious. Various investigations have shown that they can be defined properly for non-smooth systems as well and that they provide a reliable tool to describe the dynamics, see [11] for a review.

Standard bifurcation theory as well is built up on linearization techniques such as the Lyapunov–Schmidt procedure or a center manifold approach.

The change of stationary solutions to periodic motion is a frequent situation in practical applications. The corresponding mathematical result in the form of Hopf-bifurcation relies on properties of the linearized system such as the crossing of a pair of complex eigenvalues through the imaginary axis. These analytical criteria do not work for non-smooth systems since there is no linearization.

The corresponding geometric analog though suggests a suitable approach. At the bifurcation point there is a switch in the basic system from a stable focus to an unstable focus via a center. This feature exists for appropriate piecewise linear systems as well and can be used to trigger the bifurcation of periodic orbits in the form of some kind of generalized Hopf-bifurcation. This approach has been carried out first for planar systems [22–24] using a simple Poincaré map. The idea to split a piecewise nonlinear system into a piecewise linear system (PWLS) and remaining terms of higher order first used for planar systems serves as an useful approach for higher dimensional systems as well. A successful technique to analyze the dynamics

of high dimensional systems is based on reduction to lower dimensional systems which contain the essential dynamics. Usually, this is done by the construction of invariant manifolds. The center manifold approach is known as a well-established procedure for such a reduction.

There are various ways to construct invariant manifolds, which are usually defined as solution of an appropriate fixed point problem in function spaces. It is common to these approaches that they are based on properties of the linearized problem, and hence differentiality is required.

Since that situation is typically not given for non-smooth systems, new methods have to be developed.

The key idea which we pursue is based on the splitting of the piecewise smooth system into a piecewise linear part and nonlinear perturbations of higher order.

For piecewise linear systems it is easy to set up a Poincaré map, to study its properties, and to define invariant sets, typically given as invariant cones.

It can then be shown that these invariant sets remain under small perturbation in the way that they are deformed to cone like surfaces. These invariant surfaces can be seen as generalization of center manifolds for piecewise smooth systems, and they can be used to reduce investigation of the bifurcation and stability. We note that such cones have already been detected in the analysis of continuous piecewise linear systems [3, 4]. Our analysis is based on the fact that the Poincaré map for the PWLS can be split into a sum of two operators representing different differentiality properties crucial for the analysis.

The motion on the cones can easily been used to illustrate the fact well known in control theory that the combination of stable systems may lead to instability. This corresponds to the situation that the cone itself is attractive, but the dynamics on the cone is unstable. In [16] Marsden and Scheurle presented a general approach to construct invariant manifolds for smooth systems by a method based on deformations of the linear part. It would be an interesting project to investigate if that approach could be carried over with piecewise linear systems used as a base.

For 3D continuous piecewise smooth systems with two zones a normal form has been derived [3, 4]. In the case of discontinuous systems this is more complicated. Preliminary results have been obtained by Weiss [18].

In addition, discontinuous piecewise linear systems exhibit a more complicated behavior, such as sliding or constellations with multiple attractive invariant cones.

Within the sliding area the dimension is reduced anyway. For piecewise linear smooth system the dimension can be further reduced due to the special homogeneous form of the Filippov extension. This reduction allows a simplified analysis of the sliding motion dynamics; in particular for three-dimensional systems the situation turns out to be simple, since the sliding motion can be split into components separated by invariant manifolds.

The dynamics within the sliding area may as well lead to further bifurcations; various situations will be illustrated by examples. For higher dimensional systems the situation is more complicated.

The concept of generalized invariant “manifold” carries over to the case that sliding is involved. The proof is obvious if the flow in the sliding area is linear which

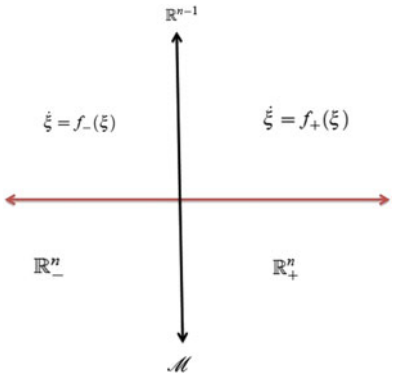

Fig. 5.1 Two half-spaces separated by a hyperplane

holds under certain conditions. The general results need some extra care, which will be carried out elsewhere [20].

5.2 General Setting

To describe the results we use a simplified setting of a piecewise smooth system (PWS) in $\mathbb{R}^n$ given in two half-spaces separated by a hyperplane $\mathscr{M} := \{\xi \in \mathbb{R}^n | \, h(\xi) = 0\}$, Fig. 5.1:

$$\dot{\xi} = \begin{cases} f_+(\xi), \, \xi \in \mathbb{R}^n_+, \\ f_-(\xi), \, \xi \in \mathbb{R}^n_-, \end{cases} \tag{5.1}$$

where $f_\pm : \mathbb{R}^n \to \mathbb{R}^n$ are sufficiently smooth functions and $\mathbb{R}^n$ is split into two regions $\mathbb{R}^n_+$ and $\mathbb{R}^n_-$ by the separation manifold $\mathscr{M}$ such that $\mathbb{R}^n = \mathbb{R}^n_+ \cup \mathscr{M} \cup \mathbb{R}^n_-$. The regions $\mathbb{R}^n_+$ and $\mathbb{R}^n_-$ are defined as

$$\mathbb{R}^n_+ = \{\xi \in \mathbb{R}^n | h(\xi) > 0\},$$
$$\mathbb{R}^n_- = \{\xi \in \mathbb{R}^n | h(\xi) < 0\}.$$

On the separating hyperplane $\mathscr{M}$ we need additional rules describing the interaction:

(a) Direct transversal crossing.
Let $\rho(\xi) = n^{\mathrm{T}}(\xi) f_+(\xi) n^{\mathrm{T}}(\xi) f_-(\xi)$, (the normal vector $n(\xi)$ perpendicular to the manifold $\mathscr{M}$ is given by $n(\xi) = \frac{\nabla h(\xi)}{\|\nabla h(\xi)\|_2}$). Then, the direct crossing set is defined as $\mathscr{M}^c = \{\xi \in \mathscr{M} | \rho(\xi) > 0\}$. Further, the direct crossing set can be partitioned into two subsets: $\mathscr{M}^c_- = \{\xi \in \mathscr{M}^c \mid n^{\mathrm{T}}(\xi) f_+(\xi) < 0\}$ and $\mathscr{M}^c_+ = \{\xi \in \mathscr{M}^c \mid n^{\mathrm{T}}(\xi) f_+(\xi) > 0\}$, Fig. 5.2a.

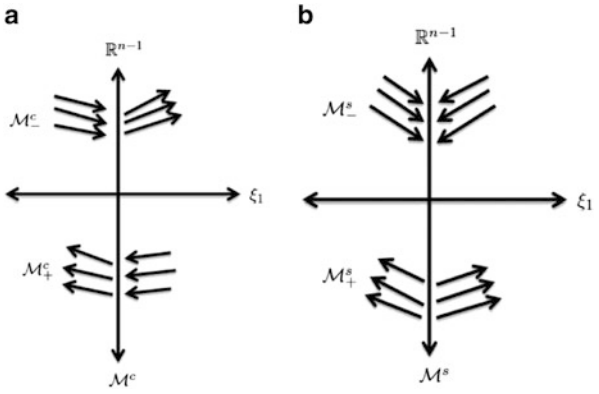

Fig. 5.2 Schematic illustration dynamics at a switching manifold. (**a**) Transversal crossing; (b) sliding mode

A simple system showing direct crossing is given by the continuous but piecewise smooth linear system.

Example 5.1 ([4]).

Set $h(\xi) = e_1^{\mathrm{T}}\xi$, $f_\pm(\xi) = A^\pm\xi$, $A^\pm = \begin{pmatrix} t^\pm & -1 & 0 \\ m^\pm & 0 & 1 \\ d^\pm & 0 & 0 \end{pmatrix}$,

here both matrices satisfy the continuity relation $A^+ - A^- = (A^+ - A^-)e_1e_1^{\mathrm{T}}$. Hence, all trajectories of this system approaching the hyperplane $\mathscr{M}$ cross it immediately and for such initial condition, there is a unique absolutely continuous solution.

(b) Sliding on $\mathscr{M}$.
The sliding mode set is defined as $\mathscr{M}^s = \{\xi \in \mathscr{M} \mid \rho(\xi) \leq 0\}$. This set is further classified as attracting $\mathscr{M}^s_-$ or repulsive $\mathscr{M}^s_+$

$$\mathscr{M}^s_- = \{\xi \in \mathscr{M}^s \mid n^{\mathrm{T}}(\xi)f_+(\xi) < 0\},$$
$$\mathscr{M}^s_+ = \{\xi \in \mathscr{M}^s \mid n^{\mathrm{T}}(\xi)f_+(\xi) > 0\}.$$

If $\xi \in \mathscr{M}^s_-$, then the vector field of both systems at ξ points toward $\mathscr{M}^s$, hence the flow cannot leave $\mathscr{M}^s$ at ξ; $\mathscr{M}^s_-$ is called attractive sliding area.

If $\xi \in \mathscr{M}^s_+$, then both vector fields are directed away from $\mathscr{M}^s$ at ξ; hence the flow in forward time is not uniquely defined at ξ and $\mathscr{M}^s_+$ is called repulsive, Fig. 5.2b. The flow in $\mathscr{M}^s$ itself is governed by Filippov's extension:

$$\dot{\xi} = \frac{n^{\mathrm{T}}(\xi)f_-(\xi) \cdot f_+(\xi) - n^{\mathrm{T}}(\xi)f_+(\xi) \cdot f_-(\xi)}{n^{\mathrm{T}}(\xi)(f_-(\xi) - f_+(\xi))}. \tag{5.2}$$

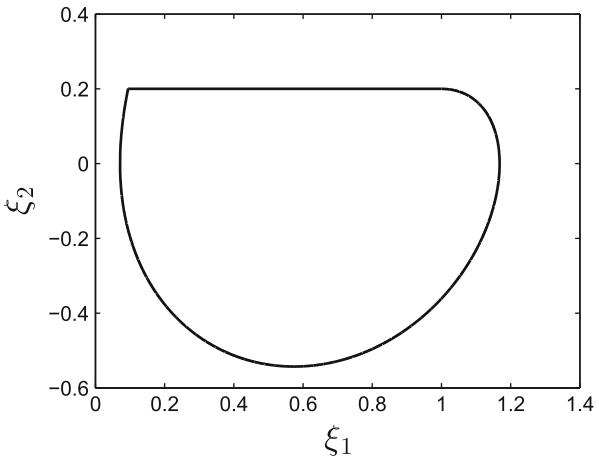

Fig. 5.3 Periodic orbit comprising a sliding segment

Example 5.2 ([7]).
Set: $h(\xi) = \xi_2 - 0.2$ and

$$f_+(\xi) = \begin{pmatrix} \xi_2 \\ -\xi_1 - \frac{1}{0.8+\xi_2} \end{pmatrix}, \quad f_-(\xi) = \begin{pmatrix} \xi_2 \\ -\xi_1 + \frac{1}{1.2-\xi_2} \end{pmatrix}.$$

We can define the sliding region as: $\mathscr{M}^s = \{\xi \in \mathscr{M}, (\xi_1 + 1)(\xi_1 - 1) < 0\}$ which is attractive, i.e., $\mathscr{M}^s = \mathscr{M}^s_-$ where $\mathscr{M}^s_- = \{\xi \in \mathscr{M}^s, \xi_1 \in (-1, 1)\}$. Therefore using the sliding vector field (5.2), we obtain F_s as:

$$F_s = \begin{pmatrix} 0.2 \\ 0 \end{pmatrix}.$$

Thus, the sliding flow in ξ_1 grows linearly within $\mathscr{M}^s_-$ until it reaches the boundary of sliding at $\xi_1 = 1$. In Fig. 5.3, we show the periodic orbit containing a segment of sliding motion.

(c) Jumps in phase space
Impacts at a specific time t^i $(i = 1, 2, \ldots)$ will cause a jump in phase space due to an impact rule

$$\xi(t^i_+) = R(\xi(t^i_-)),$$

where t_- and t_+ are the instants of time immediately before and after an impact. Usually Newton's impact rule is used and formulated as reflection of the velocity at the impact point together with some damping. Typical examples are given by impact oscillators.

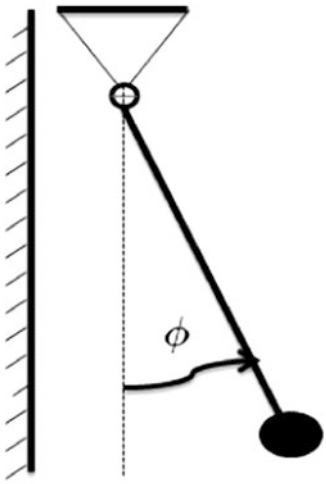

Fig. 5.4 Schematic illustration of impact

Example 5.3 (Impact Pendulum [2]).
The dynamics of the impact pendulum (Fig. 5.4) between impacting is described by the equations:

$$\ddot{\phi}(t) + \sin\phi(t) = g(t), \quad -\hat{\phi} < \phi(t),$$
$$\left.\begin{array}{c}\phi(t_+) = \phi(t_-) \\ \dot{\phi}(t_+) = -r\dot{\phi}(t_-)\end{array}\right\}, \text{if } \phi(t) = -\hat{\phi}$$

where $r \in (0, 1]$ denotes some factor reflecting damping.

Example 5.4 (Bells as Impacting System).
Bells are a nice example for impacting systems with state-dependent impacts. Following Veltmann's analysis [17] with regard to the large Emperor's bell in the Cathedral of Cologne the system of bell and clapper can be modelled as a (forced) double pendulum.

While Veltmann has derived this system to understand the curious behavior why the Cathedral of Cologne did not ring when installed in 1887, the system of bell and clapper provides an example of an impacting system showing typical behavior such as multiple impacts eventually leading to grazing. Following [12, 15] the non-dimensionalized equations of motion for a model of an impacting-contact model of a bell and clapper takes the form

$$\mathbf{M}(\Phi)\ddot{\Phi} + \mathbf{B}(\Phi)\dot{\Phi}^2 + \mathbf{C}\dot{\Phi} + \mathbf{D}(\Phi) = \mathbf{F}(\tau), \tag{5.3}$$

with impact events

$$\dot{\varphi}_2(\tau_+) - \dot{\varphi}_1(\tau_+) = \mu\big(\dot{\varphi}_2(\tau_-) - \dot{\varphi}_1(\tau_-)\big), \text{when } \varphi_2 = \varphi_1 \pm \frac{\psi}{2} \tag{5.4}$$

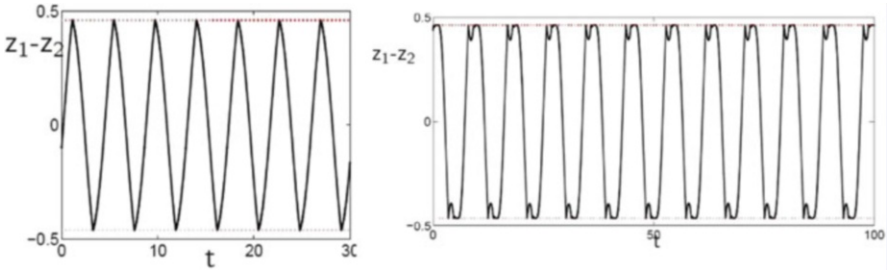

Fig. 5.5 Quasiperiodic solution associated with grazing bifurcations and multiple impact (chattering) [figures produced in cooperation with P. Piiroinen and J. Mason]

where $\Phi = [\varphi_1, \varphi_2]^{\mathrm{T}}$ represent the angular displacement of the motion of bell and the clapper, ψ is the maximum angular displacement of the clapper, $\dot{\varphi}_{1,2}(\tau_-)$ and $\dot{\varphi}_{1,2}(\tau_+)$ are the velocities immediately before and after impact, respectively, $\mathbf{C}$ is a constant damping coefficient and

$$\mathbf{M}(\Phi) = \begin{pmatrix} \alpha & \varepsilon\cos(\varphi_1 - \varphi_2) \\ \varepsilon\cos(\varphi_1 - \varphi_2) & 1 \end{pmatrix}, \quad \mathbf{B}(\Phi) = \begin{pmatrix} 0 & \varepsilon\sin(\varphi_1 - \varphi_2) \\ -\varepsilon\sin(\varphi_1 - \varphi_2) & 0 \end{pmatrix},$$

$$\mathbf{D}(\Phi) = \begin{pmatrix} \gamma\sin(\varphi_1) \\ \sin(\varphi_2) \end{pmatrix}, \quad \mathbf{F}(\tau) = \begin{pmatrix} \hat{\psi}\cos(\omega_0\tau) \\ 0 \end{pmatrix}, \dot{\Phi}^2 = (\dot{\varphi_1^2}, \dot{\varphi_1^2})^T.$$

Here, $\alpha, \varepsilon, \hat{\psi}$, and γ are constant parameters.

System (5.3) and (5.4) can be written as 5D-PWS

$$\frac{d}{dt}\mathrm{Z} = f(\mathrm{Z}(\tau)), \quad \mathrm{Z} = [\varphi_1, \varphi_2, \dot{\varphi}_1, \dot{\varphi}_2, \omega_0\tau]$$

Multiple impact and grazing bifurcation behavior is shown in Fig. 5.5.

5.3 Concept of Generalized Center Manifolds

For smooth systems $\dot{\xi} = f(\xi), \xi \in \mathbb{R}^n$ the center manifold approach can be used to determine the dynamics near a special solution $\bar{\xi}$ by reducing the system to a smaller system, which can be employed to investigate bifurcation, stability, and the dependence on critical parameters.

As an essential tool for the analysis linearization techniques are used. Since such properties are not at hand for non-smooth systems near a special solution, new methods have to be developed.

Following [25], we assume that $\bar{\xi} = 0 \in \mathscr{M}$ is a stationary solution of our system which will be written as:

$$\dot{\xi} = f_{\pm}(\xi, \lambda) = \underbrace{A^{\pm}(\lambda)\xi}_{basic\ linear\ term} + \underbrace{g^{\pm}(\xi, \lambda),}_{nonlinear\ term} \quad \lambda \in \mathbb{R},\ \pm e_1^{\mathrm{T}}\xi > 0, \tag{5.5}$$

where $A^{\pm}$ are constant (parameters dependent) matrices and $g^{\pm}$ denote smooth functions of higher order terms.

We first review the situation for planar piecewise smooth system to investigate mechanisms leading to the generation of periodic orbits. The standard procedure for smooth system is given by Hopf-bifurcation triggered by the crossing of exactly one pair of eigenvalues through the imaginary axis. For a piecewise smooth system the notion of eigenvalues is not at hand, but it turned out that instead of this analytical criterium the geometric correspondent remains available. Geometrically Hopf-bifurcation occurs, when the stationary solution changes from a stable focus to an unstable focus via a center. It turns out that this situation carries over to non-smooth systems. The formal procedure relies on the construction of a Poincaré map for the two-dimensional piecewise linear system $\dot{\xi} = A^{\pm}(\lambda)\xi$ of the form

$$P(\xi_2, \lambda) = e^{\pi b(\lambda)}\xi_2, \quad b(\lambda) = \alpha^{+}(\lambda)/\omega^{+}(\lambda) + \alpha^{-}(\lambda)/\omega^{-}(\lambda).$$

We consider the following assumptions:

(H1) $f_{\pm}(\xi, \lambda)$ are $\mathbf{C}^k$-smooth $(k \geq 2)$ for $(\xi, \lambda) \in \mathbb{R}_2^{\pm} \times \mathbb{R}$.
(H2) $f_{\pm}(0, \lambda) \equiv 0$ for $\lambda \in \mathbb{R}$.
(H3) The spectrum of $A^{\pm}(\lambda)$ consists of a pair of complex conjugate eigenvalues $\alpha^{\pm}(\lambda) \pm i\omega^{\pm}(\lambda)$, $\omega(\lambda) > 0$ for $\lambda \in \mathbb{R}$.
(H4) $a_{12}^{\pm} > 0$ or $a_{12}^{\pm} < 0$.
(H5) transversality condition, $b(0) = 0$, $\frac{db}{d\lambda}(0) \neq 0$.

Under the previous assumptions, the main result is given in the following theorem .

Theorem 5.1 ([25]).
Suppose that $(H1) - (H5)$ hold, then there bifurcates a continuous branch of periodic orbits for the planar PWS (5.5) from the origin at $\lambda = 0$.

Example 5.5 (Brake System for a Bike [25]).
The mathematical model is a system of two differential equations:

$$\begin{aligned} m\ddot{x} + d_1\dot{x} + c_1 x &= \sigma^{+}(x, \dot{x}, \lambda), && \text{if } x > 0 \\ m\ddot{x} + (d_1 + d_2)\dot{x} + (c_1 + c_2)x &= \sigma^{-}(x, \dot{x}, \lambda), && \text{if } x < 0 \end{aligned} \tag{5.6}$$

where the mass rests on a smooth surface and is connected to the walls by springs c_j and dampers d_j, $j = 1, 2$, $\sigma^{\pm}$ representing external force and λ is a free parameter.

Set $x = \mathrm{u}$, $\dot{x} = \mathrm{v}$, and $m = 1$; without loss of generality [25], we assume that the system (5.6) is of the following form

$$\begin{pmatrix} \dot{\mathrm{u}} \\ \dot{\mathrm{v}} \end{pmatrix} = \begin{pmatrix} \mathrm{v} \\ (\mathrm{b}_1^{\pm}\lambda - \mathrm{b}_0^{\pm})\mathrm{v} - a^{\pm}\mathrm{u} - \beta^{\pm}\mathrm{u}^3 \end{pmatrix}. \tag{5.7}$$

Note that the origin is always an equilibrium point, hence the eigenvalues corresponding to the linearization of (5.7) at the origin are given by $\alpha^{\pm}(\lambda) \pm i\omega^{\pm}(\lambda) = -\frac{1}{2}(b_0^{\pm} - b_1^{\pm}\lambda) \pm i\frac{1}{2}\big(4a^{\pm} - (b_0^{\pm} - b_1^{\pm}\lambda)^2\big)^{1/2}$.

The assumption (H1)–(H5) hold if $b_0^- = -(a^-/a^+)^{1/2}b_0^+$. Therefore, generalized Hopf-bifurcation occurs as the parameter λ crosses 0.

We assume that the transition at $\mathscr{M}$ is determined by the vector field of (5.1), i.e., here we first consider either direct transition or sliding but no jumps in phase space.

As an interesting example for systems of the form we use the following brake system suggested by K. Popp (1998, private communication), consisting of three coupled oscillators connected by friction forces. To capture realistic friction behavior we have slightly extended the system by allowing a general friction characteristic μ_2.

Example 5.6 (Brake System).
A brake pad 1 on a rigid frame acts on a brake disc 2. Between brake pad and brake disc there is a relative displacement with constant velocity $v > 0$, Fig. 5.6; for that reason friction forces depend only on the normal force and the kinematic friction μ_1. The coefficients of the linear viscous dampers are represented by d_1, d_2 and spring constants are denoted by c_1, c_2. Therefore, the brake pad is equipped with three mechanical degrees of freedom:

- Vertical movement x_1
- Horizontal movement x_2
- Rotation ϕ

The pad is supported via a friction contact with velocity depending friction force $R(\nu_{rel})$ by the frame where R is of the form $R(\nu_{rel}) = F_n \cdot \mu_2(\nu_{rel})$. As friction characteristic we take

$$\mu_2(\nu) = sgn(\nu)\big[\alpha_1 + \frac{\beta_1}{1 + \gamma_1|\nu|} + \delta_1\nu^2\big].$$

Note that the simple Coulomb friction characteristic is included for $\beta_1 = \delta_1 = 0$, but for $\delta_1 > 0$ care is taken to incorporate the fact that friction increases for large value of the relative velocity.

The equations of motion are given as:

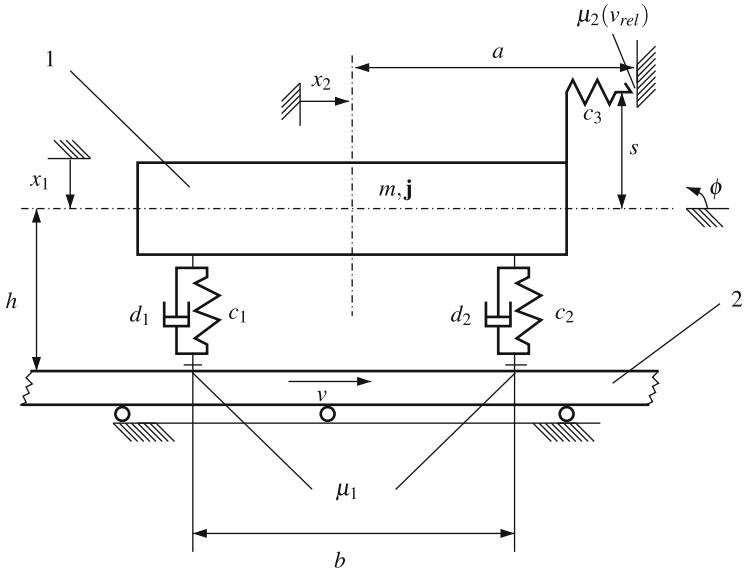

Fig. 5.6 Three-degree-of-freedom brake system model

$$m\ddot{x}_1 = -(d_1+d_2)\dot{x}_1 + \frac{b}{2}(d_2-d_1)\dot{\phi} - (c_1+c_2)x_1 + \frac{b}{2}(c_2-c_1)\phi - sgn(\dot{x}_1 - a\dot{\phi})c_3x_2\big[\alpha_1 + \frac{\beta_1}{1+\gamma_1|(\dot{x}_1 - a\dot{\phi})|} + \delta_1(\dot{x}_1 - a\dot{\phi})^2\big] \tag{5.8a}$$

$$m\ddot{x}_2 = (d_1+d_2)\mu_1\dot{x}_1 + \frac{\mu_1 b}{2}(d_1-d_2)\dot{\phi} + (c_1+c_2)\mu_1 x_1 - c_3x_2 + \frac{\mu_1 b}{2}(c_1-c_2)\phi, \tag{5.8b}$$

$$\mathbf{j}\ddot{\phi} = \big(\frac{b}{2}(d_2-d_1) + (d_1+d_2)h\mu_1\big)\dot{x}_1 - \big(\frac{b^2}{4}(d_1+d_2) + \frac{bh\mu_1}{2}(d_2-d_1)\big)\dot{\phi} - \big(\frac{b}{2}(c_1-c_2) - (c_1+c_2)h\mu_1\big)x_1 + c_3sx_2 - \big(\frac{b^2}{4}(c_1+c_2) + \frac{bh\mu_1}{2}(c_2-c_1)\big)\phi + sgn(\dot{x}_1 - a\dot{\phi})c_3ax_2\big[\alpha_1 + \frac{\beta_1}{1+\gamma_1|(\dot{x}_1 - a\dot{\phi})|} + \delta_1(\dot{x}_1 - a\dot{\phi})^2\big]. \tag{5.8c}$$

5.3.1 Brake Model as PWS

System (5.8) contains six unknown variables $(x_1, \dot{x}_1, x_2, \dot{x}_2, \phi, \dot{\phi})$ and 13 parameters. Non-smooth components enter just in two ways by the term $sgn(\dot{x}_1 - a\dot{\phi})$. It is clear that an exact analytic solution is unavailable. Our approach to such a problem is to view it as a non-smooth system. We first carry out the following transformation and scaling of t described as:

$$z_1 := x_1,\ \ z_2 := x_2,\ \ z_3 := x_1 - a\phi,\ \ z_4 := \mu_1\dot{x}_1,\ \ z_5 := \dot{x}_2,\ \ z_6 := \dot{x}_1 - a\dot{\phi},\ \ t \to ma\mu_1 t,$$

where $a, m, \mu_1 > 0$, which has no effect on the solution behavior of the model system.

To be specific, we rewrite (5.8) by using the above transformation as an equivalent six-dimensional system as follows:

$$\dot{z} = \begin{cases} A^+ z + g^+(z), & z_6 > 0, \\ A^- z + g^-(z), & z_6 < 0, \end{cases} \tag{5.9}$$

with the simple form of the matrices

$$A^{\pm} = \begin{pmatrix} 0 & 0 & 0 & a_{14} & 0 & 0 \\ 0 & 0 & 0 & 0 & a_{25} & 0 \\ 0 & 0 & 0 & 0 & 0 & a_{25} \\ a_{41} & \mp\alpha & a_{43} & a_{44} & a_{45} & a_{46} \\ a_{51} & a_{52} & a_{53} & a_{54} & 0 & a_{56} \\ a_{61} & \mp\beta & a_{63} & a_{64} & a_{65} & a_{66} \end{pmatrix}, g^{\pm} = \mp \begin{pmatrix} 0 \\ 0 \\ 0 \\ c_3 a\mu_1^2 z_2(\tilde{\epsilon} z_6 + \tilde{\tilde{\epsilon}} z_6^2) + \dots \\ 0 \\ -ac_3\mu_1(1 + \frac{ma^2}{\mathbf{j}}) z_2(\tilde{\epsilon} z_6 + \tilde{\tilde{\epsilon}} z_6^2) + \dots \end{pmatrix} \tag{5.10}$$

where
$a_{14} = ma$, $a_{25} = ma\mu_1$, $a_{41} = a\mu_1^2(c_1 + c_2) - \frac{b\mu_1^2(c_2-c_1)}{2}$, $\alpha = ac_3\mu_1^2\mu_2^0$, $a_{43} = -\frac{b\mu_1^2(c_2-c_1)}{2}$, $a_{44} = -a\mu_1(d_1 + d_2) + \frac{b\mu_1(d_2-d_1)}{2}$, $a_{46} = -\frac{b\mu_1^2}{2}(d_2 - d_1)$, $a_{51} = a\mu_1^2(c_1 + c_2) + \frac{b\mu_1^2(c_2-c_1)}{2}$, $a_{52} = -ac_3\mu_1$, $a_{53} = \frac{-b\mu_1^2(c_2-c_1)}{2}$, $a_{54} = a\mu_1(d_1 + d_2) + \frac{b\mu_1(d_1-d_2)}{2}$, $a_{56} = -\frac{b\mu_1^2}{2}(d_1 - d_2)$, $a_{61} = \frac{b\mu_1}{2}(c_2 - c_1) - a\mu_1(c_1+c_2) + \frac{ma^2\mu_1}{\mathbf{j}}(\frac{b}{2}(c_1-c_2) - (c_1+c_2)h\mu_1) + \frac{ma\mu_1}{\mathbf{j}}(\frac{b^2}{4}(c_1+c_2) + \frac{bh\mu_1}{2}(c_2-c_1))$, $\beta = \mp(ac_3\mu_1\mu_2^0 + \frac{a^3c_3m\mu_1\mu_2^0}{\mathbf{j}}) - \frac{a^2c_3sm\mu_1}{\mathbf{j}}$, $a_{63} = \frac{b\mu_1}{2}(c_1-c_2) - \frac{ma\mu_1}{\mathbf{j}}(\frac{b^2}{4}(c_1+c_2) + \frac{bh\mu_1}{2}(c_2 - c_1))$, $a_{64} = -a(d_1 + d_2) - \frac{ma^2}{\mathbf{j}}(\frac{b}{2}(d_2 - d_1) + (d_1 + d_2)h\mu_1) + \frac{ma}{\mathbf{j}}(\frac{b^2}{4}(d_1 + d_2) + \frac{bh\mu_1}{2}(d_2 - d_1))$, $a_{65} = \frac{b\mu_1}{2}(d_2 - d_1)$, $a_{66} = -\frac{b\mu_1}{2}(d_2 - d_1) - \frac{ma\mu_1}{\mathbf{j}}(\frac{b^2}{4}(d_1+d_2) + \frac{bh\mu_1}{2}(d_2-d_1))$, $\mu_2^0 = \alpha_1+\beta_1$, $\tilde{\epsilon} = -\alpha_1\gamma_1$, $\tilde{\tilde{\epsilon}} = 2(\delta_1+\alpha_1+\gamma_1^2)$.

For PWS it is necessary to know the direction of the flow of the vector field when the trajectory reaches $\mathscr{M}$. We will discuss the vector field on $\mathscr{M}$ in two main cases, namely direct crossing through $\mathscr{M}$ or sliding motion on $\mathscr{M}$ where the sliding surface is particularly important with regard to the friction coefficient.

5.3.2 Detecting Crossing and Sliding Regions

In this section we demonstrate the existence of a crossing and sliding mode from the point of view of a Filippov system. Let $\Upsilon(z) = a_{61}z_1 + a_{63}z_3 + a_{64}z_4 + a_{65}z_5$. The direct crossing in $\mathscr{M}^c$ for $z_6 = 0$ occurs if both quantities $[n^{\mathrm{T}}(z)f_{\pm}(z)]$ have the same sign. Therefore, the crossing region $\mathscr{M}^c := \{z \in \mathscr{M} | \Upsilon(z)^2 - (\beta z_2)^2 > 0\}$ is divided into two main regions, namely

$$\mathscr{M}^c_+ := \{z \in \mathscr{M}^c | \Upsilon(z) > \beta z_2\},$$
$$\mathscr{M}^c_- := \{z \in \mathscr{M}^c | \Upsilon(z) < \beta z_2\}.$$

In a similar way, we can define the sliding mode region as $\mathscr{M}^s := \{z \in \mathscr{M} | \Upsilon(z)^2 - (\beta z_2)^2 \leq 0\}$ which is divided into two main regions, namely

$$\mathscr{M}^s_- := \{z \in \mathscr{M}^s | \Upsilon(z) < \beta z_2\},$$
$$\mathscr{M}^s_+ := \{z \in \mathscr{M}^s | \Upsilon(z) > \beta z_2\},$$

where we use the notation $\mathscr{M}^s_-$ to represent the attractive sliding motion and $\mathscr{M}^s_+$ to represent repulsive sliding motion.

5.4 Piecewise Smooth Linear System

PWLS are extensively used to model many physical phenomena such as mechanical devices [6] or electronic circuits [21]. We consider the n-dimensional piecewise smooth linear system :

$$\dot{\xi} = \begin{cases} A^+\xi, & h(\xi) > 0, \\ A^-\xi, & h(\xi) < 0, \end{cases} \tag{5.11}$$

where $\xi \in \mathbb{R}^n$ and $A^{\pm}$ are $n \times n$ real matrices. For that setting the stationary solution is always located within the separating manifold. We are interested to investigate the dynamical behavior in a neighborhood of the stationary solution and in particular to study the generation of periodic orbits. Other questions concern stability and the possibility to reduce the system to a lower dimensional one.

5.4.1 Concepts of Invariant Cones

PWLS can be classified in two classes depending on the degree of smoothness properties of the associated vector field, namely continuous PWLS (non-sliding flow) and discontinuous PWLS (sliding flow).

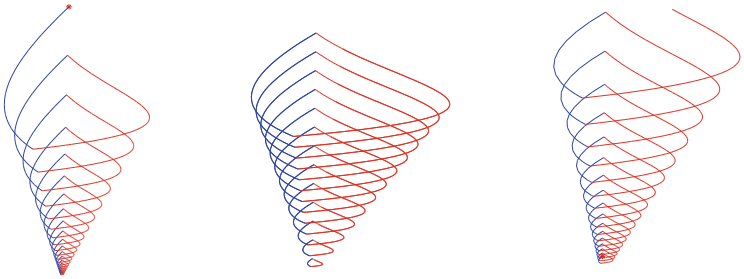

Fig. 5.7 Different dynamics on cones, $\mu_c < 1$, $\mu_c = 1$ and $\mu_c > 1$, respectively

To analyze the dynamical behavior of n-dimensional PWLS (5.11), we assume that both matrices have at least a pair of complex conjugated eigenvalues introducing rotations in the system. For initial values $\xi \in \mathscr{M}$ for which $n^{\mathrm{T}}(\xi)A^{\pm}\xi$ have both negative sign and for which $e^{t_-(\xi)A^-}\xi$ reaches $\mathscr{M}$ again for the first time $t_-(\xi)$ at η, we define the Poincaré map $P_-(\xi) := e^{t_-(\xi)A^-}\xi$ and similarly $P_+(\eta) := e^{t_+(\eta)A^+}\eta$.

If both P_- and P_+ are well defined so that $P_+(P_-(\xi))$ exists, we can study the behavior of the combined map $P_+ \circ P_-$. If there exists $\tilde{\xi} \in \mathscr{M}$ such that

$$P(\tilde{\xi}) = \mu_c \tilde{\xi},$$

for some $\mu_c > 0$, then the same holds for the half-ray $\{\lambda\tilde{\xi} \,|\lambda > 0\}$.

In that way an invariant cone is generated by the flow of (5.11).

The "eigenvalue" parameter μ_c determines the dynamics on the cone; if $\mu_c < 1$ resp. $\mu_c > 1$ the flow on the cone spirals in resp. out; for $\mu_c = 1$ the cone is foliated by periodic orbits of (5.11), Fig. 5.7. The "eigenvalue" μ_c of P is an eigenvalue of the linear operator DP evaluated at $\tilde{\xi}$ as well.

Attractivity of the invariant cone is determined by the remaining $(n-2)$ eigenvalues of $DP(\tilde{\xi})$.

Theorem 5.2 ([13]).
If there exists $\tilde{\xi} \in \mathscr{M}_-^c$ and $\mu_c > 0$ such that

$$P(\tilde{\xi}) = \mu_c \tilde{\xi},$$

then $\tilde{\xi}$ generates an invariant cone under the flow of (5.11) due to $P(\lambda\tilde{\xi}) = \lambda P(\tilde{\xi}) = \lambda\mu_c\tilde{\xi}$; moreover,

(i) If $\mu_c > 1$, then the stationary solution 0 is unstable.
(ii) If $\mu_c = 1$, then the cone consists of periodic orbits.
(iii) If $\mu_c < 1$, then the stability of 0 depends on the stability of P with respect to the complimentary directions.

The existence of an invariant cone for 3D problems has been studied in detail in [10, 14]; further a general nonlinear determining system to compute the generating vector ξ has been set up in [13].

For homogenous and continuous 3D-PWLS with two zones existence of invariant cones and their bifurcations have already been studied in [3, 4]. There also, an example is given demonstrating that the combination of two stable systems may be unstable; an effect already known in control theory.

This nevertheless surprising result can systematically be explained in our setting by the existence of an invariant attractive cone with the property that the motion on the cone is unstable, i.e., $\mu_c > 1$.

In general coexistence of several attractive cones is possible; a result which does not hold for continuous PWLS.

Example 5.7 (Existence of Multiple Invariant Cones).
Set

$$A^- = \begin{pmatrix} \lambda^- & -1 & 0 \\ 1 & \lambda^- & 0 \\ 0 & 0 & \mu^- \end{pmatrix}, \quad A^+ = \begin{pmatrix} 2\lambda^+ & -1 & \mu^+ \\ \lambda^{+2}+1 & 0 & -\lambda^{+2}+2\lambda^+\mu^+ - 1 \\ 0 & 0 & \mu^+ \end{pmatrix}.$$

The $\ominus$-system possesses an invariant plane $\xi_3 = 0$ with constant return time $t_-(\xi) = \pi$. For the $\oplus$-system the line $\xi_3 = \frac{1}{\mu^+}\xi_2$ determines the boundary of the sliding motion area in the (ξ_2, ξ_3)-plane.

Note that vanishing or appearing of a sliding area is based on one parameter μ^+.

Here, we assume that the starting point $\xi \in \mathscr{M}_+^c$ hence $\xi_2 < 0$, and the Poincaré map $P = P_- \circ P_+(\xi)$ mapping (ξ_2, ξ_3) into itself (i.e., $P(\xi) : \mathscr{M}_+^c \to \mathscr{M}_+^c$) is given by

$$P(\xi) = F \begin{pmatrix} \lambda^+ \sin(t_+) - \cos(t_+) & \sin(t_+)(1-\lambda^{+2}) + 2\lambda^+\left(\cos(t_+) - e^{(\mu^+ - \lambda^+)t_+}\right) \\ 0 & e^{(\mu^+ - \lambda^+)t_+ + \pi(\mu^- - \lambda^-)} \end{pmatrix} \begin{pmatrix} \xi_2 \\ \xi_3 \end{pmatrix},$$

where $F = e^{\lambda^+ t_+ + \pi\lambda^-}$ and the return time $t_+(\xi)$ depends on ξ in a nonlinear linear way, and it is determined by the smallest positive solution of the following equation

$$-\sin(t_+)\xi_2 + (\lambda^+ \sin(t_+) - \cos(t_+) + e^{t_+(\mu^+ - \lambda^+)})\xi_3 = 0. \tag{5.12}$$

Lemma 5.1 ([10]).
If $\xi \in \mathscr{M}_\pm^c$ and $\lambda^+ = -\lambda^-$, then the present system has, at least, two invariant cones with periodic orbits. One of them can be asymptotically stable and the other unstable or both can be unstable foci; but there is also the situation where both invariant cones are asymptotically stable.

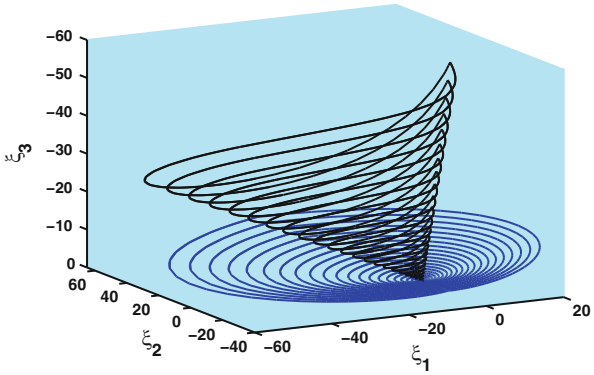

Fig. 5.8 Two attractive invariant cones, $\lambda^+ = -\lambda^- = 0.6$, $\mu^+ = -1.13$, $\mu^- = 0.4266$, where $t_+ = \pi$ for the flat cone and $t_+ = 1.1306$ for the other

Proof. Set $\xi \in \mathscr{M}_+^c$ and $\lambda^+ = -\lambda^-$. Since an invariant cone consisting of periodic orbits requires that either $\mu_{c_1} = e^{\lambda^-(\pi - t_+)}(\lambda^+ \sin(t_+) - \cos(t_+))$ or $\mu_{c_2} = e^{\mu^+ t_+ + \pi\mu^-}$ equals 1 we get

(i) $\mu_{c_1} = 1$, and $t_+ = \pi$ by direct analysis of the fixed point equation $P(\xi) = \xi$ which requires $-2\lambda^+(1 + e^{(\mu^+ - \lambda^+)\pi})\xi_3 = 0$, hence $\xi_3 = 0$. In this case we obtain a flat cone given as the invariant plane which is attractive if $\mu^+ < -\mu^-$ or repulsive if $\mu^+ > -\mu^-$.

(ii) $\mu_{c_2} = 1$, hence $t_+ = -\frac{\mu^-\pi}{\mu^+}$. The corresponding eigenvector is calculated as

$$\tilde{\xi} = \begin{pmatrix} -1 \\ -\frac{1+e^{\lambda^+ t_+ + \pi\lambda^-}\left(\lambda^+ \sin(t_+) - \cos(t_+)\right)}{e^{\lambda^+ t_+ + \pi\lambda^-}\left(\sin(t_+)(1-\lambda^{+2}) + 2\lambda^+(\cos(t_+) - e^{(\mu^+ - \lambda^+)t_+})\right)} \end{pmatrix}$$

To prove the existence of the function $t_+(\xi) \in (0, \pi)$, without loss of generality, we set $\xi \in \partial\mathscr{M}_+^s$ ($\partial\mathscr{M}_+^s$ refer to the boundary between sliding and crossing areas), then the existence of a solution for Eq. (5.12) requires $(\mu^+ - \lambda^+) < 1$. The corresponding cone is attractive, resp. repulsive if $\mu_{c_1} < 1$ resp. $\mu_{c_1} > 1$.

Figure 5.8 shows an example to illustrate the situation of two attractive invariant cones for the special choice of parameters. □

Example 5.8 (Existence of an Invariant Cone for the Linear Brake System Without Sliding Motion [10]).
For the simple Coulomb friction characteristic included by $\beta_1 = \delta_1 = 0$, the nonlinear brake system (5.8) reduces to a linear form. To simplify, we set the parameters $c := c_1 = c_2$ and $d := d_1 = d_2$. In Fig. 5.9, we fix all parameters values as in Table 5.1 and choose the friction coefficient smaller than the static

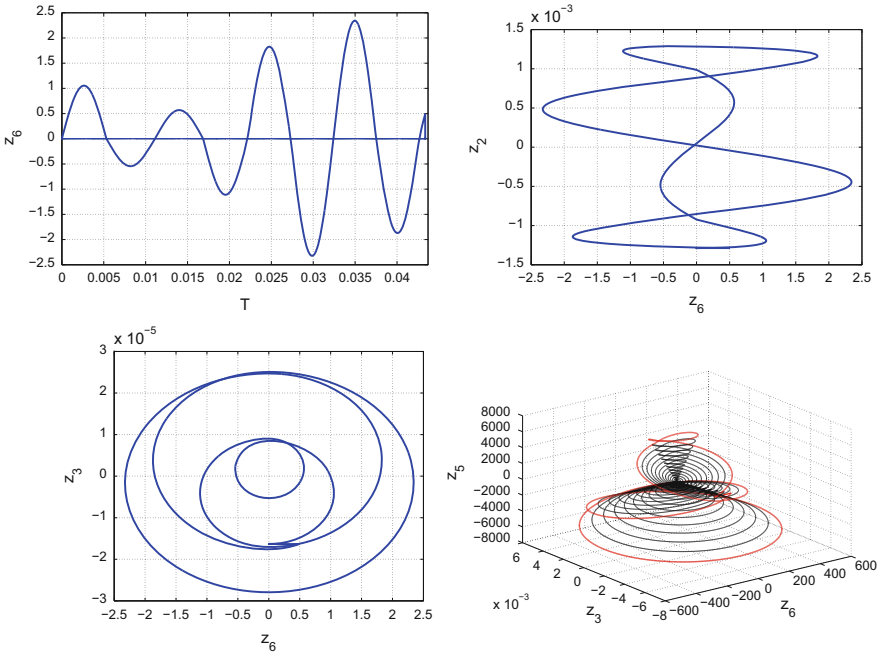

Fig. 5.9 Invariant cones and solution components for linear brake system without sliding motion

Table 5.1 Parameters are presented in K. Popp (1998, private communication)

Description	Unit	Value	Remark
m	kg	0.3	Weighed, rounded
$\mathbf{j}$	$\mathrm{kg\,m^2}$	$3 \cdot 10^{-4}$	From m and geometry
a	m	$58 \cdot 10^{-3}$	Measured
b	m	$50 \cdot 10^{-3}$	Estimated
h	m	$8 \cdot 10^{-3}$	Measured
s	m	$1 \cdot 10^{-3}$	Measured
μ_1	1	0.4	Static friction
μ_2^0	1	0.15	Kinetic friction
c_1, c_2	$\mathrm{N\,m^{-1}}$	$18 \cdot 10^8$	Spring constants
c_3	$\mathrm{N\,m^{-1}}$	$13 \cdot 10^7$	Spring constant, estimated
d_1, d_2	$\mathrm{N\,s\,m^{-1}}$	657.3	Damping coefficients

one (i.e., $\mu_2^0 \ll \mu_1$). The main reason for choosing μ_2^0 is that this choice rapidly restores the spring to a more relaxed length. Note that a change of this parameter μ_2^0 changes the control parameters $\alpha,\ \beta$ (i.e., the friction force). The parameter β in turn causes the existence of sliding and crossing regions.

Figure 5.9 shows an invariant cone, a 4-periodic orbit and solution components at $d = 0$ and $\mu_2^0 = 0.00014$ which is quite small.

5.5 PWLS with Sliding

Example 5.6 indicates that sliding motion occurs. Following Filippov the motion within $\mathscr{M}$ is determined by (5.2). For a piecewise linear system (5.11) this reduces to

$$\dot{\xi} = \frac{n^{\mathrm{T}}(\xi)A^{-}\xi \cdot A^{+}\xi - n^{\mathrm{T}}(\xi)A^{+}\xi \cdot A^{-}\xi}{n^{\mathrm{T}}(\xi)(A^{-}\xi - A^{+}\xi)} \tag{5.13}$$

In general this is a nonlinear system. Due to the homogeneity special features hold.

(a) If $\xi(t)$ is a solution, then $\lambda\xi(t)$ as well for $\lambda \geq 0$.
(b) Half-rays are mapped into half-rays, with constant time of evaluation from one half-ray to another, hence if some trajectory leaves $\mathscr{M}$ a half-ray does at the same time.
Further, if there are stationary solutions in $\mathscr{M}$, i.e., $F_s(\bar{\xi}) = 0$, then there is a half-ray of stationary solution.

For an initial position in $\mathscr{M}^s_-$ or if the flow of a subsystem of (5.11) arrives at the sliding region $\mathscr{M}^s_-$, the sliding motion can be observed along the discontinuity surface in phase space. Let $\varphi^s(t_s(\xi), \xi)$ in $\mathbf{C}^k$, $k \geq 1$, denote the sliding flow generated by solution of (5.13), and let t_s be the time spent in the $\mathscr{M}^s_-$ region.

Then we define the sliding map as

$$P_s : \mathscr{M}^s_- \to \mathscr{M}^s_-,$$
$$\xi \to P_s(\xi) = \varphi^s(t_s, \xi).$$

The existence of an invariant cone passing through the sliding region depends on the existence of an "eigenvector" $\bar{\xi} \notin \mathscr{M}^s_+$ of the nonlinear eigenvalue problem $P(\tilde{\xi}) = \mu_c\tilde{\xi}$ where P is the required composition of one or both of (P_-, P_+) and P_s.

Example 5.9 (Existence of Invariant Cone for the Linear Brake System with Sliding Motion).
For the special choice that the initial friction coefficient μ_2^0 is equal to the static $\mu_2^0 = \mu_1 = 0.4$, the complex behavior of the brake system is revealed to multiple periodic orbits including sliding. In Fig. 5.10 we show a 4-periodic orbit where a transition phase slip motions with small length appear.

Stationary solutions and invariant manifolds within $\mathscr{M}$ may strongly influence the flow in $\mathscr{M}$; in particular they can prevent trajectories to leave $\mathscr{M}$ so that the long time motion might be restricted to $\mathscr{M}$. For that reason it is worth while to investigate the flow in $\mathscr{M}$. Due to special properties the system can be reduced to a lower dimensional system or even to a linear system under additional hypotheses.

In the following we assume without restriction that $n = e_1$.
Using a suitable transformation T we can simplify system (5.13).

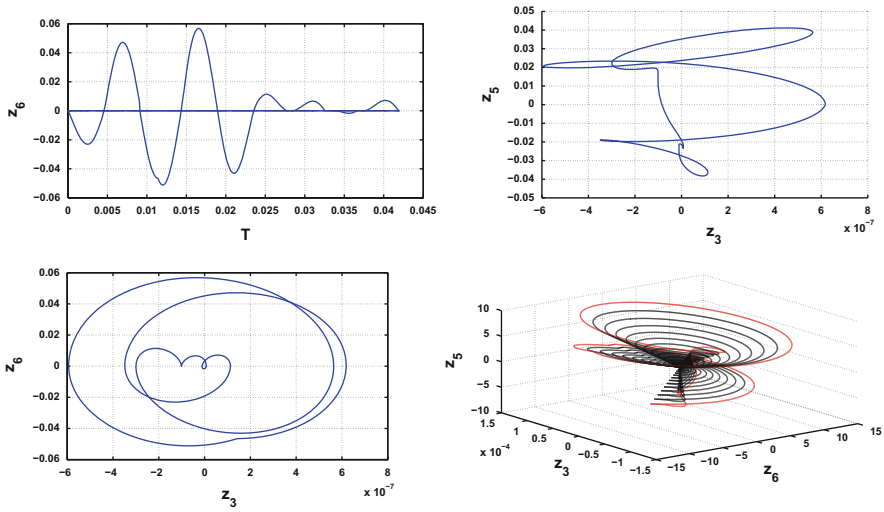

Fig. 5.10 Invariant cones and solution components, existence of 4-sliding periodic orbit when $\alpha \neq 0, \beta \neq 0$

Assume that T leaves $\mathscr{M}$ invariant, i.e., $Te_1 = e_1$ and set $T\eta := \xi$. Then

$$\dot{\eta} = \frac{1}{e_1^{\mathrm{T}}(A^- - A^+)T\eta}\big[(e_1^{\mathrm{T}}A^-T\eta)T^{-1}A^+T\eta - (e_1^{\mathrm{T}}A^+T\eta)T^{-1}A^-T\eta\big].$$

Further we can arrange that for $\eta \in \mathscr{M}$

$$e_1^{\mathrm{T}}(A^- - A^+)T\eta = \eta_j,$$

for some $j \in \{2, \ldots, n\}$; we assume without restriction that $j = n$. Then

$$\dot{\eta} = \frac{1}{\eta_n}\big[(e_1^{\mathrm{T}}A^-T\eta)T^{-1}A^+T\eta - (e_1^{\mathrm{T}}A^+T\eta)T^{-1}A^-T\eta\big].$$

Define slopes $s_i = \eta_i/\eta_n, (i = 1, \ldots, n)$. Then for $\eta \in \mathscr{M}$: $\dot{s}_1 = 0$ and $\dot{s}_n = 0$; for $i \in \{2, \ldots, n-1\}$ we obtain

$$\dot{s}_i = \big[\dot{\eta}_i\eta_n - \eta_i\dot{\eta}_n\big]/\eta_n^2.$$

By using the differential equations of $\dot{\eta}$, we obtain a reduced system for the slopes describing the motion in $\mathscr{M}$.

Lemma 5.2. *The flow in $\mathscr{M}$ is governed by the evolution of the slopes $s_i, (i = 2, \ldots, n-1)$:*

$$\begin{aligned}\dot{s}_i = {} & e_1^{\mathrm{T}} A^- Tse_i^{\mathrm{T}} T^{-1} A^+ Ts - e_1^{\mathrm{T}} A^+ Tse_i^{\mathrm{T}} T^{-1} A^- Ts \\ & -s_i \left[e_1^{\mathrm{T}} A^- Tse_n^{\mathrm{T}} T^{-1} A^+ Ts - e_1^{\mathrm{T}} A^+ Tse_n^{\mathrm{T}} T^{-1} A^- Ts \right].\end{aligned} \tag{5.14}$$

Remark 5.1. Since the right-hand side of (5.14) consists of quadratic resp. cubic (polynomial) forms statements can be drawn concerning for example the number of stationary lines.

As a special situation consider $n = 3$. Then

$$\dot{s}_1 = 0,$$

$$\begin{aligned}\dot{s}_2 = {} & e_1^{\mathrm{T}} A^- Tse_2^{\mathrm{T}} T^{-1} A^+ Ts - e_1^{\mathrm{T}} A^+ Tse_2^{\mathrm{T}} T^{-1} A^- Ts \\ & -s_2 \left[e_1^{\mathrm{T}} A^- Tse_3^{\mathrm{T}} T^{-1} A^+ Ts - e_1^{\mathrm{T}} A^+ Tse_3^{\mathrm{T}} T^{-1} A^- Ts \right] = g(s_2).\end{aligned}$$

$$\dot{s}_3 = 0.$$

Since $g(s_2)$ is either a quadratic resp. cubic polynomial in s_2, the number of possible stationary solutions is limited to at most 2 resp. 3; in a similar way stability can be obtained via the derivative of g.

Example 5.10. As a simple example to illustrate possible features resulting of sliding motion we consider a situation where A^+, A^- are chosen such that $T = I$ and $h(x) = \xi_1$. We take

$$A^+ = \begin{pmatrix} \lambda^+ & -1 & 0 \\ 1 & \lambda^+ & a_{23}^+ \\ 0 & 0 & \mu^+ \end{pmatrix}, \quad A^- = \begin{pmatrix} \lambda^- & -1 & 1 \\ a_{21}^- & 0 & a_{23}^- \\ 0 & a_{32}^- & \mu^- \end{pmatrix}.$$

Then, for $\mathscr{M} = \{\xi \in \mathbb{R}^3 \mid \xi_1 = 0\}$ the attractive sliding motion area is given by

$$\mathscr{M}_-^s = \{\xi \in \mathbb{R}^3 \mid \xi_1 = 0, \xi_3 \geq \xi_2 \geq 0\}.$$

In Fig. 5.11 we show $\mathscr{M}_-^s$ is bounded by the half-rays G_1 and G_2.

The sliding motion dynamics is governed for $s_1 = 0,\ s_3 = 1$ by :

$$\dot{s}_2 = -a_{32}^- s_2^3 + (\mu^+ - \mu^- - \lambda^+) s_2^2 + (\lambda^+ - \mu^+ + a_{23}^- - a_{23}^+) s_2 + a_{23}^+ =: \mathbf{g}(s_2),$$

$(0 \leq s_2 \leq 1)$.

Remark 5.2. (i) For the 3-dimensional system the dynamics within the sliding motion area is described by a simple equation $\dot{s}_2 = \mathbf{g}(s_2)$.
Since $\mathbf{g}$ is a polynomial of degree 3 there are at most 3 stationary solutions.

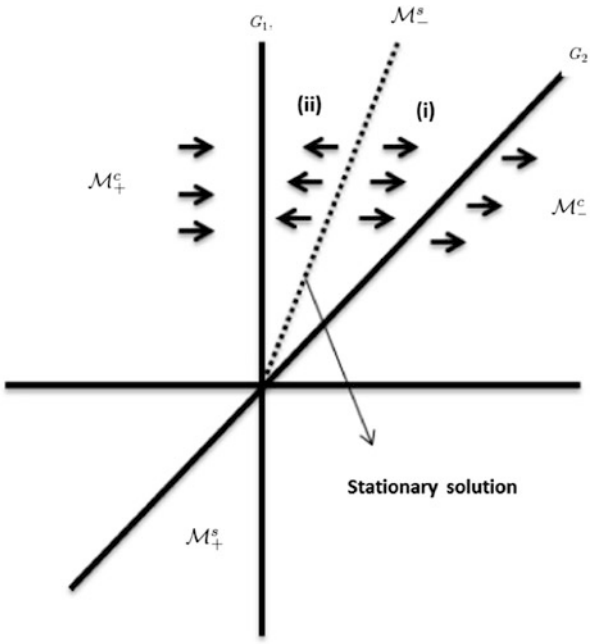

Fig. 5.11 Schematic illustration location of attractive sliding region $\mathscr{M}^s_-$ consists of stationary solution (pseudo-equilibrium line)

The degree of $\mathbf{g}$ is equal to 3 if $a^-_{32} \neq 0$. In case $a^-_{32} = 0$ there are at most 2 zeros of $\mathbf{g}$. The flow on the boundary of the sliding motion area is determined by $\dot{s}_2 = \mathbf{g}(s_2)$ for $s_2 = 0$ resp. $s_2 = 1$, hence by

$$\mathbf{g}(0) = a^+_{23},$$
$$\mathbf{g}(1) = a^+_{23} - a^-_{32} - \mu^-.$$

If $a^+_{23} > 0$ then the flow enters $\mathscr{M}^s_-$ through G_1, if $a^+_{23} < 0$ the flow leaves $\mathscr{M}^s_-$ through G_1, and if $a^+_{23} = 0$ then G_1 is invariant. In a similar way the flow on G_2 can be characterized by $\mathbf{g}(1) = a^+_{23} - a^-_{32} - \mu^-$.

(ii) If $0 \leq a^+_{23}$ and $a^+_{23} > a^-_{32} + \mu^-$, then the flow enters $\mathscr{M}^s_-$ through G_1 and leaves it through G_2 and there are either no zero of $\mathbf{g}$ in $[0, 1]$ or one stable and one unstable one.

(iii) If $0 \leq a^+_{23} < a^-_{32} + \mu^-$, then the flow enters $\mathscr{M}^s_-$ through G_1 and G_2, and there is exactly one stable zero of $\mathbf{g}$ in $[0, 1]$.

(iv) Stationary solutions of $\dot{s}_2 = \mathbf{g}(s_2)$ correspond to invariant lines for the system $\dot{\xi} = F_s(\xi)$ which can be a stable, unstable, or a center manifold separating the planar phase space.

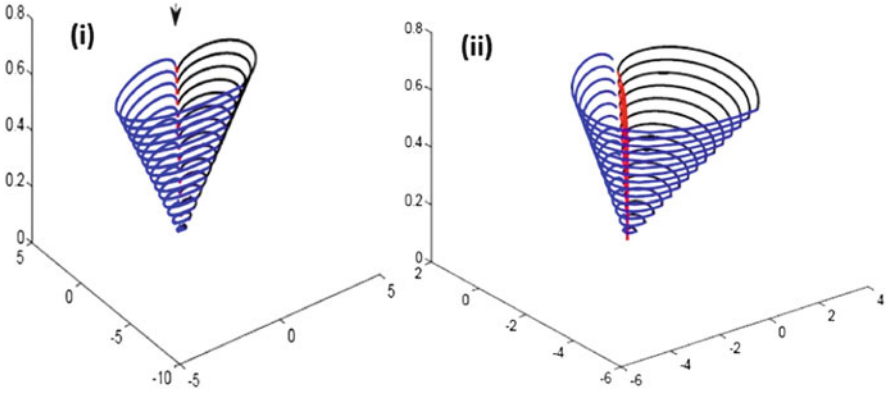

Fig. 5.12 Bifurcation of invariant cone involving sliding depending on the location of the return map, the cone is closed (consists of periodic orbits, see (i)) or destroyed with solution remaining in $\mathscr{M}_-^s$ (see (ii))

In higher dimensional systems those lines do not separate the phase space; so it is interesting to investigate if separating manifolds can be used to structure $\mathscr{M}_-^s$ in dimensions greater than 3.

(v) If the line G_1 is mapped back to $\mathscr{M}_-^s$ by the flow of the system than it depends on the structure in $\mathscr{M}_-^s$ and the location of the return if the flow remains in $\mathscr{M}_-^s$ for all times or if an invariant closed cone will be generated. Invariant lines in $\mathscr{M}_-^s$ will serve as separatrix and lead to bifurcation. Using the classification in (ii) and (iii) together with properties of P_+ and P_-, parameters can be chosen appropriately.
Illustrative examples are shown in Fig. 5.12.

(vi) The relationship

$$(A^+ - A^-)\left(\begin{array}{c|c} 0 & 0 \\ \hline 0 & I \end{array}\right) = x\, y^{\mathrm{T}},$$

for suitable vectors x and y which leads to a linear flow in $\mathscr{M}_-^s$ holds if $\lambda^+ = 0$ and a_{32}^-, x and y can be chosen as $y_1 = y_2 = 0$, $y_3 = 1$, $x_1 = -1, x_2 = a_{23}^+ - a_{23}^-$.

In special situations such as for the brake system the sliding motion is governed by a linear equation:

$$\dot{z} = \begin{pmatrix} a_{14}z_4 \\ a_{25}z_5 \\ 0 \\ (a_{41} - \frac{\alpha}{\beta}a_{61})z_1 + (a_{43} - \frac{\alpha}{\beta}a_{63})z_3 + (a_{44} - \frac{\alpha}{\beta}a_{64})z_4 + (a_{45} - \frac{\alpha}{\beta}a_{65})z_5 \\ a_{51}z_1 + a_{52}z_2 + a_{53}z_3 + a_{54}z_4 \end{pmatrix}.$$

This simplification holds under conditions which are described in [20]:

Theorem 5.3. *Assume that* $Te_1 = e_1$ *and that for suitable vectors* x, y *the relation*

$$(A^+ - A^-)T[I - e_1 e_1^{\mathrm{T}}] = xy^{\mathrm{T}},$$

holds. Then the flow within $\mathscr{M}$ *is described by the linear system*

$$\dot{\eta} = T^{-1}A^+T\eta - \frac{1}{e_1^{\mathrm{T}}x}e_1^{\mathrm{T}}A^+T\eta T^{-1}x.$$

Proof. Using

$$\begin{aligned}
\dot{\xi} &= \frac{1}{e_1^{\mathrm{T}}[A^- - A^+]\xi}\left[(e_1^{\mathrm{T}}A^-\xi)A^+\xi - (e_1^{\mathrm{T}}A^+\xi)A^-\xi\right] \\
&= \frac{1}{e_1^{\mathrm{T}}[A^- - A^+]\xi}\left[(e_1^{\mathrm{T}}A^-\xi)A^+\xi - (e_1^{\mathrm{T}}A^+\xi)A^+\xi + (e_1^{\mathrm{T}}A^+\xi)A^+\xi\right. \\
&\quad \left. -(e_1^{\mathrm{T}}A^+\xi)A^-\xi\right] \\
&= A^+\xi + \frac{(e_1^{\mathrm{T}}A^+\xi)}{e_1^{\mathrm{T}}[A^- - A^+]\xi}[A^+ - A^-]\xi
\end{aligned}$$

and

$$\begin{aligned}
\dot{\eta} &= T^{-1}A^+T\eta - \frac{(e_1^{\mathrm{T}}A^+T\eta)}{e_1^{\mathrm{T}}[A^- - A^+]T\eta}T^{-1}[A^+ - A^-]T\eta \\
&= T^{-1}A^+T\eta - \frac{(e_1^{\mathrm{T}}A^+T\eta)}{e_1^{\mathrm{T}}xy^{\mathrm{T}}\eta}T^{-1}xy^{\mathrm{T}}\eta \\
&= T^{-1}A^+T\eta - \frac{(e_1^{\mathrm{T}}A^+T\eta)}{e_1^{\mathrm{T}}x}T^{-1}x.
\end{aligned}$$

□

5.6 Nonlinear Piecewise Smooth Systems (PWNS)

Recently [19], the existence of cone-like invariant manifolds as an extension to nonlinear perturbations of certain n-dimensional non-smooth systems under appropriate conditions in the case without ***sliding motion*** carrying the essential dynamics of the full system has been proved. To see this we introduce the following hypotheses:

(a) We assume

$$\dot{\xi} = f_{\pm}(\xi) = \underbrace{A^{\pm}\xi}_{basic\ linear\ term} + \underbrace{g^{\pm}(\xi)}_{nonlinear\ term}, \quad \pm e_1^{\mathrm{T}}\xi > 0, \xi \in \mathbb{R}^n, \tag{5.15}$$

with constant matrices $A^{\pm}$ and nonlinear C^k-parts $g^{\pm}(\xi) = o(\|\xi\|)$, $k \geq 1$.

(b) Direct transition between $\mathbb{R}^n_-$ and $\mathbb{R}^n_+$ through $\mathscr{M}$, hence, without loss of generality, $\xi \in \mathscr{M}^c_-$.

(c) Existence of $\mu_c > 0$ and $\bar{\xi}$ such that $P(\bar{\xi}) = \mu_c\bar{\xi}$ for linear PWS.

(d) The attractivity condition is satisfied, i.e., the remaining (n-2) eigenvalues of $\lambda_-, \ldots, \lambda_{n-2}$ of DP satisfy $|\lambda_j| < \min\{1, \mu_c\}, (j = 1, \ldots, n-2)$.

Theorem 5.4 ([19]).
Under the previous hypotheses on the corresponding PWLS and $g_{\pm}$, there exists a sufficiently small δ and a C^1-function $H : [0, \delta) \to \mathscr{M}$ satisfying $H(0) = 0$ and $\frac{\partial}{\partial u}H(0) = \bar{\xi}$ such that

$$\{H(u) \mid 0 \leq u < \delta\}$$

is locally invariant and attractive under the Poincaré map of system (1). For $k = 2$ the function H is C^k in case of $\mu_c \geq 1$ and $C^{\min(k,j)}$ in case of $\mu_c < 1$ and $\alpha < \mu_c^j$.

Example 5.11 (Class of 3D-PWNS).
Set:

$$A^{\pm} = (\mathbf{S}_{\pm})^{-1}A_N^{\pm}\mathbf{S}_{\pm}, A_N^{\pm} = \begin{pmatrix} \lambda^{\pm} & -\omega^{\pm} & 0 \\ -\omega^{\pm} & \lambda^{\pm} & 0 \\ 0 & 0 & \mu^{\pm} \end{pmatrix}, (S_-)^{-1} = \begin{pmatrix} 1 & \frac{-\alpha(\alpha+1)}{2} & -\alpha \\ -\delta & 1 & 0 \\ 0 & -\delta & 1 \end{pmatrix},$$

$$(S_+)^{-1} = I, g^+(\xi) = \rho_+ \begin{pmatrix} 0 \\ 0 \\ \xi_1^2 + \xi_2^2 \end{pmatrix}, \; g^-(\xi) = \rho_- \begin{pmatrix} \xi_3^2 \\ 0 \\ 0 \end{pmatrix}.$$

Attractivity of the cone is guaranteed if $|\mu_1| < \min\{1, \mu_c\}$ and the invariant "eigenvector" $\bar{\xi}$ satisfying $P(\bar{\xi}) = \mu_c\bar{\xi}$ in PWLS is chosen as $\bar{\xi} = (\bar{y}, \bar{z})^{\mathrm{T}} = (1, m)^{\mathrm{T}}$ with m as slope of the invariant line.

If $\alpha = \rho_- = 0$, system (5.15) has an invariant curve given by

$$H(y) = my + \frac{b_2}{\mu_c^2 - \mu_1}y^2 + \ldots,$$

where $b_2 = \frac{\rho_+\mu_c\left(e^{2\lambda^+\pi/\omega^+} - e^{\mu^+\pi/\omega^+}\right)}{2\lambda^+ - \mu^+}$.

Figure 5.13 shows that an invariant cone is generated by $H(y)$ with parameters set as $\omega^{\pm} = 1.0$, $\lambda^+ = -\lambda^- = 1.0$, $\mu^+ = 0.02$, $\rho_+ = 12.3$, $t_{\pm} = \pi$.

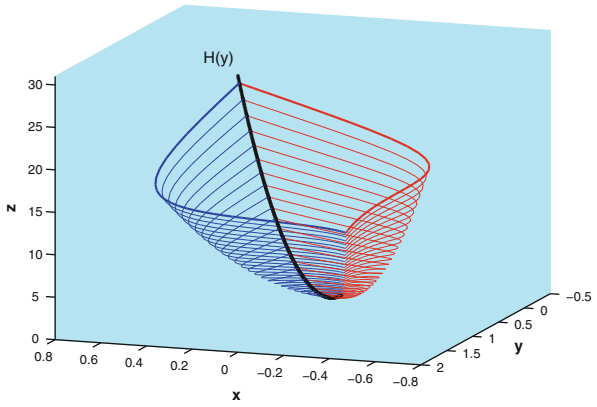

Fig. 5.13 An attractive cone generated by invariant curve $H(y)$ of PWNS

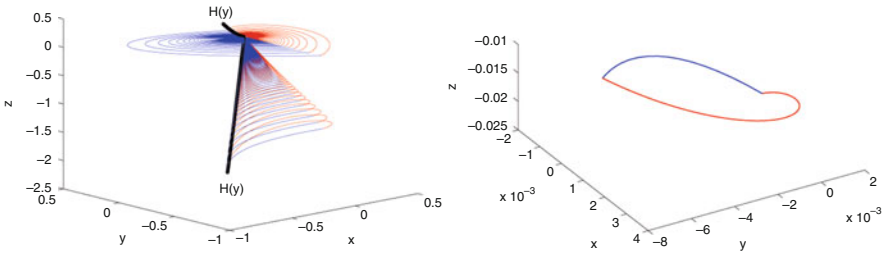

Fig. 5.14 Two generalized center manifolds of PWNS ($\rho^- = -0.01, \rho^+ = 0.1$) for $\mu^- = \mu_0^-$ (*left*), stable periodic orbit of PWNS for $\mu^- = -1.06 > \mu_0^-$ (*right*)

Figure 5.14 (left) shows another situation, where the system (5.15) has two invariant curves, hence there are two attractive invariant cones. A periodic orbit on the manifold generated by Hopf-bifurcation is shown in Fig. 5.14 (right). The simulation is done with parameters set at $\lambda^+ = -0.5, \lambda^- = 0.5, \mu^+ = 0.2, \alpha = 0.5$, $t_+ = \pi, \omega^+ = \omega^- = 1.0, \rho_- = -0.01, \rho_+ = 0.1$ and bifurcation parameter μ^- close to $\mu_0^- := -\mu^+ t_+ / t_-(\bar{\xi}) \approx -1.0604$, where $t_-(\bar{\xi}) \approx 0.5928$.

Extension of these results to situations involving sliding motion are given in [20].

References

1. Acary, V., Brogliato, B.: Numerical Methods for Nonsmooth Dynamical Systems. Applications in Mechanics and Electronics. Springer, Berlin (2008)
2. Budd, C.J., Dux, F.J.: Chattering and related behaviour in impact oscillators. Phil. Trans. R. Soc. Lond. **A347**, 365–389 (1994)
3. Carmona, V., Freire, E., Ponce, E., Torres, F.: Invariant manifolds of periodic orbits for piecewise linear three-dimensional systems. IMA J. APPl. Math. **69**, 71–91 (2004)

4. Carmona, V., Freire, E., Ponce, E., Torres, F.: Bifurcation of invariant cones in piecewise linear homogeneous systems. Int. J. Bifur. Chaos **15**(8), 2469–2484 (2005)
5. di Bernardo, M., Budd, C., Champneys, A.R., Kowalczyk, P.: Piecewise-Smooth Dynamical Systems: Theory and Applications. Applied Mathematics Series, vol. 163. Springer, Berlin (2008)
6. di Bernardo, M., Budd, C., Champneys, A.R., Kowalczyk, P., Nordmark, A.B., Olivar, G., Piiroinen, P.T.: Bifurcations in nonsmooth dynamical systems. SIAM Rev. **50**(4), 629–701 (2008)
7. Dieci, L., Lopez, L.: Sliding motion in Filippov differential systems: theoretical results and a computational approach. SIAM J. Numer. Anal. **47**, 2023–2051 (2009)
8. Filippov, A.F.: Differential equations with discontinuous right-hand side. Am. Math. Soc. Trans. **2**(42), 199–231 (1964)
9. Filippov, A.F.: Differential equations with discontinuous right-hand sides. In: Mathematics and Its Applications. Kluwer Academic, Dordrecht (1988)
10. Hosham, H.A.: Cone-like invariant manifolds for nonsmooth systems. Ph.D. Thesis. Universität zu Köln (2011)
11. Kunze, M., Küpper, T.: Non-smooth dynamical systems: an overview. In: Fiedler, B. (ed.) Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems. Bericht über Projekt im DFG-Schwerpunktprogramm 1999, pp. 431–452. Springer, Berlin (2001)
12. Köker, S.: Zur Dynamik des Glockenläutens. Diplomarbeit, Universitt zu Köln (2009)
13. Küpper T.: Invariant cones for non-smooth systems. Math. Comput. Simul. **79**, 1396–1409 (2008)
14. Küpper, T., Hosham, H.A.: Reduction to invariant cones for non-smooth systems. Math. Comput. Simul. **81**, 980–995 (2011)
15. Küpper, T., Hosham, H.A., Dudtschenko, K.: The dynamics of bells as impacting system. J. Mech. Eng. Sci. **225**(10), 2436–2443 (2011)
16. Marsden, J.E., Scheurle, J.: The construction and smoothness of invariant manifolds by the deformation method. SIAM J. Math. Anal. **18**(5), 1261–1274 (1987)
17. Veltmann, W.: Ueber die Bewegung einer Glocke. Dinglers Polytechnisches J. **22**, 481–494 (1876)
18. Weiss, D.: Existence and stability of invariant cones (2013, in preparation)
19. Weiss, D., Küpper, T., Hosham, H.A.: Invariant manifolds for nonsmooth systems. Physica D **241**(22), 1895–1902 (2012)
20. Weiss, D., Küpper, T., Hosham, H.A.: Invariant manifolds for nonsmooth systems with sliding mode (2013, submitted)
21. Wu, C.W., Chua, L.O.: On the generality of the unfolded chua's circuit. Int. J. Bifur. Chaos **6**(5), 801–832 (1996)
22. Zou, Y., Küpper, T.: Generalized Hopf bifurcation emanated from a corner. Nonlinear Anal. TAM **62**(1), 1–17 (2005)
23. Zou, Y., Küpper, T.: Generalized Hopf bifurcation for nonsmooth planar dynamical systems: the corner case. Northeast. Math. J. **17**(4), 383–386 (2001)
24. Zou, Y., Küpper, T.: Generalized Hopf bifurcation emanated from a corner for piecewise smooth planar systems. Nonlinear Anal. **62**, 1–17 (2005)
25. Zou, Y., Küpper, T., Beyn, W.J.: Generalized Hopf bifurcation for planar Filippov systems continuous at the origin. J. Nonlinear Sci. **16**(2), 159–177 (2006)

Chapter 6
Homoclinic Flip Bifurcations in Conservative Reversible Systems

Björn Sandstede

Abstract In this paper, flip bifurcations of homoclinic orbits in conservative reversible systems are analyzed. In such systems, orbit-flip and inclination-flip bifurcations occur simultaneously. It is shown that multi-pulses either do not bifurcate at all at flip bifurcation points or else bifurcate simultaneously to both sides of the bifurcation point. An application to a fifth-order model of water waves is given to illustrate the results, and open problems regarding the PDE stability of multi-pulses are outlined.

6.1 Introduction

In this paper, we discuss flip bifurcations of homoclinic orbits in conservative reversible systems. Our main motivation for studying these bifurcations comes from the observation that spatially localized traveling waves of partial differential equations (PDEs) in one-dimensional extended domains can be found as homoclinic orbits of the underlying ordinary differential equation (ODE) that describes traveling waves. To illustrate this principle, consider the fifth-order PDE

$$u_t + \frac{2}{15}u_{xxxxx} - bu_{xxx} + 3uu_x + 2u_xu_{xx} + uu_{xxx} = 0, \qquad x \in \mathbb{R}, \tag{6.1}$$

which arises as the weakly nonlinear long-wave approximation to the classical gravity-capillary water-wave problem [1, 2]. Here, $u(x,t)$ is the surface elevation measured with respect to the underlying normal water height, and the parameter b

B. Sandstede (✉)
Division of Applied Mathematics, Brown University, Providence, RI 02912, USA
e-mail: bjorn_sandstede@brown.edu

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_6,

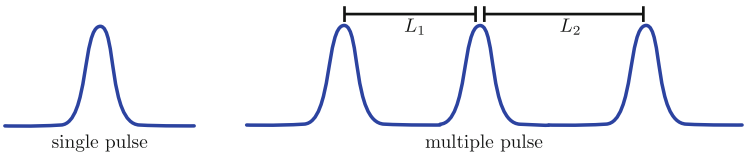

Fig. 6.1 The *left panel* illustrates the shape of a typical localized pulse $u(x)$: plotted is $u(x)$ vertically against the horizontal spatial variable x. Multiple pulses resemble several well-separated copies of a single pulse as indicated in the *right panel* for a 3-pulse, consisting of three copies. The distances L_1 and L_2 between consecutive pulses in the 3-pulse can be used to distinguish different multi-pulses

is the offset of the Bond number, which measures surface tension, from the value $\frac{1}{3}$. Traveling waves $u(x,t) = u(x+ct)$ of (6.1) satisfy the fourth-order equation

$$\frac{2}{15}u^{\mathrm{iv}} - bu'' + cu + \frac{3}{2}u^2 - \frac{1}{2}[u']^2 + [uu']' = 0, \tag{6.2}$$

where c denotes the wave speed. Localized wave profiles $u(x)$ of (6.1) that satisfy $\lim_{x\to\pm\infty} u(x) = 0$, which we will refer to as pulses, correspond therefore to homoclinic orbits of the first-order system obtained from the ODE (6.2).

Assume now that we found a pulse $u(x)$, which corresponds to a localized wave of elevation or suppression. In this case, it is of interest to see whether several copies of the pulse can be glued together to create a traveling pulse that consists of several regions of elevation or suppression as indicated in Fig. 6.1. Several bifurcation scenarios are known at which multi-pulses of the form described above emerge, and we refer to [5] for a comprehensive survey. This paper focuses on homoclinic flip bifurcations, which come in two varieties. Orbit-flip bifurcations arise if the pulse is more localized than expected from the spatial eigenvalue structure of the equilibrium $u = 0$ of the ODE (6.2). Inclination-flip bifurcations, on the other hand, arise as follows: let $\mathscr{L}$ be the linearization of the PDE (6.1), formulated in a comoving frame, about the pulse, then this operator has an eigenvalue at the origin due to translation symmetry. Let $\psi(x)$ denote the associated eigenfunction of the adjoint operator $\mathscr{L}^*$; this eigenfunction has a natural interpretation as a solution of the adjoint variational equation of the ODE (6.2) about the pulse $u(x)$. An inclination-flip arises if the adjoint eigenfunction is more localized than expected.

Whether, and in what form, multi-pulses bifurcate at an orbit- or inclination flip bifurcation depends strongly on whether the underlying traveling-wave system (6.2) has additional structure. Here, we focus on two possible structures that commonly arise. The first structure is equivariance of (6.2) under the reflection $x \mapsto -x$, which we will refer to as reversibility. The second relevant structure is whether the traveling-wave system admits a first integral, that is, a real-valued quantity H that does not change when evaluated along solutions: we refer to such systems as conservative.

It was shown in [12] that orbit-flip bifurcations of nonconservative reversible systems lead to N-pulses for each N. A similar result was shown in [13] for nonreversible conservative systems. It turns out that the water-wave problem (6.2) is both reversible and conservative (we will show this in Sect. 6.4). Neither of the aforementioned results therefore applies to (6.2), and this paper focuses on deriving bifurcation results for this case. As we will see, the results for reversible conservative systems are quite different from those for systems that admit one but not both of these structures.

The main open issue is the stability of the multi-pulses found in this and other bifurcation scenarios with respect to the underlying partial differential equation. The fifth-order model given above is a Hamiltonian PDE, and stability for such equations is subtle. We will comment in detail on the outstanding issues in the conclusions section at the end of this paper.

This paper is structured as follows. The precise setting and the main results are formulated in Sect. 6.2. Our results are proved in Sect. 6.3, and we consider the application to the water-wave problem in Sect. 6.4. Conclusions and open problems are presented in Sect. 6.5.

6.2 Main Results

In this section, we state the setting, assumptions, and main results more formally. We consider the ordinary differential equation

$$u' = f(u, \mu), \qquad (u, \mu) \in \mathbb{R}^{2n} \times \mathbb{R}, \tag{6.3}$$

where f is a smooth nonlinearity. We assume that (6.3) is reversible and conservative in the following sense.

Hypothesis (H1) (Reversibility). *There exists a linear map $R : \mathbb{R}^{2n} \to \mathbb{R}^{2n}$ such that $R^2 = \mathrm{id}$, the fixed-point space $\mathrm{Fix}(R)$ of the reverser R satisfies $\dim \mathrm{Fix}(R) = n$, and $Rf(u, \mu) = -f(Ru, \mu)$ for all $(u, \mu) \in \mathbb{R}^{2n} \times \mathbb{R}$.*

We call a solution $u(x)$ reversible or symmetric with respect to the reverser R if $u(0) \in \mathrm{Fix}(R)$. Any symmetric solution u automatically satisfies

$$u(-x) = Ru(x), \qquad x \in \mathbb{R}.$$

We also assume that (6.3) is conservative, that is, it admits a conserved quantity or first integral that is compatible with the reverser R.

Hypothesis (H2) (Conservative System). *There exists a smooth function $H : \mathbb{R}^{2n} \times \mathbb{R} \to \mathbb{R}$ such that $H_u(u, \mu) f(u, \mu) = 0$ for all $(u, \mu) \in \mathbb{R}^{2n} \times \mathbb{R}$, and $H_u(u, \mu) = 0$ only at a discrete set of points in $\mathbb{R}^{2n}$ for each fixed $\mu \in \mathbb{R}$. Furthermore, we assume that H is invariant under the reverser R, that is, $H(Ru, \mu) = H(u, \mu)$ for all (u, μ).*

If Hypothesis (H2) is met, then $H(u(x),\mu) = H(u(0),\mu)$ along any solution $u(x)$ of (6.3). Hamiltonian systems given by

$$u' = JH_u(u,\mu), \qquad J = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \qquad u \in \mathbb{R}^n \times \mathbb{R}^n$$

are a particular example of conservative systems.

Throughout, we assume that zero is a hyperbolic equilibrium of (6.3) for all μ near zero so that $f(0,\mu) = 0$ and $f_u(0,\mu)$ is hyperbolic for μ near zero. We also assume that (6.3) admits a reversible homoclinic solution $h_0(x)$ to the origin for $\mu = 0$.

Hypothesis (H3). *There is a solution $h_0(x)$ of (6.3) for $\mu = 0$ with $h_0 \not\equiv 0$ such that*

1. $\lim_{x\to\pm\infty} h_0(x) = 0$;
2. $T_{h_0(0)}W^{\mathrm{s}}(0) \cap T_{h_0(0)}W^{\mathrm{u}}(0) = \mathbb{R}h_0'(0)$;
3. $h_0(0) \in \mathrm{Fix}(R)$.

Hypotheses (H1) and (H2) each imply that the spectrum of $f_u(0,\mu)$ is symmetric with respect to the imaginary axis; see, for instance, [5, 14]. We assume that the leading eigenvalues of the origin are real.

Hypothesis (H4). *The spectrum of the equilibrium $u = 0$ is given by*

$$\mathrm{spec}(f_u(0,\mu)) = \sigma^{\mathrm{s}} \cup \{\pm\lambda^{\mathrm{u}}(\mu), \pm\lambda^{\mathrm{uu}}(\mu)\} \cup \sigma^{\mathrm{u}}$$

where $\pm\lambda^{\mathrm{u}}(\mu)$ and $\pm\lambda^{\mathrm{uu}}(\mu)$ are simple eigenvalues with $0 < \lambda^{\mathrm{u}}(\mu) < \lambda^{\mathrm{uu}}(\mu)$. We also assume that there is a constant λ^{r} with $\lambda^{\mathrm{r}} > \lambda^{\mathrm{uu}}(\mu)$ so that $\mathrm{Re}\,\sigma^{\mathrm{s}} < -\lambda^{\mathrm{r}}$ and $\mathrm{Re}\,\sigma^{\mathrm{u}} > \lambda^{\mathrm{r}}$ for all μ near zero.

Hypothesis (H3) implies that there exists a smooth one-parameter family $h_\mu(x)$ of homoclinic solutions for μ close to zero that all satisfy Hypothesis (H3); see again [5, 14], for instance. We assume that these homoclinic orbits undergo an orbit-flip bifurcation at $\mu = 0$:

Hypothesis (H5). *We assume that $h_0(x) \in W^{\mathrm{uu}}(0)$ for $\mu = 0$ and that*

1. $\lim\limits_{x\to-\infty} \mathrm{e}^{-\lambda^{\mathrm{uu}}x} h_0(x) = v_{\mathrm{uu}} \neq 0$;
2. $\lim\limits_{x\to-\infty} \mathrm{e}^{-\lambda^{\mathrm{u}}x} \frac{\mathrm{d}}{\mathrm{d}\mu} h_\mu(x)|_{\mu=0} = v_{\mathrm{u}} \neq 0$.

It follows that v_{u} and v_{uu} are eigenvectors of $f_u(0,0)$ that belong to the eigenvalues $\lambda^{\mathrm{u}}(0)$ and $\lambda^{\mathrm{uu}}(0)$, respectively. The quantities

$$b^j := \langle H_{uu}(0,0)v_j, Rv_j\rangle = H_{uu}(0,0)[v_j, Rv_j], \qquad j = \mathrm{u}, \mathrm{uu} \tag{6.4}$$

will play an important role in our result. In four dimensions, the sign of the product $b^{\mathrm{u}}b^{\mathrm{uu}}$ has the following geometric interpretation. First, using that H cannot change along the stable or unstable manifolds of the origin, we can show that

$$b^j = \langle H_{uu}(0,0)v_j, Rv_j\rangle = \langle H_{uu}(0,0)(v_j + Rv_j), (v_j + Rv_j)\rangle, \qquad j = \mathrm{u}, \mathrm{uu}.$$

Thus, b^{u} measures how the energy changes if we move along the direction $v_{\mathrm{u}} + Rv_{\mathrm{u}}$, which is the spine of the cone formed by the eigenvectors v_{u} and Rv_{u} belonging to the eigenvalues λ^{u} and $-\lambda^{\mathrm{u}}$, respectively, of $f_u(0,0)$. In particular, $b^{\mathrm{u}} > 0$ indicates that the energy increases in this direction, while $b^{\mathrm{u}} < 0$ means the energy decreases. The quantity b^{uu} has same interpretation for the cone in the strong stable and unstable eigenspace. The product $b^{\mathrm{u}}b^{\mathrm{uu}}$ is therefore positive if the energy increases or decreases in both cones, while $b^{\mathrm{u}}b^{\mathrm{uu}} < 0$ means that the energy increases in one cone and decreases in the other cone. In order to be able to pass from $h_\mu(L)$ to $h_\mu(-L)$, we need to pass through the product of these cones near the equilibrium. Since energy is conserved, this seems possible only if the zero energy level set intersects these cones: this happens only if $b^{\mathrm{u}}b^{\mathrm{uu}} < 0$. Thus, at least in four dimensions, we expect N-pulses to exist only when $b^{\mathrm{u}}b^{\mathrm{uu}}$ is negative, but not if $b^{\mathrm{u}}b^{\mathrm{uu}}$ is positive. Our main result confirms this intuition, not just for the four-dimensional but also for the general case.

Theorem 6.1. *Assume that Hypotheses (H1)–(H5) are met. For each $N > 1$, there exist numbers $\mu_N > 0$ and $L_N \gg 1$ with the following properties.*

1. *If $b^{\mathrm{u}}b^{\mathrm{uu}} > 0$, then (6.3) with $|\mu| < \mu_N$ does not admit a homoclinic orbit that makes N distinct loops near the primary orbit $h_0(x)$, where each return time is larger than L_N.*
2. *If $b^{\mathrm{u}}b^{\mathrm{uu}} < 0$, then (6.3) has, for each $\mu \neq 0$ with $|\mu| < \mu_N$, a unique homoclinic orbit that makes N distinct loops near the primary orbit $h_0(x)$, where each return time is larger L_N. This orbit is reversible, and the return times between consecutive pulses are given, to leading order, by*

$$L = \frac{\ln|\mu|}{\lambda^{\mathrm{u}} - \lambda^{\mathrm{uu}}} + L_*$$

for some constant $L_* \in \mathbb{R}$.

In other words, N-pulses either do not emerge at all or else emerge to either side of $\mu = 0$. This is in contrast to many other homoclinic flip bifurcations, where solutions bifurcate either sub- or super-critically. In particular, N-pulses bifurcate to one side only at orbit-flip bifurcations in non-conservative reversible systems [12] and in non-reversible conservative systems [13].

6.3 Proof of Theorem 6.1

We apply Lin's method [8] to prove the existence and nonexistence of the N-pulses near a homoclinic flip bifurcation. This method is explained in detail in [12], see also [5], and we shall follow here the same strategy and use the same notation as in [12].

Before we can state the results from [8, 10, 12], we introduce additional notation. Recall that we assumed that our system (6.3) is conservative. As shown for instance in [14], this property implies that the functions

$$\psi_\mu(x) = \nabla_u H(h_\mu(x), \mu) \tag{6.5}$$

are nontrivial bounded solutions to the adjoint variational equations

$$w' = -f_u(h_\mu(x), \mu)^* w \tag{6.6}$$

associated with the homoclinic orbits $h_\mu(x)$. We can now state the results from [8, 10, 12] that express conditions for the existence of N-pulses. Fix any natural number $N > 1$, then the results in [8, 10, 12] state that there are numbers μ_* and L_* such that Eq. (6.3) with $|\mu| < \mu_*$ has an N-homoclinic orbit that is pointwise close to the orbit of $h_0(x)$ and follows $h_0(x)$ N-times if, and only if, the equations

$$\langle \psi_\mu(-L_{j-1}), h_\mu(L_{j-1})\rangle - \langle \psi_\mu(L_j), h_\mu(-L_j)\rangle + R_j(L, \mu) = 0 \tag{6.7}$$

with $j = 1, \ldots, N-1$ has a solution $L = (L_j)_{j=1,\ldots,N-1}$ with $L_j \geq L_*$ and $L_0 = \infty$. The numbers L_j are the return times of consecutive homoclinic loops to a fixed section at $h_0(0)$ or, alternatively, the distances between consecutive pulses in the corresponding multi-pulse. The functions $R_j(L, \mu)$ are higher-order terms, which we will estimate below. In restricting the index j in (6.7) to the set $j = 1, \ldots, N-1$, we have used [12, Lemma 3.2] which asserts that the Nth equation will be satisfied automatically due to the energy constraint H provided the first $N-1$ equations are met.

Before stating the estimates for the remainder terms $R_j(L, \mu)$, we simplify (6.7) further. Reversibility of $h_\mu(x)$ and compatibility of R and H imply that

$$\begin{aligned} h_\mu(x) &= Rh_\mu(-x) \\ \psi_\mu(-x) &= \nabla_u H(h_\mu(-x), \mu) = \nabla_u H(Rh_\mu(x), \mu) = R^* \nabla_u H(h_\mu(x), \mu) = R^* \psi_\mu(x). \end{aligned}$$

Thus, (6.7) can be written as

$$\langle \psi_\mu(L_{j-1}), h_\mu(-L_{j-1})\rangle - \langle \psi_\mu(L_j), h_\mu(-L_j)\rangle + R_j(L, \mu) = 0, \quad j = 1, \ldots, N-1.$$

Using that $L_0 = \infty$, we can recursively add the $(j-1)$th equation to the jth equation to obtain the new equivalent system

$$\langle \psi_\mu(L_j), h_\mu(-L_j)\rangle - R_j(L,\mu) - R_{j-1}(L,\mu) = 0, \quad j = 1,\dots,N-1$$

or, equivalently,

$$\langle \nabla_u H(h_\mu(L_j),\mu), h_\mu(-L_j)\rangle - R_j(L,\mu) - R_{j-1}(L,\mu) = 0, \quad j = 1,\dots,N-1 \tag{6.8}$$

with $R_0 \equiv 0$.

Next, we derive expressions for the scalar product that appears in (6.8). Since we assumed that $f_u(0,\mu)$ is hyperbolic, and therefore in particular invertible, it follows easily from differentiating $H_u(u,\mu)f(u,\mu)$ with respect to x that $H_u(0,\mu) = 0$. Using this property together with [12, (3.8)] and (6.5), we obtain the expansions

$$\begin{aligned} h_\mu(-x) &= \mu \mathrm{e}^{-\lambda^{\mathrm{u}} x} v_{\mathrm{u}} + \mathrm{e}^{-\lambda^{\mathrm{uu}} x} v_{\mathrm{uu}} + v_{\mathrm{r}}(x) \\ &\quad + \mathrm{O}\left(|\mu|(|\mu|\mathrm{e}^{-\lambda^{\mathrm{u}} x} + \mathrm{e}^{-2\lambda^{\mathrm{u}} x} + \mathrm{e}^{-\lambda^{\mathrm{uu}} x}) + \mathrm{e}^{-2\lambda^{\mathrm{uu}} x} \right) \\ \nabla_u H(h_\mu(x),\mu) &= H_{uu}(0,\mu) h_\mu(x) + \mathrm{O}(|h_\mu(x)|^2) \\ &= H_{uu}(0,\mu) R h_\mu(-x) + \mathrm{O}(|h_\mu(-x)|^2) \end{aligned} \tag{6.9}$$

in the limit $x \to \infty$, where the function $v_{\mathrm{r}}(x)$ lies in the eigenspace associated with σ^{u} and decays faster than $\mathrm{e}^{-\lambda^{\mathrm{r}} x}$ as $x \to \infty$. Using these expansions, recalling the definition (6.4) of the quantities b^j, and using the fact that $H_{uu}(0,0)v$ is an eigenvector of $f_u(0,0)^*$ belonging to the eigenvalue $-\lambda$ whenever v is an eigenvector of $f_u(0,0)$ belonging to the eigenvalue λ, a straightforward calculation shows that

$$\begin{aligned} \langle \nabla_u H(h_\mu(x),\mu), h_\mu(-x)\rangle &= \mu^2 b^{\mathrm{u}} \mathrm{e}^{-2\lambda^{\mathrm{u}} x} + b^{\mathrm{uu}} \mathrm{e}^{-2\lambda^{\mathrm{uu}} x} \\ &\quad + \mathrm{O}\Big(\mathrm{e}^{-2\lambda^{\mathrm{r}} x} + |\mu|[\mathrm{e}^{-(2\lambda^{\mathrm{u}} + \lambda^{\mathrm{uu}})x} + \mathrm{e}^{-2\lambda^{\mathrm{uu}} x}] \\ &\quad + |\mu|^2[\mathrm{e}^{-(\lambda^{\mathrm{u}} + \lambda^{\mathrm{uu}})x} + \mathrm{e}^{-3\lambda^{\mathrm{u}} x}] + |\mu|^3 \mathrm{e}^{-2\lambda^{\mathrm{u}} x} \Big) \end{aligned} \tag{6.10}$$

uniformly in μ near zero and $x \gg 1$. It remains to derive estimates on the remainder terms R_j.

Lemma 6.1. *Under the hypotheses of Theorem 6.1, the error terms $R_j(L,\mu)$ satisfy*

$$R_j(L,\mu) = \mathrm{O}\left(\left(\mathrm{e}^{-\lambda^{\mathrm{u}} L_{j-1}} + \mathrm{e}^{-\lambda^{\mathrm{u}} L_j} \right) \sum_{k=1}^{N-1} \left(\mu^2 \mathrm{e}^{-2\lambda^{\mathrm{u}} L_k} + \mathrm{e}^{-2\lambda^{\mathrm{uu}} L_k} \right) \right), \tag{6.11}$$

and the error terms can be differentiated.

Proof. The estimates given in [12, Theorem 3] are not sufficient to get the statement of the theorem. We therefore return to [10, §3.3.2] where the relevant expression for the bifurcation equations are recorded on [10, top of page 99]. The integral terms appearing in [10, top of page 99] can be estimated by

$$\left(\mathrm{e}^{-\lambda^{\mathrm{u}}L_{j-1}}+\mathrm{e}^{-\lambda^{\mathrm{u}}L_j}\right)\sum_{k=1}^{N-1}\left(\mu^2\mathrm{e}^{-2\lambda^{\mathrm{u}}L_k}+\mathrm{e}^{-2\lambda^{\mathrm{uu}}L_k}\right)$$

using [10, Lemma 3.20]. The scalar products appearing in [10, top of page 99] are given by

$$\langle\psi_\mu(-L_j),h_\mu(L_j)\rangle+\mathrm{O}\left(\left(\mathrm{e}^{-\lambda^{\mathrm{u}}L_{j-1}}+\mathrm{e}^{-\lambda^{\mathrm{u}}L_j}\right)\sum_{k=1}^{N-1}\left(\mu^2\mathrm{e}^{-2\lambda^{\mathrm{u}}L_k}+\mathrm{e}^{-2\lambda^{\mathrm{uu}}L_k}\right)\right)$$

once [10, (3.42), (3.43) and (3.39)] are used, which completes the proof. □

As in [10, 12], we replace the variables L_j and μ by the new variables

$$\mu=\pm r^{\beta/2},\qquad a_{jr}=\mathrm{e}^{-2\lambda^{\mathrm{u}}L_j} \tag{6.12}$$

for $a_j>0$ and $r\geq 0$, where the exponent $\beta>0$ will be chosen later. We also define

$$\alpha=\frac{\lambda^{\mathrm{uu}}}{\lambda^{\mathrm{u}}}-1=\frac{\lambda^{\mathrm{uu}}-\lambda^{\mathrm{u}}}{\lambda^{\mathrm{u}}}.$$

Substituting the expansion (6.10) and the estimates (6.11) into the bifurcation equations (6.8) and rewrite them using (6.12), we arrive, after some tedious but straightforward manipulations, at the system

$$a_jr^{1+\beta}b^{\mathrm{u}}+a_j^{1+\alpha}r^{1+\alpha}b^{\mathrm{uu}}+\mathrm{O}\left(\max\left\{r^{1+\alpha+\gamma},r^{1+\beta+\gamma}\right\}\right)=0 \tag{6.13}$$

where $j=1,\ldots,N-1$. Here, $\gamma>0$ is a positive constant that depends only on the quantities λ^{u}, λ^{uu}, and λ^{r} but not on L or μ. Observe that nontrivial solutions of (6.13) can exist only in the scaling $\alpha=\beta$ for which (6.13) becomes

$$a_jr^{1+\alpha}b^{\mathrm{u}}+a_j^{1+\alpha}r^{1+\alpha}b^{\mathrm{uu}}+\mathrm{O}(r^{1+\alpha+\gamma})=0,\qquad j=1,\ldots,N-1.$$

Dividing by $r^{1+\alpha}$, we obtain

$$a_j\left(b^{\mathrm{u}}+a_j^{\alpha}b^{\mathrm{uu}}\right)+\mathrm{O}(r^{\gamma})=0,\qquad j=1,\ldots,N-1. \tag{6.14}$$

If $b^{\mathrm{u}}b^{\mathrm{uu}}>0$, it is not difficult to see that (6.14) cannot have any solutions other than $a_j=0$ for all j which corresponds to the persisting homoclinic orbit $h_\mu(x)$.

Thus, let us assume from now on that $b^{\mathrm{u}}b^{\mathrm{uu}} < 0$. In this case, (6.14) has the positive solution

$$a_j^* = \left(-\frac{b^{\mathrm{u}}}{b^{\mathrm{uu}}}\right)^{1/\alpha} > 0, \qquad j = 1, \ldots, N-1 \tag{6.15}$$

at $r = 0$, and we can solve (6.14) near this solution for $r > 0$ by the implicit function theorem.

This completes the existence part of Theorem 6.1. The obtained N-homoclinic orbit is reversible since we could simply have solved the equations for $j = 1, \ldots, [N/2]$, with $[x]$ being the largest integer smaller than x, and setting $a_{N-j} := a_j$ for $j = 1, \ldots, [N/2]$. Applying [12, Lemma 3.1] then shows that any solution to this truncated system corresponds to a reversible N-homoclinic orbit of (6.3). Proceeding as above though, we find the same solution (6.15), which therefore must be symmetric. Lastly, the uniqueness of the N-homoclinic orbits can be proved as in [12, Lemma 3.6].

6.4 Application to a Fifth-Order Model for Water Waves

We now apply Theorem 6.1 to Eq. (6.1), now written as

$$u_t + \partial_x \left(\frac{2}{15}u_{xxxx} - bu_{xx} + \frac{3}{2}u^2 + \frac{1}{2}u_x^2 + uu_{xx}\right) = 0, \qquad x \in \mathbb{R}, \tag{6.16}$$

which, as mentioned in the introduction, arises as a long-wave approximation to the gravity-capillary water-wave problem [2]. Localized traveling waves $u(x,t) = u(x+ct)$ of (6.16) satisfy the fourth-order equation

$$\frac{2}{15}u^{\mathrm{iv}} - bu'' + cu + \frac{3}{2}u^2 + \frac{1}{2}[u']^2 + uu'' = 0. \tag{6.17}$$

Note that (6.17) is reversible under the reflection $x \mapsto -x$. This equation is also Hamiltonian, and hence conservative: indeed, as shown in [2], the variables

$$q_1 = u, \quad q_2 = u', \quad p_1 = -\frac{2}{15}u''' + bu' - uu', \quad p_2 = \frac{2}{15}u''$$

make (6.17) Hamiltonian with respect to the energy

$$H = -\frac{1}{2}q_1^3 - \frac{c}{2}q_1^2 + p_1q_2 - \frac{b}{2}q_2^2 + \frac{15}{4}p_2^2 + \frac{1}{2}q_1q_2^2$$

and the symplectic operator J

$$J: \quad (q_1, q_2, p_1, p_2) \longmapsto (p_1, p_2, -q_1, -q_2).$$

In these coordinates, the reverser R becomes

$$R: \quad (q_1, q_2, p_1, p_2) \longmapsto (q_1, -q_2, -p_1, p_2),$$

and we have $H \circ R = H$ as required. In particular, Hypotheses (H1) and (H2) are met.

As shown in [2, (4.3)], Eq. (6.17) has the explicit localized solution

$$u_*(x) = 3\left(b + \frac{1}{2}\right) \operatorname{sech}^2\left(\sqrt{3(2b+1)}\,\frac{x}{2}\right) \tag{6.18}$$

for $b > -\frac{1}{2}$, with wave speed given by

$$c = c_*(b) = \frac{3}{5}(2b+1)(b-2). \tag{6.19}$$

Linearizing (6.17) about $u = 0$, we find that its eigenvalues satisfy the relation

$$\frac{2}{15}\lambda^4 - b\lambda^2 + c = 0. \tag{6.20}$$

Substituting $c = c_*(b)$ from (6.19) and computing the unstable spatial eigenvalues, we find that they are given by

$$\lambda^{\mathrm{u}} = \frac{1}{2}\sqrt{6(b-2)}, \qquad \lambda^{\mathrm{uu}} = \sqrt{3(2b+1)}. \tag{6.21}$$

In particular, the equilibrium $u = 0$ is hyperbolic when $b > 2$, and this will be the region we shall focus on from now on. We are interested in applying Theorem 6.1 to the system (6.17), where the speed c, varied near $c = c_*(b)$, plays the role of the parameter μ that appears in Sect. 6.2. We now discuss the validity of the hypotheses required for Theorem 6.1 to hold.

Comparing with (6.17), we see that the homoclinic orbit u_* is indeed in an orbit-flip configuration for all such b, and we see that Hypotheses (H3)(i)+(iii), (H4), and (H5)(i) are satisfied.

Thus, it remains to discuss Hypotheses (H3)(ii) and (H5)(ii): we do not have an analytical proof of their validity but describe now how they can be checked numerically. Restated in a more convenient formulation, Hypothesis (H3)(ii) assumes that $u_*'(x)$ is the only bounded solution of the variational equation

$$\mathscr{L}v := \frac{2}{15}v^{\mathrm{iv}} - bv'' + c_*(b)v + 3u_*(x)v + u_*'(x)v' + u_*(x)v'' + vu_*''(x) = 0. \tag{6.22}$$

In other words, this hypothesis requires that the eigenvalue $\lambda_{\mathrm{pde}} = 0$ of the operator $\mathscr{L}$ posed on $L^2(\mathbb{R})$ is simple. Discretizing the derivatives in the operator $\mathscr{L}$ by

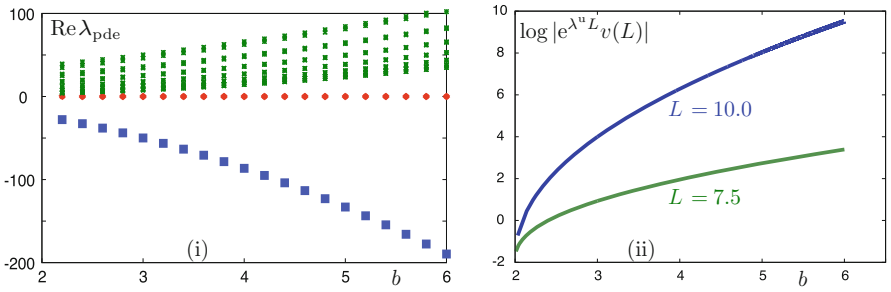

Fig. 6.2 The numerical computations presented in panel (*i*) indicate that $\lambda_{\rm pde} = 0$ is simple as an eigenvalue of the linearization $\mathscr{L}$ about the pulse; this indicates that Hypothesis (H3)(ii) is met. In panel (*ii*), we plot b versus $e^{\lambda^u L} \frac{d}{dc} u_*(x;c)|_{c=c_*(b)}$ for different values of L: this quantity is not decreasing as L increases, thus indicating that Hypothesis (H5)(ii) is also satisfied. We refer to the main text for details on how these computations were carried out

centered finite differences and calculating its spectrum numerically in MATLAB using the sparse eigenvalue solver EIGS, we find that $\lambda_{\rm pde} = 0$ is indeed simple as an eigenvalue of $\mathscr{L}$; see Fig. 6.2a for the spectrum of $\mathscr{L}$ for values of b in the range $[2, 6]$. These results therefore indicate that Hypothesis (H3)(ii) is met, so that there is a family $u_*(x;c)$ of pulses for c near $c_*(b)$ for each fixed $b > 2$.

Next, Hypothesis (H5)(ii) assumes that the function

$$v(x) := \frac{d}{dc} u_*(x;c)\Big|_{c=c_*(b)}$$

does not decay faster than exponentially with rate λ^u as $x \to \infty$; in other words, $e^{\lambda^u x} v(x)$ should not converge to zero. Differentiating (6.17) with respect to c and evaluating at $c = c_*(b)$, we see that the function $v(x)$ satisfies the system

$$\mathscr{L} v + u_*(x) = 0.$$

We calculated this solution in AUTO on the interval $[0, L]$ for with Neumann boundary conditions on either end for different values of L and plotted $e^{\lambda^u L} v(L)$ for some of the values of L in Fig. 6.2b. The results indicate that Hypothesis (H5)(ii) is also met.

In summary, a combination of analytical verification and numerical computations indicates that Theorem 6.1 applies to the fifth-order water-wave problem (6.17). It remains to evaluate the constants

$$b^j := \langle H_{uu}(0,0) v_j, R v_j \rangle, \qquad j = \mathrm{u}, \mathrm{uu}$$

from (6.4), where we can take $v_{\rm u}$ and $v_{\rm uu}$ to be any eigenvectors of the linearization

$$JH_{uu}(0) = \begin{pmatrix} 0 & 0\,1\,0 \\ 0 & 0\,0\,1 \\ -1 & 0\,0\,0 \\ 0 & -1\,0\,0 \end{pmatrix} \begin{pmatrix} -c & 0 & 0 & 0 \\ 0 & -b & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{15}{2} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \frac{15}{2} \\ c & 0 & 0 & 0 \\ 0 & b & -1 & 0 \end{pmatrix}$$

about $u = 0$ at $c = c_*(b)$. The eigenvector v belonging to a real eigenvalue λ is given by

$$v = \left(1, \lambda, \frac{c}{\lambda}, \frac{2\lambda^2}{15}\right).$$

Thus, upon using (6.20), we obtain

$$\langle H_{uu}(0)v, Rv\rangle = b\lambda^2 - 2c,$$

and substituting the eigenvalues from (6.21), we obtain

$$b^{\mathrm{u}} = -\frac{3}{10}(3b+4)(b-2), \qquad b^{\mathrm{uu}} = \frac{3}{5}(3b+4)(2b+1).$$

Thus, b^{uu} is positive for $b > -\frac{1}{2}$, while b^{u} is positive for $-\frac{1}{2} < b < 2$ and negative for $b > 2$. In summary, we expect N-pulses to bifurcate from the primary pulse for $b > 2$, while our theory does not apply to $-\frac{1}{2} < b < 2$ as the origin is not hyperbolic in this parameter range.

The analytical predictions from Theorem 6.1 (communicated by me to the authors of [2] prior to publication of [2]) were confirmed in the numerical computations of 2-pulses presented in [2, Figures 23–24] below and above the bifurcation point $c = c_*(b)$.

6.5 Open Problems

One of the issues not addressed here, or elsewhere, is the stability of the multi-pulses we found above under the time evolution of the fifth-order model (6.16)

$$u_t + \partial_x\left(\frac{2}{15}u_{xxxx} - bu_{xx} + cu + \frac{3}{2}u^2 + \frac{1}{2}u_x^2 + uu_{xx}\right) = 0, \qquad x \in \mathbb{R}, \tag{6.23}$$

now written in a comoving frame. This is a difficult question as the PDE (6.23) is Hamiltonian when posed on appropriate function spaces since it can be written as

$$u_t = -\partial_x E'(u) \tag{6.24}$$

where $J := -\partial_x$ is skew-symmetric and

$$E(u) = \frac{1}{2}\int_{\mathbb{R}} \left(\frac{2}{15}u_{xx}^2 + bu_x^2 + cu^2 + u^3 - uu_x^2 \right) \mathrm{d}x.$$

It is worthwhile to point out that the L^2-norm

$$N(u) = \frac{1}{2}\int_{\mathbb{R}} u(x)^2 \, \mathrm{d}x,$$

is invariant under the time evolution of (6.23). More generally, many other fifth-order equations, such as the fifth-order Korteweg–de Vries equation

$$u_t + \partial_x \left(u_{xxxx} - u_{xx} + cu + u^2 \right), \qquad x \in \mathbb{R}, \tag{6.25}$$

which are of the form (6.24) with the same conserved quantity $N(u)$, are known to exhibit solitary waves and multi-pulses: it is therefore natural, and indeed important for applications, to investigate the temporal stability of their pulses and multi-pulses.

Before we discuss multi-pulses, we briefly review stability results for the underlying primary solitary wave $u_*(x)$ given in (6.18) of (6.23); recall that this profile is in a flip configuration and gives rise to multi-pulses as discussed in the previous section. Given the Hamiltonian nature of (6.23), it is natural to construct stable stationary solutions of (6.23), which correspond to traveling waves with speed c of the original equation (6.1), by seeking minimizers of the Hamiltonian E. However, solitary waves usually do not minimize the functional E. Instead, they can be thought of as constrained minimizers of $\tilde{E}(u) = E(u) - cN(u)$ under the constraint $N(u) = \mathrm{const}$; in this formulation, the wave speed c arises as a Lagrange multiplier. Typically, the energy will decrease as one moves along the family $u_*(\cdot; c)$ of solitary waves with c varying, while b is kept fixed. As shown in [3, Theorem 2.1] and the references therein, the pulse $u_*(x)$ will be stable for (6.23) if the Hessian, or second variation, $E''(u_*) = \mathscr{L}$ of E has a simple eigenvalue at the origin and only one negative eigenvalue and if furthermore $\frac{\mathrm{d}}{\mathrm{d}c}N(u_*(\cdot; c)) > 0$. For (6.23), the first hypothesis is checked numerically in Fig. 6.2(i), while Fig. 6.3 indicates the second assumption is met as well. These numerical calculations therefore indicate that the underlying primary pulses (6.18) are indeed stable, and it is natural to discuss whether the multi-pulses emerging from them are stable, too.

First, we consider the Hessian $\mathscr{L}_N$ of the PDE energy E at an N-pulse solution. It follows from [11] that $\mathscr{L}_N$ has N eigenvalues near the negative eigenvalue of $\mathscr{L}_1$ and exactly N small eigenvalues near the origin. For (6.23), preliminary computations that follow [11] indicate that $N-1$ of these small eigenvalues are negative, while the remaining small eigenvalue is at the origin, as dictated by translation invariance of the energy; see Fig. 6.4. In particular, there are now $2N-1$

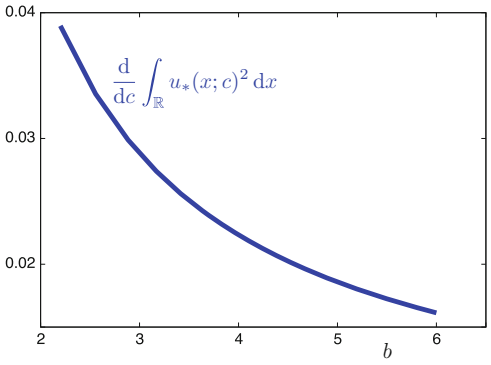

Fig. 6.3 Shown is the dependence of $\frac{\mathrm{d}}{\mathrm{d}c}\int_{\mathbb{R}} u_*(x;c)^2\,\mathrm{d}x$ on the parameter b. This quantity is computed numerically using AUTO

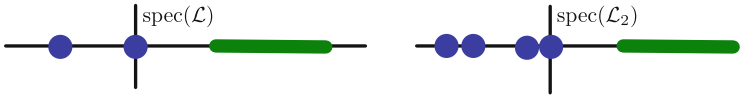

Fig. 6.4 Shown are the anticipated spectra of the Hessian of the energy $\mathscr{E}$ evaluated at the primary pulse (*left*) and the 2-pulse (*right*)

direction along with the energy decreases, and the single known conserved quantity $N(u)$ cannot compensate for them when $N > 1$.

Next, consider the linearization $J\mathscr{L}_N = -\partial_x \mathscr{L}_N$ of the PDE (6.24) about an N-pulse. The anticipated spectra of $J\mathscr{L}$ and $J\mathscr{L}_N$ are shown in Fig. 6.5 for $N = 2$. Indeed, the results in [11], applied in an appropriate exponentially weighted norm to the linearization $J\mathscr{L}_N$, show that $J\mathscr{L}$ will have $2N$ eigenvalues near the origin, and two of these will reside at the origin due to translational symmetry. For a 2-pulse, the remaining two eigenvalues may reside on the real or imaginary axis, or move off the imaginary axis. In the latter case, there will be another pair of eigenvalues on the other side of the imaginary axis as the spectrum of Hamiltonian linearizations is symmetric with respect to reflections across the imaginary axis. The reason that the two extra eigenvalues are not included in the count of $2N$ is that their eigenfunctions are not bounded under the exponential weighted norm used to locate them. We claim that the case shown in Fig. 6.5(iii) can actually not occur given that the spectrum of $\mathscr{L}_2$ looks as shown in Fig. 6.4. Indeed, there should be three directions along which the energy decreases near the 2-pulse: one of these directions corresponds again to changing the speed of the 2-pulse, and the other two directions must therefore be associated with the eigenspace of the two real nonzero eigenvalues; however, the dynamics on this eigenspace is of saddle-type, and the energy therefore decreases only along one and not both of these directions. More generally, we expect that an expression of the form

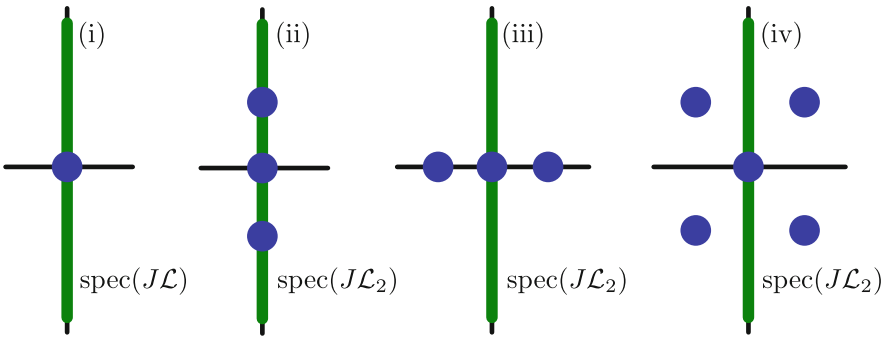

Fig. 6.5 Panel (*i*) shows the spectrum of the linearization of (6.24) about a stable primary pulse. Panels (*ii*)–(*iv*) show the anticipated three possibilities for the spectrum of the PDE linearization about a 2-pulse. The origin is an algebraically double eigenvalue in all panels

$$n(\mathscr{L}) - 1 = k_{\mathrm{r}} + 2k_{\mathrm{i}}^{-} + 2k_{\mathrm{c}}$$

holds, where $n(\mathscr{L})$ is the number of strictly negative eigenvalues of $\mathscr{L}$, while k_{r}, k_{i}^{-}, and k_{c} denote the number of pairs of real eigenvalues, pairs of purely imaginary eigenvalues with negative Krein signature (meaning that $\mathscr{L}$ is negative definite on the associated eigenspace), and quadruplets of genuinely complex eigenvalues, respectively. Theorems of this type can be found, for instance, in [7] and [4], though results of this type are not known for fifth-order KdV equations posed on the real line as the symplectic operator ∂_x is not bounded (and boundedness is a crucial assumption needed in [7] and references therein). To conclude at least spectral stability, it therefore remains to exclude the case shown in Fig. 6.5(iv). This is difficult for the following reason: the computations arising from [11] show that the quadruplet lies, to leading order, on the imaginary axis, and it is not clear how a refined analysis could exclude the possibility of a very small yet nonzero real part for those eigenvalues. Furthermore, even if the eigenvalues start out to be purely imaginary, as shown in Fig. 6.5(ii), then these eigenvalues with negative Krein signature can move off the imaginary axis as soon as they collide with eigenvalues of positive Krein signature.[1] Since the essential spectrum that occupies the imaginary axis has positive Krein signature, there does not seem to be an immediate structural reason that confines these eigenvalues to the imaginary axis. I believe that the eigenvalues lie on the imaginary axis, and one possible way of ascertaining that they do is to use the recent Krein-matrix formalism developed in [6]: this formalism shows that there can be hidden structural reasons that prevent eigenvalues from leaving the imaginary axis even when eigenvalues of opposite

[1]The Hessian of the energy restricted to the eigenspace associated with a quadruplet off the imaginary axis must decrease and increase in two transverse planes; thus eigenvalues can leave the imaginary axis only when the energy restricted to their combined eigenspace is indefinite, that is, the eigenvalues have opposite Krein signatures; see [7] and references therein.

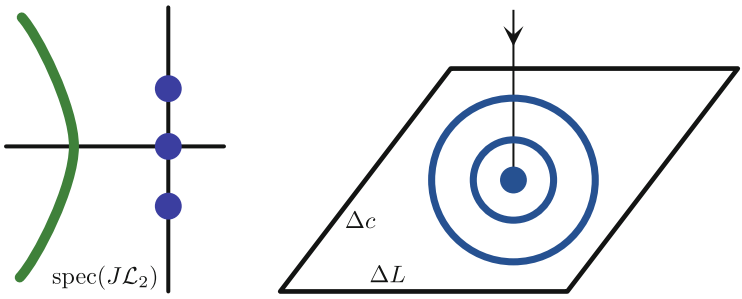

Fig. 6.6 The *left panel* indicates the anticipated spectrum of $J\mathscr{L}_2$ in an appropriate exponential weight. The *right panel* illustrates the anticipated flow on the four-dimensional center manifold, with the two directions that correspond to translation and speed taken out. The energy is still conserved so that the flow must consist of periodic orbits that surround the 2-pulse

Krein signature collide. Currently, the formalism in [6] applies only to problems with discrete spectrum, and it remains an open problem whether it can be extended to PDEs of the form (6.23).

Finally, I discuss briefly what type of nonlinear stability one might expect if the 2-pulses turn out to be spectrally stable. Following [9], the idea is to work in an appropriate exponentially weighted space in which the spectrum of $J\mathscr{L}_2$ will look as shown in Fig. 6.6. It should then be possible to extend the analysis in [9] to prove the existence of a local four-dimensional center manifold near the 2-pulse the contains the two-parameter family of 2-pulses which is parametrized by their location and speed. The other two directions consist of functions that resemble two copies of the 1-pulse whose distances and relative speeds differ by ΔL and Δc from those of the 2-pulse. This manifold will be exponentially attracting, and nearby solutions will converge to it and asymptotically follow solutions on the center manifold. Since the spectrum of the Hessian is negative definite on the eigenspace associated with the pair of purely imaginary eigenvalues shown in Fig. 6.5(ii), it follows that the flow on the nontrivial part of the center manifold consists of periodic orbits that surround the 2-pulses. Thus, the 2-pulses are expected to be nonlinearly stable in this setting, though, in contrast to the 1-pulse setting of [9], they are not asymptotically stable in the exponential weight.

Most of the discussion above is, of course, highly speculative, though I also believe that some progress on the program outlined above can be made given the recent advances on Krein-signature analyses.

Acknowledgements This paper is dedicated to Jürgen Scheurle on the occasion of his 60th birthday: I am deeply grateful for his expressions of encouragement and support when I began my career as a graduate student and postdoc.

References

1. Champneys, A.R.: Homoclinic orbits in reversible systems and their applications in mechanics, fluids and optics. Physica D **112**, 158–186 (1998)
2. Champneys, A.R., Groves, M.D.: A global investigation of solitary-wave solutions to a two-parameter model for water waves. J. Fluid Mech. **342**, 199–229 (1997)
3. Chugunova, M., Pelinovsky, D.: Two-pulse solutions in the fifth-order KdV equation: rigorous theory and numerical approximations. Discrete Contin. Dyn. Syst. Ser. B **8**, 773–800 (2007)
4. Haragus, M., Kapitula, T.: On the spectra of periodic waves for infinite-dimensional Hamiltonian systems. Physica D **237**, 2649–2671 (2008)
5. Homburg, A.J., Sandstede, B.: Homoclinic and heteroclinic bifurcations in vector fields. In: Broer, H., Takens, F., Hasselblatt, B. (eds) Handbook of Dynamical Systems, vol. III, pp. 379–524. Elsevier, Amsterdam (2010)
6. Kapitula, T.: The Krein signature, Krein eigenvalues, and the Krein oscillation theorem. Indiana Univ. Math. J. **59**, 1245–1275 (2010)
7. Kapitula, T., Kevrekidis, P.G., Sandstede, B.: Counting eigenvalues via the Krein signature in infinite-dimensional Hamiltonian systems. Physica D **195**, 263–282 (2004)
8. Lin, X.-B.: Using Melnikov's method to solve Silnikov's problems. Proc. R. Soc. Edinburgh A **116**, 295–325 (1990)
9. Pego, R.L., Weinstein, M.I.: Asymptotic stability of solitary waves. Commun. Math. Phys. **164**, 305–349 (1994)
10. Sandstede, B.: Verzweigungstheorie homokliner Verdopplungen. PhD Thesis. University of Stuttgart (1993)
11. Sandstede, B.: Stability of multiple-pulse solutions. Trans. Am. Math. Soc. **350**, 429–472 (1998)
12. Sandstede, B., Jones, C.K.R.T., Alexander, J.C.: Existence and stability of N-pulses on optical fibers with phase-sensitive amplifiers. Physica D **106**, 167–206 (1997)
13. Turaev, D.V.: Multi-pulse homoclinic loops in systems with a smooth first integral. In: Ergodic Theory, Analysis, and Efficient Simulation of Dynamical Systems, pp. 691–716. Springer, Berlin (2001)
14. Vanderbauwhede, A., Fiedler, B.: Homoclinic period blow-up in reversible and conservative systems. Z. Angew. Math. Phys. **43**, 292–318 (1992)

Chapter 7
Local Lyapunov Functions for Periodic and Finite-Time ODEs

Peter Giesl and Sigurdur Hafstein

Dedicated to Jürgen Scheurle on the occasion of his 60th birthday

Abstract Lyapunov functions for general systems are difficult to construct. However, for autonomous linear systems with exponentially stable equilibrium, there is a classical way to construct a global Lyapunov function by solving a matrix equation. Consequently, the same function is a local Lyapunov function for a nonlinear system.

In this paper, we generalise these results to time-periodic and, in particular, finite-time systems with an exponentially attractive zero solution. We show the existence of local Lyapunov functions for nonlinear systems. For finite-time systems, we consider a generalised notion of a Lyapunov function, which is not necessarily continuously differentiable, but just locally Lipschitz continuous; the derivative is then replaced by the Dini derivative.

7.1 Introduction

Lyapunov functions were introduced by Lyapunov in 1892 [22] to study stability of equilibria or other invariant sets. They can also be used to study the basin of attraction of attractors by their sublevel sets. For simplicity, we will, in the following, focus on an equilibrium or the zero solution as an attractor. The main

P. Giesl (✉)
Department of Mathematics, University of Sussex, Falmer Campus, Brighton BN1 9QH, UK
e-mail: p.a.giesl@sussex.ac.uk

S. Hafstein
School of Science and Engineering, Reykjavik University, Menntavegi 1, 101 Reykjavik, Iceland
e-mail: sigurdurh@ru.is

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_7,

features of a Lyapunov function are that it (a) decreases (strictly) along solutions and (b) attains its minimum on the attractor.

Lyapunov functions characterise certain attractivity properties and the basin of attraction; the necessity, i.e. the existence of Lyapunov functions, has been shown in so-called converse theorems. However, the construction of a Lyapunov function still remains a challenging problem. Recently, several algorithmic methods have been proposed to construct a Lyapunov function for a given system. Many of these methods face difficulties near the equilibrium or zero solution, since here the Lyapunov function does not decrease, but is constant.

Let us consider three modern construction methods, the so-called SOS method (sum of squares) [23–25], where a Lyapunov function that is presentable as a sum of squares of polynomials is constructed by convex optimisation, the CPA (continuous and piecewise affine) method [10–12, 16], where a Lyapunov function that is continuous and locally affine on each simplex of a suitable triangulation is constructed by linear programming, and the RBF (radial basis functions) method using radial basis functions to numerically solve the Zubov equation [7]. All three methods can compute Lyapunov functions on compact neighbourhoods of exponentially stable equilibria of autonomous systems and include the equilibrium in the domain of the Lyapunov function computed, given that the equilibrium is exponentially stable. These methods are, however, very different in nature. The SOS method is basically a local method, where the domain of the Lyapunov function can be enhanced by increasing the order of the polynomial Lyapunov function at the cost of greater computational complexity. The CPA and RBF methods are not local in nature and have no problems computing Lyapunov functions with large domains, if an arbitrary small neighbourhood of the equilibrium is excluded [7, 16], respectively.

The problem of including the equilibrium at the origin in the domain of Lyapunov functions for CPA and RBF for the nonlinear system $\dot{x} = f(x)$ can be overcome by studying the linearised problem $\dot{x} = Ax$, where $A = Df(0)$. For such a linear equation, there is a classical method to construct a Lyapunov function $V(x) = x^T Qx$. This function is a **local** Lyapunov function for the nonlinear system $\dot{x} = f(x)$, i.e. V decreases along solutions only in a (small) neighbourhood of the origin. Hence, the local Lyapunov function can be used to determine a local basin of attraction and close the gap between the implications of the nonlocal Lyapunov function and the local behaviour. Moreover, it can be combined with a global construction method to construct a Lyapunov function which is a true Lyapunov function even near the equilibrium. For the RBF method this was done in [8].

For the CPA method it was shown that a modified CPA method can always compute a CPA Lyapunov function including the equilibrium in its domain, first for planar systems [10, 11] and then for general n-dimensional systems [12, 13]. The key to the existence of a CPA Lyapunov function close to the equilibrium was to use the Lyapunov function $W(x) = \sqrt{x^T Qx}$, which satisfies $A\|x\|_2 \leq W(x)$ and $W'(x) \leq -B\|x\|_2$, and to interpolate this function on the edges of a suitably fine triangulation around the origin.

This paper generalises these ideas to **time-periodic** systems of the form $\dot{x} = f(t,x)$, where $f(t,x)$ is a T-periodic function, i.e. $f(t+T,x) = f(t,x)$, as well as to **finite-time** systems of the form $\dot{x} = f(t,x)$, considered over the finite-time interval $[0,T]$. The reason why we are enhancing the Lyapunov stability theory in this direction is because the CPA method has some nice properties like only assuming $f \in C^2$ and is extendable to switched systems [17] and differential inclusions [1] in a straightforward manner. Hence, the results of this paper will, besides the theoretical insight into Lyapunov functions, provide the starting point to develop a CPA construction method for Lyapunov functions for time-periodic and finite-time systems on domains, which include the attractive solution.

Lyapunov functions for periodic systems are functions $v(t,x)$ where the orbital derivative $v'(t,x) = \nabla_x v(t,x) \cdot f(t,x) + v_t(t,x)$ is negative. Such a Lyapunov function can be considered to be T-periodic without loss of generality.

Finite-time systems consider a nonautonomous equation $\dot{x} = f(t,x)$ over a finite-time interval $\mathbb{I} = [0,T]$. Finite-time dynamics were first studied in applications, in particular in fluid dynamics. The first mathematical theory was introduced by George Haller, who defined a ***Lagrangian coherent structure*** [20], i.e. time-evolving surfaces which can serve as boundaries of attraction areas. The relation of Lagrangian coherent structures to ***finite-time Lyapunov exponents*** as well as computational aspects are studied in [18, 19]. Furthermore, hyperbolicity and stable/unstable cones, which adapt the classical, infinite-time concepts of hyperbolicity and stable/unstable manifolds to the finite-time case, have been studied in [2, 5, 6].

While in the definition of hyperbolicity, attractivity is supposed to occur at every instance within the time domain under consideration, in [14, 26], a concept of attraction has been introduced, which allows that trajectories near an attracting solution move away from it, provided they return before the end of the time period. In this paper, we will use this notion of attractivity, where the distance of a solution $x(t)$ to the zero solution at time T is smaller than the distance of the solution at time 0, i.e. $\|x(T)\| < \|x(0)\|$. Note that for finite-time stability, the chosen norm is crucial, since different norms lead to different notions of attractivity; this is not the case for autonomous or periodic systems with infinite time because all norms on $\mathbb{R}^n$ are equivalent. Lyapunov functions for general nonautonomous systems have been studied in [15, 21], whereas Lyapunov functions for finite-time systems have been considered in [9, 14].

To characterise stability of zero solutions in periodic systems, one can use Floquet theory. We will show that Floquet theory is also helpful in the finite-time case; however, similar results to the periodic case can only be obtained under conditions that are stronger than assuming the attractivity of the zero solution and only for a specific type of vector norm. It turns out that in the general case a finite-time Lyapunov function can be constructed by a different approach.

An autonomous system can be regarded as a periodic system, and a periodic system can also be considered over a finite-time interval; hence, we can compare the different notions of attractivity in these cases. It turns out that finite-time

attractivity implies periodic-time attractivity, whereas the notions for autonomous and periodic-time are equivalent.

The paper is structured in the following way: in Sects. 7.2–7.4 we study autonomous, periodic, and finite-time systems, respectively. In each section, we start with linear systems, characterise exponential stability of the zero solution (equilibrium in the autonomous case), and show the existence of global Lyapunov functions. Furthermore, we consider nonlinear systems and prove similar results for local Lyapunov functions. While the results in the autonomous case are classical, parts of the periodic case are new. The main advance of the paper is the study of the finite-time case. In Sect. 7.5, we compare the notion of attractivity in periodic systems with the same system regarded as a finite-time system, and then we compare all three notions for an autonomous system. We end the paper with conclusions and an outlook for further work and applications of the results.

Notations

Definition 7.1. Consider a matrix $A \in \mathbb{C}^{n\times n}$.

1. A is called Hurwitz if all its eigenvalues have a strictly negative real part.
2. A is called Hermitian if A is equal to its conjugate transpose $A^* := \overline{A^T}$.
3. A is called positive definite if all its eigenvalues are real-valued and strictly positive.

Note that a Hermitian matrix A has real eigenvalues and $v^T Av$ is a real number for all $v \in \mathbb{R}^n$.

We denote by $\|\cdot\|_2 \colon \mathbb{C}^n \to \mathbb{R}$ the Euclidean norm $\|v\|_2 = \sqrt{\langle v, v\rangle}$, where $\langle v, w\rangle = \overline{v}^T w$, and by $\|\cdot\| \colon \mathbb{C}^n \to \mathbb{R}$ an arbitrary vector norm on $\mathbb{C}^n$. As usual, for $A \in \mathbb{C}^{n\times n}$, we denote by $\|A\|$ the induced matrix norm

$$\|A\| := \max_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\|,$$

so that $\|Ax\| \le \|A\|\|x\|$ holds for all $x \in \mathbb{C}^n$. For $x_0 \in \mathbb{R}^n$ and $\eta > 0$, we define the open ball with respect to a norm on $\mathbb{R}^n$ by $B_\eta(x_0) := \{x \in R^n \mid \|x - x_0\| < \eta\}$.

7.2 Autonomous System

The section about autonomous systems does not contain any new results, but collects classical results that are needed for the periodic and finite-time case. It is included for the convenience of the reader, and for comparison with the other two cases.

Lemma 7.1. *Let $C \in \mathbb{C}^{n\times n}$ be a Hermitian, positive definite matrix, and $L \in \mathbb{C}^{n\times n}$ be Hurwitz. Then there is a unique solution $Q \in \mathbb{C}^{n\times n}$ of the matrix equation*

$$QL + L^*Q = -C$$

and Q is Hermitian and positive definite. If C and L are real-valued, then so is Q.

The proof is similar to the real case, cf. [21, Theorem 4.6].

Definition 7.2. A (strict) local Lyapunov function for the equilibrium at the origin of system $\dot{x} = f(x)$, where $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$ with $f(0) = 0$ is a function $V \in C(U, \mathbb{R}) \cap C^1(U \setminus \{0\}, \mathbb{R})$, where $U \subset \mathbb{R}^n$ is an open neighbourhood of 0, which satisfies

1. $V(x) > 0$ for all $x \in U \setminus \{0\}$ and $V(0) = 0$ and
2. $V'(x) < 0$ for all $x \in U \setminus \{0\}$

where the orbital derivative is defined by $V'(x) = \nabla V(x) \cdot f(x)$. If $U = \mathbb{R}^n$, then the Lyapunov function is called global.

Note that the condition on differentiability of V can be dropped if the orbital derivative is replaced by the Dini derivative; this will be considered in Sect. 7.4.1. A Lyapunov function gives important information about the stability and the basin of attraction of the equilibrium 0.

Theorem 7.1. *Let V be a local Lyapunov function. Then the equilibrium 0 is asymptotically stable and any compact set $V^{-1}([0, c])$ with $c > 0$, contained in U, is a subset of the basin of attraction of 0.*

Theorem 7.2. *Consider the autonomous, linear system*

$$\dot{x} = Ax, \quad \text{where} \quad A \in \mathbb{R}^{n\times n}. \tag{7.1}$$

Every fundamental matrix solution $\Phi(t)$ of (7.1) can be expressed in the form

$$\Phi(t) = e^{tA}P_0$$

where $P_0 \in \mathbb{R}^{n\times n}$ and the zero solution of (7.1) is globally exponentially stable if and only if A is Hurwitz.

Let $C \in \mathbb{R}^{n\times n}$ be a symmetric, positive definite matrix and $Q \in \mathbb{R}^{n\times n}$ be the solution of the matrix equation $QA + A^TQ = -C$ given by Lemma 7.1; note that this implies that Q is also symmetric and positive definite.

Then $V\colon \mathbb{R}^n \to \mathbb{R}$, $V(x) := x^TQx$, and $W\colon \mathbb{R}^n \to \mathbb{R}$, $W(x) := \sqrt{V(x)} = \sqrt{x^TQx}$, are both global Lyapunov functions for (7.1), satisfying

$$\begin{aligned} a_1\|x\|_2^2 \le V(x) \le b_1\|x\|_2^2, \qquad & V'(x) \le -c_1\|x\|_2^2, \\ a_2\|x\|_2 \le W(x) \le b_2\|x\|_2, \qquad & W'(x) \le -c_2\|x\|_2. \end{aligned}$$

for all $x \in \mathbb{R}^n \setminus \{0\}$ with constants $a_1, b_1, c_1, a_2, b_2, c_2 > 0$.

In the nonlinear case, we have the following theorem, cf. [21, Corollary 4.3, Proof of Theorem 4.7].

Theorem 7.3. *Consider the autonomous, nonlinear system*

$$\dot{x} = f(x) \tag{7.2}$$

with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, $f(0) = 0$ *and* $A := Df(0)$.

The equilibrium 0 *of (7.2) is locally exponentially stable if and only if the equilibrium* 0 *of (7.1) is globally exponentially stable, i.e. by Theorem 7.2 if* A *is Hurwitz. The functions* V *and* W *from Theorem 7.2 are local Lyapunov functions for (7.2) in some open neighbourhood* U *of* 0 *and satisfy the same inequalities as in Theorem 7.2.*

7.3 Periodic Time

Most results of this section are classical; however, the explicit form of the Lyapunov functions in Theorems 7.6 and 7.7 using Floquet theory is, to the best of our knowledge, new. We start with a fundamental lemma, concerning the matrix logarithm, cf. [3, Theorem 2.47].

Lemma 7.2. *Let* $M \in \mathbb{R}^{n\times n}$ *be invertible. Then the matrix equation*

$$e^X = M \tag{7.3}$$

has a solution $X \in \mathbb{C}^{n\times n}$.

It is important to notice that, in general, even if the matrix M is real-valued, the matrix X can be complex-valued. Moreover, the solution X is not unique. A characterisation of all real-valued matrices M for which the matrix equation (7.3) has a real solution is given in [4].

7.3.1 Linear Systems

The classical Floquet Theorem gives a representation of the fundamental solution in terms of complex matrices, even if $A(t)$ is real, cf. [3, Theorem 2.48].

Theorem 7.4. *Consider the* T*-periodic system*

$$\dot{x} = A(t)x \tag{7.4}$$

where $A(t) \in C(\mathbb{R}, \mathbb{R}^{n\times n})$ *is* T*-periodic.*

Then every fundamental matrix solution $\Phi(t)$ of (7.4) can be expressed in the form

$$\Phi(t) = P(t)e^{tL} \tag{7.5}$$

where $P(t)$ is continuously differentiable and T-periodic, $P(t) \in \mathbb{C}^{n\times n}$ is invertible for all $t \in \mathbb{R}$ and $L \in \mathbb{C}^{n\times n}$.

Definition 7.3. A T-periodic (strict) local Lyapunov function for the zero solution of system $\dot{x} = f(t, x)$, where $f \in C^1(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n)$ with $f(t, 0) = 0$ for all $t \in \mathbb{R}$ and $f(t + T, x) = f(t, x)$ for all $(t, x) \in \mathbb{R} \times \mathbb{R}^n$, is a function $V \in C(\mathbb{R} \times U, \mathbb{R}) \cap C^1(\mathbb{R} \times U \setminus \{0\}, \mathbb{R})$, where $U \subset \mathbb{R}^n$ is an open neighbourhood of 0, which satisfies

1. $V(t + T, x) = V(t, x)$ for all $x \in U$ and $t \in \mathbb{R}$, i.e. V is T-periodic,
2. $V(t, x) > 0$ for all $x \in U \setminus \{0\}$ and $V(t, 0) = 0$ for all $t \in \mathbb{R}$ and
3. $V'(t, x) < 0$ for all $x \in U \setminus \{0\}$ for all $t \in \mathbb{R}$

where the orbital derivative is defined by

$$V'(t, x) = \nabla_x V(t, x) \cdot f(t, x) + V_t(t, x).$$

If $U = \mathbb{R}^n$, then the Lyapunov function is called global.

Theorem 7.5. *Let V be a T-periodic local Lyapunov function. Then the zero solution is asymptotically stable and any compact set $V^{-1}([0, c])\big|_{[0,T]\times\mathbb{R}^n}$ with $c > 0$, contained in $[0, T] \times U$, is a subset of the basin of attraction of the zero solution.*

In the following theorem we construct a T-periodic Lyapunov function for the linear system (7.4). Note that this Lyapunov function is the same as the one constructed in [21, Theorem 4.12] for the general nonautonomous case. In the periodic case, to which we restrict ourselves here, however, one can drop some assumptions on the uniformity with respect to t and, moreover, we can give a more explicit expression for V, using Floquet theory.

Theorem 7.6. *Consider the T-periodic linear equation*

$$\dot{x} = A(t)x \tag{7.6}$$

where $A(t) \in C(\mathbb{R}, \mathbb{R}^{n\times n})$ is T-periodic.

Then the zero solution of (7.6) is globally exponentially stable if and only if L is Hurwitz, where L is defined in Theorem 7.4.

Let $C \in \mathbb{C}^{n\times n}$ be a Hermitian, positive definite matrix and $Q \in \mathbb{C}^{n\times n}$ be the solution of the matrix equation $QL + L^*Q = -C$, see Lemma 7.1; note that also Q is Hermitian and positive definite.

Then $V, W : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$,

$$V(t,x) := x^T (P^{-1}(t))^* Q P^{-1}(t) x$$
$$W(t,x) := \sqrt{V(t,x)} = \sqrt{x^T (P^{-1}(t))^* Q P^{-1}(t) x}$$

are both T*-periodic global Lyapunov functions for (7.6), satisfying*

$$a_1 \|x\|_2^2 \le V(t,x) \le b_1 \|x\|_2^2, \qquad V'(t,x) \le -c_1 \|x\|_2^2,$$
$$a_2 \|x\|_2 \le W(t,x) \le b_2 \|x\|_2, \qquad W'(t,x) \le -c_2 \|x\|_2,$$

for all $x \in \mathbb{R}^n \setminus \{0\}$ *and* $t \in \mathbb{R}$ *with constants* $a_1, b_1, c_1, a_2, b_2, c_2 > 0$.

Proof. Using the transformation $y = P^{-1}(t)x$, the system is transformed into the autonomous system $\dot{y} = Ly$; the characterisation of exponential stability now follows from Theorem 7.2.

Using Theorem 7.4 we express the fundamental matrix solution with initial condition $\Phi(0) = I$ by

$$\Phi(t) = P(t) e^{tL}$$

where $P(0) = P(T) = I$, $P(t+T) = P(t)$ and $P(t), L \in \mathbb{C}^{n \times n}$. Note that since $\Phi(t)$ is a solution, we have

$$\dot{\Phi}(t) = A(t)\Phi(t) = A(t)P(t)e^{tL}$$

On the other hand,

$$\dot{\Phi}(t) = \dot{P}(t)e^{tL} + P(t)Le^{tL},$$
$$\text{which yields } \dot{P}(t) = -P(t)L + A(t)P(t).$$

Moreover, since $0 = \frac{d}{dt}\left(P(t)P^{-1}(t)\right) = \dot{P}(t)P^{-1}(t) + P(t)\dot{P}^{-1}(t)$, we have

$$\dot{P}^{-1}(t) = -P^{-1}(t)\dot{P}(t)P^{-1}(t) = LP^{-1}(t) - P^{-1}(t)A(t). \tag{7.7}$$

Note that

$$V(t,x) = x^T (P^{-1}(t))^* Q P^{-1}(t) x$$

is T-periodic and real-valued since $(P^{-1}(t))^* Q P^{-1}(t)$ is Hermitian. Moreover, since Q is positive definite and $P^{-1}(t)$ is non-singular and T-periodic, there are constants $a_1, b_1 > 0$ such that $a_1 \|x\|_2^2 \le V(t,x) \le b_1 \|x\|_2^2$ for all $x \in \mathbb{R}^n$ and $t \in \mathbb{R}$.

We show the statement for $V'(t,x)$. Using (7.7), we have

$$\begin{aligned}
V'(t,x) &= x^T(A(t))^T(P^{-1}(t))^* QP^{-1}(t)x + x^T(\dot{P}^{-1}(t))^* QP^{-1}(t)x \\
&\quad + x^T(P^{-1}(t))^* Q\dot{P}^{-1}(t)x + x^T(P^{-1}(t))^* QP^{-1}(t)A(t)x \\
&= x^T[(A(t))^T(P^{-1}(t))^* + (P^{-1}(t))^* L^* - (A(t))^*(P^{-1}(t))^*]QP^{-1}(t)x \\
&\quad + x^T(P^{-1}(t))^* Q[LP^{-1}(t) - P^{-1}(t)A(t) + P^{-1}(t)A(t)]x \\
&= x^T(P^{-1}(t))^*[L^*Q + QL]P^{-1}(t)x \\
&= -(P^{-1}(t)x)^* CP^{-1}(t)x \\
&\le -c_1\|x\|_2^2
\end{aligned}$$

for a suitable $c_1 > 0$, since C is positive definite. For the function W, we use

$$W'(t,x) = \frac{1}{2W(t,x)}V'(t,x) \le \frac{1}{2\sqrt{b_1}\|x\|_2}(-c_1\|x\|_2^2) = -\frac{c_1}{2\sqrt{b_1}}\|x\|_2. \qquad \square$$

7.3.2 Nonlinear Systems

Theorem 7.7. *Consider the T-periodic nonlinear equation*

$$\dot{x} = f(t,x) \tag{7.8}$$

where $f \in C^1(\mathbb{R}\times\mathbb{R}^n,\mathbb{R}^n)$, $f(t+T,x) = f(t,x)$, $f(t,0) = 0$ for all $t \in \mathbb{R}$ and $A(t) := D_x f(t,0)$.

Consider (7.6) with the same $A(t)$. The zero solution of (7.8) is locally exponentially stable if and only if the zero solution of (7.6) is globally exponentially stable, i.e. L is Hurwitz, where L is defined in Theorem 7.6.

The functions V and W defined in Theorem 7.6 are local Lyapunov functions for (7.8) and satisfy the same inequalities as in Theorem 7.6.

Proof. The zero solution is exponentially stable if and only if L is Hurwitz, cf. e.g. [21, Theorem 4.15]—note that the assumptions of that theorem, which holds in the more general nonautonomous case, can be relaxed, since we are focussing on the periodic case. In particular, $D_x f(t,x)$ is bounded uniformly in t, since it is periodic in t, and the Lipschitz continuity, which was used in Taylor's Theorem, can be dropped by the following argument: Using Taylor's Theorem, we can write $f(t,x) = A(t)x + \psi(t,x)$, where $\psi(t,x) = (D_x f(t,\theta x) - D_x f(t,0))\,x$ by the mean value theorem, where $\theta \in [0,1]$, i.e. $\psi(t,x) = o(\|x\|)$ as $x \to 0$, uniformly in t, since $D_x f(t,x)$ is continuous and T-periodic. Hence, for all $\epsilon > 0$ there is a $r > 0$ such that $\|\psi(t,x)\|_2 \le \epsilon\|x\|_2$ holds for all $\|x\|_2 < r$ and all $t \in \mathbb{R}$.

We show that V is a local Lyapunov function, fulfilling the inequalities. Note that the inequalities on $V(t,x)$ are clear, by Theorem 7.6, so that we only have to prove $V'(t,x) \leq -c_1\|x\|_2^2$; note that in the nonlinear case, $V'(t,x)$ is different to the linear case.

Since C is Hermitian and positive definite, there is a smallest eigenvalue $\lambda > 0$ of C such that $y^T Cy \geq \lambda\|y\|_2^2$ for all $y \in \mathbb{R}^n$. Set $\epsilon := \frac{\lambda}{4\|Q\|_2 \max_{t\in[0,T]}(\|P^{-1}(t)\|_2\|P(t)\|_2)}$ and choose $r > 0$ as above such that $\|\psi(t,x)\|_2 \leq \epsilon\|x\|_2$ holds for all $\|x\|_2 < r$ and all $t \in \mathbb{R}$. Then, similar to the theorem in the linear case, we have, using (7.7)

$$\begin{aligned}
V'(t,x) &= x^T(A(t))^T(P^{-1}(t))^* QP^{-1}(t)x + \psi(t,x)^T(P^{-1}(t))^* QP^{-1}(t)x \\
&\quad + x^T(\dot{P}^{-1}(t))^* QP^{-1}(t)x + x^T(P^{-1}(t))^* Q\dot{P}^{-1}(t)x \\
&\quad + x^T(P^{-1}(t))^* QP^{-1}(t)A(t)x + x^T(P^{-1}(t))^* QP^{-1}(t)\psi(t,x) \\
&\leq x^T[(A(t))^T(P^{-1}(t))^* + (P^{-1}(t))^* L^* - (A(t))^*(P^{-1}(t))^*]QP^{-1}(t)x \\
&\quad + x^T(P^{-1}(t))^* Q[LP^{-1}(t) - P^{-1}(t)A(t) + P^{-1}(t)A(t)]x \\
&\quad + 2\|Q\|_2\|P^{-1}(t)\|_2\|P^{-1}(t)x\|_2\|\psi(t,x)\|_2 \\
&= x^T(P^{-1}(t))^*[L^*Q + QL]P^{-1}(t)x \\
&\quad + 2\epsilon\|Q\|_2\|P^{-1}(t)\|_2\|P^{-1}(t)x\|_2\|x\|_2 \\
&= -(P^{-1}(t)x)^* CP^{-1}(t)x \\
&\quad + 2\epsilon\|Q\|_2 \max_{t\in[0,T]}(\|P^{-1}(t)\|_2\|P(t)\|_2)\|P^{-1}(t)x\|_2^2 \\
&\leq \left(-\lambda + \frac{\lambda}{2}\right)\|P^{-1}(t)x\|_2^2 \\
&\leq -c_1\|x\|_2^2
\end{aligned}$$

for all $\|x\|_2 < r$ with a suitable $c_1 > 0$. The argumentation for W is as in Theorem 7.6. □

7.4 Finite Time

For this section, we fix an arbitrary norm $\|\cdot\|$ on $\mathbb{R}^n$. We consider the nonautonomous ODE

$$\dot{x} = f(t,x) \tag{7.9}$$

where $f \in C^1([0,T]\times\mathbb{R}^n,\mathbb{R}^n)$ over the finite-time interval $\mathbb{I} = [0,T]$. We denote the solution of (7.9) with initial value $x(t_0) = x_0$ by $\varphi(t,t_0,x_0) := x(t)$ and

assume that it exists in the whole interval $[0,T]$. This is e.g. the case if $D_x f(t,x)$ is bounded. We will later assume that $\mu(t)=0$ is a solution, i.e. $f(t,0)=0$ for all $t\in\mathbb{I}$. We use the following definition of finite-time attractivity from [14,26].

Definition 7.4 (Finite-Time Attractivity, Domain of Attraction). Let $\mu : \mathbb{I} \to \mathbb{R}^n$ be a solution of (7.9).

1. μ is called ***attractive on*** $\mathbb{I}$ with respect to the norm $\|\cdot\|$ if there exists an $\eta>0$ such that

$$\|\varphi(T,0,\xi)-\mu(T)\| < \|\xi-\mu(0)\| \;\forall \xi\in B_\eta(\mu(0))\setminus\{\mu(0)\}\,.$$

2. μ is called ***exponentially attractive on*** $\mathbb{I}$ with respect to the norm $\|\cdot\|$ if

$$\limsup_{\eta\searrow 0}\frac{1}{\eta}\sup_{\xi\in B_\eta(0)}\left(\|\varphi\left(T,0,\xi\right)-\mu(T)\|\right)<1\,,$$

   and the negative number

$$\frac{1}{T}\ln\left(\limsup_{\eta\searrow 0}\frac{1}{\eta}\sup_{\xi\in B_\eta(0)}\left(\|\varphi\left(T,0,\xi\right)-\mu(T)\|\right)\right)$$

   is called ***rate of exponential attraction***.
3. Let $\mu:\mathbb{I}\to\mathbb{R}^n$ be an attractive solution on $\mathbb{I}$. Then a connected and invariant nonautonomous (i.e. $G_\mu(t):=\{x\in\mathbb{R}^n \mid (t,x)\in G_\mu\}$ is nonempty for all $t\in\mathbb{I}$) set $G_\mu\subset\mathbb{I}\times\mathbb{R}^n$ is called ***domain of attraction of*** μ if

$$\|\varphi(T,0,x)-\mu(T)\| < \|x-\mu(0)\| \text{ holds for all } x\in G_\mu(0)\setminus\{\mu(0)\}\,,$$

   and G_μ is the maximal set containing $\mathrm{graph}(\mu)$ with this property.

In order to study the local properties of linear and nonlinear systems, we can use a Floquet-like theorem to define a local Lyapunov function. The following theorem is similar to the classical Floquet Theorem, but does not require $A(t)$ to be periodic. Thus, $P(t)$ is not periodic either, but we can still show that $P(0)=P(T)$ holds.

Theorem 7.8. *Consider the nonautonomous linear system*

$$\dot{x}=A(t)x \tag{7.10}$$

where $A\in C(\mathbb{I},\mathbb{R}^{n\times n})$. The principal solution, i.e. satisfying $\Phi(0)=I$, of (7.10) can be expressed in the form

$$\Phi(t)=P(t)e^{tL}$$

where $P(t)$ is continuously differentiable, $P(t)\in\mathbb{C}^{n\times n}$ is invertible for all $t\in\mathbb{I}$, $P(0)=P(T)=I$ and $L\in\mathbb{C}^{n\times n}$.

Proof. Define $M := \Phi^{-1}(0)\Phi(T)$. By Lemma 7.2, there is a matrix $L \in \mathbb{C}^{n\times n}$ such that $e^{TL} = M$. With $P(t) := \Phi(t)e^{-tL}$ we have

$$P(T) = \Phi(T)e^{-TL} = \Phi(T)M^{-1} = \Phi(0) = P(0) = I$$

using $\Phi(0) = I$, and $P(t)$ fulfills all the stated properties. □

We will first reprove the characterisation of finite-time exponential stability which was given in [14] for the Euclidean norm, now for a general norm $\|\cdot\|$.

Theorem 7.9. *Denote by $F_T\colon \mathbb{R}^n \to \mathbb{R}^n$ the time–T map of (7.9), which is defined by $F_T(x) := \varphi(T, 0, x)$. Moreover, let $\mu : \mathbb{I} \to \mathbb{R}^n$, $\mu(t) = 0$ be a solution of (7.9). Then μ is exponentially attractive on $\mathbb{I}$ if and only if $\|DF_T(0)\| < 1$, where DF_T is the Jacobian of F_T with respect to x. The rate of exponential attraction is given by*

$$\frac{1}{T}\ln\|DF_T(0)\|.$$

If the principal solution $\Phi(t)$ of the linearised equation

$$\dot{x} = D_x f(t, 0)x$$

with $\Phi(0) = I$ is expressed $\Phi(t) = P(t)e^{tL}$ as in Theorem 7.8, we have $DF_T(0) = e^{TL}$.

In particular the zero solution μ is exponentially attractive on $\mathbb{I}$ if and only if $\|e^{TL}\| < 1$.

Proof. We consider $\mu(t) = 0$ and the solution $\varphi(t, 0, w)$ starting in $w \in \mathbb{R}^n$. Using Taylor's Theorem, we obtain

$$\varphi(T, 0, w) - 0 = F_T(w) - F_T(0) = DF_T(0)w + \psi(w),$$

where $\lim_{\|w\|\to 0} \frac{\psi(w)}{\|w\|} = 0$. Thus,

$$\limsup_{\|w\|\to 0} \frac{\|\varphi(T, 0, w)\|}{\|w\|} = \limsup_{\|w\|\to 0} \frac{\|DF_T(0)w\|}{\|w\|} = \|DF_T(0)\|. \quad (7.11)$$

$$\text{Now } \frac{1}{\eta}\sup_{\xi\in B_\eta(0)} \|\varphi(T, 0, \xi)\| = \sup_{\|w\|<\eta} \frac{\|\varphi(T, 0, w)\|}{\|w\|}\frac{\|w\|}{\eta}. \quad (7.12)$$

From (7.12) we can conclude

$$\frac{1}{\eta}\sup_{\xi\in B_\eta(0)} \|\varphi(T, 0, \xi)\| \le \sup_{\|w\|<\eta} \frac{\|\varphi(T, 0, w)\|}{\|w\|}$$

which implies with (7.11) that

$$\limsup_{\eta \searrow 0} \frac{1}{\eta} \sup_{\xi \in B_\eta(0)} \|\varphi(T,0,\xi)\| \le \limsup_{\|w\| \to 0} \frac{\|\varphi(T,0,w)\|}{\|w\|} = \|DF_T(0)\|.$$

Furthermore, (7.12) and (7.11) yield for all fixed $\theta \in (0,1)$

$$\frac{1}{\eta} \sup_{\xi \in B_\eta(0)} \|\varphi(T,0,\xi)\| \ge \sup_{\|w\| = \theta\eta} \frac{\|\varphi(T,0,w)\|}{\|w\|} \theta$$

and

$$\limsup_{\eta \searrow 0} \frac{1}{\eta} \sup_{\xi \in B_\eta(0)} \|\varphi(T,0,\xi)\| \ge \limsup_{\|w\| \to 0} \frac{\|\varphi(T,0,w)\|}{\|w\|} \theta = \theta \|DF_T(0)\|.$$

Since this inequality holds for all $\theta \in (0,1)$, we have

$$\limsup_{\eta \searrow 0} \frac{1}{\eta} \sup_{\xi \in B_\eta(0)} \|\varphi(T,0,\xi)\| \ge \|DF_T(0)\|.$$

This shows $\limsup_{\eta \searrow 0} \frac{1}{\eta} \sup_{\xi \in B_\eta(0)} \|\varphi(T,0,\xi)\| = \|DF_T(0)\|$.

Furthermore, we can relate DF_T to the solution of the linearised equation. Denote by $F_t \colon \mathbb{R}^n \to \mathbb{R}^n$ the time-t map of (7.9), which is defined by $F_t(x) := \varphi(t,0,x)$. Then $\Phi(t,x) = DF_t(x)$ solves the first variation equation

$$\dot{\Phi}(t,x) = D_x f(t, \varphi(t,0,x)) \Phi(t,x).$$

In particular, as $\mu(t) = 0$ is a solution of (7.9), we obtain

$$\dot{\Phi}(t,0) = D_x f(t,0) \Phi(t,0) = A(t) \Phi(t,0)$$

with solution $\Phi(t,0) = P(t) e^{tL}$. Thus, $DF_T(0) = \Phi(T,0) = P(T) e^{TL} = e^{TL}$. Hence, the zero solution of the nonlinear equation is exponentially stable if and only if $\|e^{TL}\| < 1$. □

7.4.1 *Dini Derivative*

Due to the general norm $\|\cdot\|$, the assumption that a Lyapunov function $V(t,x)$ is C^1 is too restrictive. For example, for the system $\dot{x} = -x$, $x \in \mathbb{R}$ and $\|x\| := |x|$, the function $V(t,x) = |x|$ is not C^1 at 0, but it is a Lyapunov function in the sense that it is decreasing along trajectories. We will give a precise definition in Definition 7.5. We only assume that a Lyapunov function is continuous and locally Lipschitz in x, and we have to replace the orbital derivative by a weaker notion, the Dini derivative. Note that this can also be done in the autonomous and periodic case.

We define a finite-time Lyapunov function. The definition is similar to the periodic case, but V is fixed at times 0 and T by the norm.

Definition 7.5. A finite-time (strict) local Lyapunov function for the zero solution of the system (7.9) is a continuous function $V : \mathbb{I} \times U \to \mathbb{R}^n$, where $U \subset \mathbb{R}^n$ is an open neighbourhood of 0, which satisfies the following properties:

1. $V(t,x)$ is locally Lipschitz in x.
2. $V(0,x) = \|x\|^p$ and $V(T,x) = \|x\|^p$ for all $x \in U$, where $p \geq 1$.
3. $V(t,x) > 0$ for all $x \in U \setminus \{0\}$ and $V(t,0) = 0$ for all $t \in \mathbb{I}$.
4. $V^+(t,x) < 0$ for all $x \in U \setminus \{0\}$ and all $t \in \mathbb{I} \setminus \{T\} = [0,T)$.

Here the orbital derivative is defined by the Dini derivative

$$V^+(t,x) = \limsup_{h \searrow 0} \frac{V(t+h, x+h \cdot f(t,x)) - V(t,x)}{h}. \tag{7.13}$$

If $U = \mathbb{R}^n$, then the function is a global finite-time Lyapunov function.

Remark 7.1. Locally Lipschitz in x in 1. is defined as follows: for every compact $C \subset \mathbb{I} \times U$ there exists a constant $L > 0$, such that $|V(t,x) - V(t,y)| \leq L\|x-y\|$ for all $(t,x),(t,y) \in C$. It is needed to define the orbital derivative using $f(t,x)$ in 4. by (7.13) as shown on page 48 in [17]. Without this property

$$V^+(t,x) = \limsup_{h \searrow 0} \frac{V(t+h, \varphi(t+h,t,x)) - V(t,x)}{h}$$

is **not** necessarily true.

Remark 7.2. If V is differentiable along orbits, i.e. the limit

$$\lim_{h \to 0} \frac{V(t+h, \varphi(t+h,t,x)) - V(t,x)}{h}$$

exists for all relevant t and x, then clearly $V^+(t,x)$ is equal to this limit. Note, however, that this does not imply that $\nabla_x V(t,x)$ or $V_t(t,x)$ exist, e.g. consider the example at the beginning of Sect. 7.4.1.

Theorem 7.10. *Let V be a finite-time local Lyapunov function for the system (7.9). Then the zero solution of (7.9) is attractive and any compact set $V^{-1}([0,C])$ with $C > 0$, contained in $\mathbb{I} \times U$, is a subset of the domain of attraction of the zero solution.*

Let $V(t,x)$ additionally fulfill: There exist constants $b \geq 1$ and $c > 0$ such that

1. *$V(t,x) \leq b\|x\|^p$ for all $t \in \mathbb{I}$ and all $x \in U$, where p is the same as in Definition 7.5.*
2. *$V^+(t,x) \leq -c\|x\|^p$ for all $t \in [0,T) = \mathbb{I} \setminus \{T\}$ and all $x \in U$.*

Then the zero solution of (7.9) is exponentially attractive with rate of exponential attraction $\leq -c/(bp)$.

Proof. Since $x(t) = 0$ is a solution, there is an open neighbourhood $U' \subset U$ of 0 such that $x \in U'$ implies $\varphi(t, 0, x) \in U$ for all $t \in \mathbb{I}$.

Because V is locally Lipschitz it follows by [17, p. 48 and Corollary 3.10] that $t \mapsto V(t, \varphi(t, 0, x))$ is a strictly decreasing function on $\mathbb{I}$, if $x \neq 0$. Hence,

$$\|x\|^p = V(0, x) > V(T, \varphi(T, 0, x)) = \|\varphi(T, 0, x)\|^p$$

for all $x \in U' \setminus \{0\}$ and the zero solution is attractive.

The same argument shows that $V^{-1}([0, C])$, if it is contained in $\mathbb{I} \times U$, is positively invariant. Now let $(t_0, x_0) \in V^{-1}([0, C])$ with $x_0 \neq 0$. Then either the trajectory $\varphi(t, t_0, x_0)$ stays in $V^{-1}([0, C])$ for all $t \in \mathbb{I}$, or there is a $\tau \in \mathbb{I}$ such that $\varphi(\tau, t_0, x_0) \notin V^{-1}([0, C])$. In the first case, (t_0, x_0) is in the domain of attraction by the fact that $t \mapsto V(t, \varphi(t, t_0, x_0))$ is a strictly decreasing function. In the second case, note that $\varphi(\tau, t_0, x_0) \notin V^{-1}([0, C])$ implies that $\varphi(0, t_0, x_0) \notin V^{-1}([0, C])$, since $V^{-1}([0, C])$ is positively invariant. Again, by the positive invariance, we have $\varphi(T, t_0, x_0) \in V^{-1}([0, C])$. Hence,

$$\|\varphi(T, t_0, x_0)\|^p = V(T, \varphi(T, t_0, x_0)) \leq C < V(0, \varphi(0, t_0, x_0)) = \|\varphi(0, t_0, x_0)\|^p$$

which shows that also in this case (t_0, x_0) lies in the domain of attraction. This proves the first claim of the theorem.

Now, assume that 1. and 2. are also fulfilled. Then V fulfills the Dini differential inequality

$$V^+(t, \varphi(t, 0, x)) \leq -\frac{c}{b} V(t, \varphi(t, 0, x))$$

and by taking Lemma 6.10 in [17] into consideration, we get $V(T, \varphi(T, 0, x)) \leq V(0, x)e^{-cT/b}$ and thus $\|\varphi(T, 0, x)\|^p \leq \|x\|^p e^{-cT/b}$ so for $x \neq 0$

$$\frac{\|\varphi(T, 0, x)\|}{\|x\|} \leq e^{-cT/(bp)} < 1$$

and by Definition 7.4, 2., the zero solution is exponentially attractive with rate of exponential attraction $\leq -c/(bp)$. □

Remark 7.3. Note the obvious error in the statement of Lemma 6.10 in [17]. Of course $L_{\mathscr{C}}$ is a local Lipschitz constant for s and not y as should be clear from the text and from the proof. y does not have to be locally Lipschitz. This is an important point for otherwise $V(t, x)$ would have to be locally Lipschitz in (t, x) and not only x.

Note that the Dini derivative does in general not obey the chain-rule. To see this consider e.g. $f(x) = |x|$, $g(x) = -x$ and $h(x) = (f \circ g)(x) = |-x|$. Then

$$h^+(0) = \limsup_{\eta \searrow 0} \frac{|-\eta| - |0|}{\eta} = 1 \text{ but}$$

$$f^+(g(0)) \cdot g^+(0) = \limsup_{\eta \searrow 0} \frac{|-0+\eta| - |-0|}{\eta} \cdot \limsup_{\eta \searrow 0} \frac{-\eta - (-0)}{\eta} = 1 \cdot (-1) = -1.$$

However, for our needs the following simple lemma suffices.

Lemma 7.3. *Let $f : \mathbb{R} \to \mathbb{R}$ be a function such that $\limsup_{h \searrow 0} f(h) = S < 0$. Then there is a $\tau > 0$ such that $f(x) < 0$ for all $x \in (0, \tau)$. If $g : \mathbb{R} \to \mathbb{R}$ is a further function such that $\lim_{h \searrow 0} g(h) = L \neq 0$, then $\limsup_{h \searrow 0} f(h)g(h) = SL$.*

Proof. Assume there is no such $\tau > 0$. Then

$$\limsup_{h \searrow 0} f(h) = \lim_{h \searrow 0} [\sup\{f(x) \mid x \in (0, h)\}] \geq \lim_{h \searrow 0} 0 = 0,$$

which is a contradiction to $\limsup_{h \searrow 0} f(h) = S < 0$.

Now assume $L > 0$ and let $0 < \varepsilon < L/2$ be arbitrary. Then, for all $\tau > h > 0$ small enough, we have $0 < L - \varepsilon < g(h) < L + \varepsilon$, i.e. $(L+\varepsilon)f(h) \leq g(h)f(h) \leq (L-\varepsilon)f(h)$, and therefore

$$(L+\varepsilon)S \leq \limsup_{h \searrow 0} f(h)g(h) \leq (L-\varepsilon)S,$$

i.e. $\limsup_{h \searrow 0} f(h)g(h) = SL$ by lack of alternatives. The case $L < 0$ follows similarly. □

7.4.2 Linear Systems

Let us first focus on the linear case, i.e.

$$\dot{x} = A(t)x. \tag{7.14}$$

We can use the construction method in [14], which uses linear interpolation along a trajectory between the values at times 0 and T, to construct finite-time Lyapunov functions in the following two theorems.

Theorem 7.11. *Let the zero solution of*

$$\dot{x} = A(t)x, \tag{7.15}$$

where $A \in C(\mathbb{I}, \mathbb{R}^{n \times n})$, be exponentially stable.

Then there exists a finite-time Lyapunov function which satisfies

$$a_1\|x\|^2 \le V(t,x) \le b_1\|x\|^2, \qquad V^+(t,x) \le -c_1\|x\|^2 \tag{7.16}$$

for all $x \in \mathbb{R}^n$ *and all* $t \in \mathbb{I} \setminus \{T\}$, *where* $a_1, b_1, c_1 > 0$.

Proof. We define $V(t, \varphi(t,0,x))$ by linear interpolation of the values at time T and 0. Note that the principal fundamental solution $\Phi(t)$ (with $\Phi(0) = I$) of (7.15) can be, by Theorem 7.8, expressed in the form $\Phi(t) = P(t)e^{tL}$, where $P(t)$ is continuously differentiable, $P(t) \in \mathbb{C}^{n\times n}$ is invertible for all $t \in \mathbb{I}$, $P(0) = P(T) = I$ and $L \in \mathbb{C}^{n\times n}$. Hence, $\varphi(t_1,t_2,x) = \Phi(t_1)\Phi^{-1}(t_2)x = P(t_1)e^{L(t_1-t_2)}P^{-1}(t_2)x$. We define

$$\begin{aligned} V(t,x) &:= [\|\varphi(T,t,x)\|^2 - \|\varphi(0,t,x)\|^2]\frac{t}{T} + \|\varphi(0,t,x)\|^2 \qquad (7.17)\\ &= \frac{t}{T}\|\varphi(T,t,x)\|^2 + \left(1-\frac{t}{T}\right)\|\varphi(0,t,x)\|^2 \\ &= \frac{t}{T}\|e^{TL}\Phi^{-1}(t)x\|^2 + \left(1-\frac{t}{T}\right)\|\Phi^{-1}(t)x\|^2 \qquad (7.18)\end{aligned}$$

It is easy to see that $V(0,x) = V(T,x) = \|x\|^2$. Since $e^{TL} \neq 0$, $\Phi^{-1}(t)$ is nonsingular and continuous for $t \in \mathbb{I}$ and the norm $\|\cdot\|$ is continuous, the mappings $(t,x) \mapsto \|e^{TL}\Phi^{-1}(t)x\|$ and $(t,x) \mapsto \|\Phi^{-1}(t)x\|$ are both continuous functions from the compact set $\mathbb{I} \times \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ into the real numbers. Hence, there are $a_1, b_1 > 0$ such that

$$a_1\|x\|^2 \le \|e^{TL}\Phi^{-1}(t)x\|^2 \le b_1\|x\|^2 \tag{7.19}$$

$$a_1\|x\|^2 \le \quad \|\Phi^{-1}(t)x\|^2 \quad \le b_1\|x\|^2 \tag{7.20}$$

for all $t \in \mathbb{I}$. Together with (7.18) this shows the first part of (7.16).

To show that $V(t,x)$ is locally Lipschitz in x let $\eta > 0$ be an arbitrary constant and $x, y \in \overline{B_\eta(0)}$ and $t \in \mathbb{I}$. By (7.18), (7.19) and (7.20) we get

$$\begin{aligned} &|V(t,x) - V(t,y)| \\ &= \Big|\frac{t}{T}(\|e^{TL}\Phi^{-1}(t)x\| + \|e^{TL}\Phi^{-1}(t)y\|)(\|e^{TL}\Phi^{-1}(t)x\| - \|e^{TL}\Phi^{-1}(t)y\|) \\ &\quad + \left(1-\frac{t}{T}\right)(\|\Phi^{-1}(t)x\| + \|\Phi^{-1}(t)y\|)(\|\Phi^{-1}(t)x\| - \|\Phi^{-1}(t)y\|)\Big| \\ &\le \sqrt{b_1}(\|x\| + \|y\|)\left[\frac{t}{T}\|e^{TL}\Phi^{-1}(t)(x-y)\| + \left(1-\frac{t}{T}\right)\|\Phi^{-1}(t)(x-y)\|\right] \end{aligned}$$

$$
\begin{aligned}
&\leq \sqrt{b_1}(\|x\| + \|y\|)\left[\frac{t}{T}\sqrt{b_1}\|x-y\| + \left(1-\frac{t}{T}\right)\sqrt{b_1}\|x-y\|\right] \\
&= b_1(\|x\| + \|y\|)\|x-y\| \qquad (7.21)\\
&\leq 2\eta b_1\|x-y\|,
\end{aligned}
$$

which proves that $V(t,x)$ is locally Lipschitz in x.

Finally, we show the second part of (7.16). For every $(t_0,x_0) \in \mathbb{I}\times\mathbb{R}^n$ define the function $\psi_{(t_0,x_0)}(t) := V(t,\varphi(t,t_0,x_0))$ on $\mathbb{I}\setminus\{T\}$. We claim that for every $(t_0,x_0) \in \mathbb{I}\times\mathbb{R}^n$ the function $\psi_{(t_0,x_0)}(t)$ is differentiable with respect to t. To see this note that by (7.17) and the semigroup property $\varphi(t_1,t_2,\varphi(t_2,t_3,x)) = \varphi(t_1,t_3,x)$

$$
\begin{aligned}
\psi_{(t_0,x_0)}(t) &= [\|\varphi(T,t,\varphi(t,t_0,x_0))\|^2 - \|\varphi(0,t,\varphi(t,t_0,x_0))\|^2]\frac{t}{T} \\
&\quad +\|\varphi(0,t,\varphi(t,t_0,x_0))\|^2 \\
&= [\|\varphi(T,t_0,x_0)\|^2 - \|\varphi(0,t_0,x_0)\|^2]\frac{t}{T} + \|\varphi(0,t_0,x_0)\|^2
\end{aligned}
$$

so that

$$
\psi'_{(t_0,x_0)}(t) = [\|\varphi(T,t_0,x_0)\|^2 - \|\varphi(0,t_0,x_0)\|^2]\frac{1}{T}.
$$

By Theorem 7.9 we have $\|e^{TL}\| =: \nu \in (0,1)$, since the zero solution is exponentially stable. Since $V(t,x)$ is locally Lipschitz in x we have by Remark 7.2, the product rule for differentiation, (7.17) and (7.20) that

$$
\begin{aligned}
V^+(t_0,x_0) &= \limsup_{h\searrow 0}\frac{V(t_0+h,\varphi(t_0+h,t_0,x_0)) - V(t_0,x_0)}{h} \\
&= \limsup_{h\searrow 0}\frac{\psi_{(t_0,x_0)}(t_0+h) - \psi_{(t_0,x_0)}(t_0)}{h} \\
&= \psi'_{(t_0,x_0)}(t_0) \;=\; \frac{1}{T}\left(\|\varphi(T,t_0,x_0)\|^2 - \|\varphi(0,t_0,x_0)\|^2\right) \\
&= \frac{1}{T}\left(\|e^{TL}\Phi^{-1}(t_0)x_0\|^2 - \|\Phi^{-1}(t_0)x_0\|^2\right) \\
&\leq \frac{\nu^2-1}{T}\|\Phi^{-1}(t_0)x_0\|^2 \\
&\leq -\frac{1-\nu^2}{T}a_1\|x_0\|^2
\end{aligned}
$$

Hence, with $c_1 := (1-\nu^2)a_1/T > 0$, the rest of (7.16) is shown. □

We will show the existence of another Lyapunov function; note that this is particularly useful if one wants to approximate it by a continuous piecewise

affine function. Indeed, the authors showed in [12] that such a function can be approximated by a continuous piecewise affine function so closely that the approximation is a true CPA Lyapunov function for autonomous systems, even in a neighbourhood of the equilibrium.

Theorem 7.12. *Let the zero solution of the system (7.15) be exponentially stable. Then there exists a finite-time Lyapunov function $W(t,x)$ which satisfies*

$$a_2\|x\| \le W(t,x) \le b_2\|x\|, \qquad W^+(t,x) \le -c_2\|x\| \tag{7.22}$$

for all $x \in \mathbb{R}^n$ and $t \in \mathbb{I} \setminus \{T\}$, where $a_2, b_2, c_2 > 0$. W is globally Lipschitz in x.

Proof. We define $W(t,x) := \sqrt{V(t,x)}$, where $V(t,x)$ is the function from Theorem 7.11, and notice that immediately from (7.16) $\sqrt{a_1}\|x\| \le W(t,x) \le \sqrt{b_1}\|x\|$ follows. It is easy to see that $W(0,x) = W(T,x) = \|x\|$. To show the second part of (7.22) fix $t \in [0,T)$. The case $x = 0$ follows from

$$\frac{\sqrt{V(t+h,0+h\cdot A(t)0)} - \sqrt{V(t,0)}}{h} = \frac{0-0}{h} = 0,$$

i.e. $W^+(t,0) = 0$. If $x \neq 0$ we have by Lemma 7.3

$$\begin{aligned} W^+(t,x) &= \limsup_{h\searrow 0} \frac{\sqrt{V(t+h,x+h\cdot A(t)x)} - \sqrt{V(t,x)}}{h} \\ &= \limsup_{h\searrow 0} \frac{V(t+h,x+h\cdot A(t)x) - V(t,x)}{h\cdot(\sqrt{V(t,x+h\cdot A(t)x)} + \sqrt{V(t,x)})} \\ &\le \frac{-c_1\|x\|^2}{2\sqrt{V(t,x)}} \le -\frac{c_1}{2\sqrt{b_1}}\|x\|. \end{aligned}$$

It remains to show that $W(t,x)$ is globally Lipschitz in x. The case $x = y = 0$ is trivial and otherwise, by (7.21) and (7.16), we have

$$\begin{aligned} |W(t,x) - W(t,y)| &= \left|\sqrt{V(t,x)} - \sqrt{V(t,y)}\right| = \frac{|V(t,x) - V(t,y)|}{\sqrt{V(t,x)} + \sqrt{V(t,y)}} \\ &\le \frac{b_1(\|x\| + \|y\|)}{\sqrt{a_1}\cdot(\|x\| + \|y\|)} \cdot \|x-y\| \le \frac{b_1}{\sqrt{a_1}} \cdot \|x-y\|. \end{aligned}$$

□

Remark 7.4. A different function W with the properties as in Theorem 7.12 is

$$W_2(t,x) := [\|\varphi(T,t,x)\| - \|\varphi(0,t,x)\|]\frac{t}{T} + \|\varphi(0,t,x)\|.$$

One can prove the properties, following the proof of Theorem 7.11, dropping the squares.

We give an example that these two definitions lead to two different functions W_1 and W_2. Consider $\dot{x} = -x$ on the interval $[0, 1]$ with the Euclidean norm. Then

$$V(t,x) = \left(te^{2(t-1)} + (1-t)e^{2t}\right) x^2,$$
$$W_1(t,x) = \sqrt{te^{2(t-1)} + (1-t)e^{2t}}|x|,$$
$$W_2(t,x) = \left(te^{t-1} + (1-t)e^{t}\right)|x|.$$

7.4.3 Nonlinear Systems

Now we consider the nonlinear system

$$\dot{x} = f(t,x) \tag{7.23}$$

over the finite-time interval $\mathbb{I} = [0,T]$ and show the existence of local finite-time Lyapunov functions.

Theorem 7.13. *Consider the nonlinear system*

$$\dot{x} = f(t,x) \tag{7.24}$$

where $f \in C^1([0,T] \times \mathbb{R}^n, \mathbb{R}^n)$, $f(t,0) = 0$ for all $t \in [0,T]$ over the finite-time interval $\mathbb{I} = [0,T]$. Define $A(t) := D_x f(t,0)$.

Consider (7.15) with the same $A(t)$. Then the Lyapunov functions V and W in Theorems 7.11 and 7.12 respectively are also finite-time Lyapunov functions for (7.24) satisfying $V^+(t,x) \le -c_v\|x\|^2$ and $W^+(t,x) \le -c_w\|x\|$, $c_v, c_w > 0$, for all $t \in \mathbb{I} \setminus \{T\}$ and all x in some open neighbourhood $U \subset \mathbb{R}^n$ of 0.

Proof. It suffices to show that $V^+(t,x) \le -c_v\|x\|^2$ and $W^+(t,x) \le -c_w\|x\|$ for all $t \in \mathbb{I} \setminus \{T\}$ and all x in some open neighbourhood $U \subset \mathbb{R}^n$ of 0.

We first show $W^+(t,x) \le -c_w\|x\|$. By Taylor's Theorem we can write $f(t,x) = A(t)x + \psi(t,x)$, where for all $\epsilon > 0$ there is a $\eta > 0$ such that $\|\psi(t,x)\| \le \epsilon\|x\|$ holds for all $\|x\| < \eta$ and all $t \in \mathbb{I}$, cf. the proof of Theorem 7.7. Because $W(t,x)$ is globally Lipschitz in x by Theorem 7.12, there is a constant $L > 0$ such that $|W(t,x) - W(t,y)| \le L\|x-y\|$ for all $t \in \mathbb{I}$ and $x, y \in \mathbb{R}^n$. Hence by (7.22)

$$\begin{aligned} W^+(t,x) &= \limsup_{h \searrow 0} \frac{W(t+h, x+h\cdot f(t,x)) - W(t,x)}{h} \\ &\le \limsup_{h \searrow 0} \frac{W(t+h, x+h\cdot[A(t)x+\psi(t,x)]) - W(t+h, x+h\cdot A(t)x)}{h} \end{aligned}$$

$$
\begin{aligned}
&+ \limsup_{h \searrow 0} \frac{W(t+h, x+h \cdot A(t)x) - W(t,x)}{h} \\
&\leq \limsup_{h \searrow 0} \frac{L\|h\psi(t,x)\|}{h} - c_2\|x\| \\
&\leq L\|\psi(t,x)\| - c_2\|x\|.
\end{aligned}
$$

With $\varepsilon := c_2/(2L)$ and $c_w := c_2/2$ it follows that there exists an $\eta > 0$ such that $W^+(t,x) \leq -c_w\|x\|$ for all $t \in \mathbb{I} \setminus \{T\}$ and all $x \in B_\eta(0) =: U$.

Now consider $V(t,x) = [W(t,x)]^2$. The case $x = 0$ is trivial. For $x \neq 0$ we have by Lemma 7.3 and the above estimate

$$
\begin{aligned}
V^+(t,x) &= \limsup_{h \searrow 0} \frac{W^2(t+h, x+h \cdot f(t,x)) - W^2(t,x)}{h} \\
&\leq \limsup_{h \searrow 0} [W(t+h, x+h \cdot f(t,x)) + W(t,x)] \frac{W(t+h, x+h \cdot f(t,x)) - W(t,x)}{h} \\
&= 2W(t,x) \cdot [-c_w\|x\|] \leq -a_2 c_w \|x\|^2
\end{aligned}
$$

for all $t \in \mathbb{I} \setminus \{T\}$ and all $x \in B_\eta(0) =: U$. That is $V^+(t,x) \leq -c_v\|x\|^2$ with $c_v = a_2 c_w > 0$. □

7.4.4 *Norm* $\|x\|^2 = x^T Nx$

In this section we restrict ourselves to the class of norms $\|x\|^2 = x^T Nx$, where $N \in \mathbb{R}^{n \times n}$ is a symmetric, positive definite matrix.

In the following Theorem 7.14 we consider a nonlinear system and give a sufficient condition for the exponential stability of the zero solution. The construction of the Lyapunov function is similar to the periodic-time case. Note that the assumptions of Theorem 7.14 are sufficient, but not necessary for the exponential attraction of the zero solution, see Theorem 7.15, number 2.

Theorem 7.14. *Consider*

$$\dot{x} = f(t,x), \tag{7.25}$$

where $f \in C^1([0,T] \times \mathbb{R}^n, \mathbb{R}^n)$, $f(t,0) = 0$ *for all* $t \in [0,T]$ *over the finite-time interval* $\mathbb{I} = [0,T]$. *Define* $A(t) := D_x f(t,0)$. *Let* L *be defined as in Theorem 7.8.*

Let the norm $\|\cdot\|$ *be defined by*

$$\|x\|^2 = x^T Nx,$$

where $N \in \mathbb{R}^{n \times n}$ *is a symmetric, positive definite matrix.*

If the Hermitian matrix $L^*N + NL$ is Hurwitz, then the zero solution of (7.25) is exponentially stable. In this case,

$$V(t,x) := x^T(P^{-1}(t))^*NP^{-1}(t)x \text{ and } W(t,x) = \sqrt{V(t,x)}$$

are finite-time local Lyapunov function, satisfying

$$\begin{aligned} a_1\|x\|^2 \le V(t,x) \le b_1\|x\|^2, \qquad & V'(t,x) \le -c_1\|x\|^2, \\ a_2\|x\| \le W(t,x) \le b_2\|x\|, \qquad & W'(t,x) \le -c_2\|x\|, \end{aligned}$$

for all $x \in U \setminus \{0\}$ and $t \in \mathbb{I} \setminus \{T\}$, where U is an open neighbourhood of 0, with constants $a_1, b_1, c_1, a_2, b_2, c_2 > 0$.

Proof. Using Theorem 7.8, we express the fundamental matrix solution with initial condition $\Phi(0) = I$ by

$$\Phi(t) = P(t)e^{tL}$$

where $P(0) = P(T) = I$ and $P(t), L \in \mathbb{C}^{n\times n}$. The Hermitian matrix $(L^*N + NL)$ is negative definite. Denote the maximal eigenvalue by $-\nu < 0$, which gives us

$$z^T(L^*N + NL)z \le -\nu\|z\|_2^2 \tag{7.26}$$

for all $z \in \mathbb{C}^n$.

We define the functions V and W as in the theorem. The inequalities for $V(t,x)$ and $W(t,x)$ follow from the fact that $P^{-1}(t)$ is non-singular and N is positive definite. As $P(0) = P(T) = I$, we have $V(0,x) = V(T,x) = \|x\|^2$ and $W(0,x) = W(T,x) = \|x\|$.

Now we show the inequality for $V'(t,x)$. We use $\dot{P}^{-1}(t) = LP^{-1}(t) - P^{-1}(t)A(t)$, which is shown as in the periodic case (see (7.7) in the proof of Theorem 7.6) and $(A(t))^* = (A(t))^T$ since $A(t) \in \mathbb{R}^{n\times n}$. Furthermore, by Taylor $f(t,x) = A(t)x + \psi(x)$, where for $\epsilon := \frac{\nu}{2\|N\|_2}$ there is $r > 0$ such that $\frac{\|P^{-1}(t)\psi(t,x)\|_2}{\|P^{-1}(t)x\|_2} < \epsilon$ for all $t \in \mathbb{I}$ and $\|x\|_2 < r$. Hence, we obtain

$$\begin{aligned} V'(t,x) &= x^T(A(t))^T(P^{-1}(t))^*NP^{-1}(t)x + (\psi(t,x))^*(P^{-1}(t))^*NP^{-1}(t)x \\ &\quad + x^T(\dot{P}^{-1}(t))^*NP^{-1}(t)x + x^T(P^{-1}(t))^*N\dot{P}^{-1}(t)x \\ &\quad + x^T(P^{-1}(t))^*NP^{-1}(t)A(t)x + x^T(P^{-1}(t))^*NP^{-1}(t)\psi(t,x) \\ &= x^T[(A(t))^T(P^{-1}(t))^*N + (P^{-1}(t))^*L^*N - (A(t))^*(P^{-1}(t))^*N]P^{-1}(t)x \\ &\quad + x^T(P^{-1}(t))^*[NLP^{-1}(t) - NP^{-1}(t)A(t) + NP^{-1}(t)A(t)]x \\ &\quad + 2\|P^{-1}(t)x\|_2\|N\|_2\|P^{-1}(t)\psi(t,x)\|_2 \end{aligned}$$

$$
\begin{aligned}
&= ((P^{-1}(t))x)^*[L^*N + NL]P^{-1}(t)x + \frac{\nu}{2}\|P^{-1}(t)x\|_2^2 \\
&\leq -\frac{\nu}{2}\|P^{-1}(t)x\|_2^2 \\
&\leq -c_1\|x\|^2
\end{aligned}
$$

for a suitable $c_1 > 0$ and for all $t \in \mathbb{I}$ and $0 < \|x\|_2 < r$ by (7.26). The proof for W follows as in Theorem 7.6. □

7.5 Relations Between Autonomous, Periodic and Finite-Time Systems

7.5.1 Periodic Systems as Finite-Time Systems

If we consider a time-periodic system

$$\dot{x} = f(t, x),$$

then we can also regard this system as a finite-time system. We discuss the stability of the zero solution with respect to the different notions.

Theorem 7.15. *Consider a T-periodic system $\dot{x} = f(t,x)$ with $f \in C^1(\mathbb{R} \times \mathbb{R}^n, \mathbb{R}^n)$, $f(t,0) = 0$ for all $t \in \mathbb{R}$. We can also consider the system as a finite-time system over the interval $\mathbb{I} = [0,T]$.*

If the zero solution is exponentially stable with respect to the finite-time case, then it is exponentially stable with respect to the periodic-time case.

By Theorems 7.4 and 7.8 there is a matrix $L \in \mathbb{C}^{n\times n}$ such that the principal solution of the linearised equation with $\Phi(0) = I$ can be expressed as $\Phi(t) = P(t)e^{tL}$, where $P(0) = P(T) = I$. Now the following statements hold true for L:

1. *Let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^n$. If $\|e^{TL}\| < 1$, then L is Hurwitz.*
2. *Let $N \in \mathbb{R}^{n\times n}$ be a symmetric, positive definite matrix and let $\|\cdot\|$ be the induced matrix norm corresponding to the vector norm $\|x\|^2 = x^T Nx$.*
   If $L^*N + NL$ is Hurwitz, then $\|e^{TL}\| < 1$ for all $T > 0$.

Proof. Assume that the zero solution is exponentially stable with respect to the finite-time case. Then, by Theorem 7.9 we have $\|DF_T(0)\| = \mu \in (0,1)$.

Since $F_T(x) = DF_T(0)x + \psi(x)$ with a function $\psi(x) = o(\|x\|)$ as $x \to 0$, there exists $\eta > 0$ such that $\|\psi(x)\| \leq \frac{1-\mu}{2}\|x\|$ for all $x \in B_\eta(0)$. Thus,

$$\|F_T(x)\| \leq \|DF_T(0)\|\|x\| + \|\psi(x)\| \leq \mu\|x\| + \frac{1-\mu}{2}\|x\| = \frac{1+\mu}{2}\|x\|.$$

Denoting $\nu := \frac{1+\mu}{2} \in (0,1)$, since $\mu \in (0,1)$, we have now

$$\|F_T(x)\| \leq \nu \|x\|$$

for all $x \in B_\eta(x)$. This is also the Poincaré map $P\colon \mathbb{R}^n \to \mathbb{R}^n$, $x \to \varphi(T,0,x)$ of the periodic system and thus

$$\|P^k x\| \leq \nu^k \|x\|$$

for all $k \in \mathbb{N}$, which implies the exponential stability with respect to the periodic system as $\nu \in (0,1)$. Part 2 is a direct consequence of Theorems 7.14, 7.7 and 7.9. □

7.5.2 *Autonomous Systems as Periodic and Finite-Time Systems*

An autonomous system

$$\dot{x} = f(x)$$

can be considered as a periodic system with any period T, and also on a finite-time interval $[0,T]$ with any $T > 0$. We discuss the relations between the different notions of attractivity for such a system. We start with an example, and we later prove a general theorem.

Example 7.1. Consider the linear system with $f(x) = Ax$, where $A := \begin{pmatrix} -1 & c \\ 0 & -1 \end{pmatrix}$ with $c \in \mathbb{R}$ and $x \in \mathbb{R}^2$. Then the principal solution can be written as $\Phi(t) = e^{tL}$, where in the previous notation we have $P(t) = I$ and $L = \begin{pmatrix} -1 & c \\ 0 & -1 \end{pmatrix}$. The eigenvalues of L are -1 and have negative real part. Thus, as an autonomous example, the origin is exponentially asymptotically stable for all $c \in \mathbb{R}$ and so is the zero solution if we regard it as a periodic system.

Now we consider the system as a finite-time system on the interval $\mathbb{I} = [0,T]$ with the Euclidean norm $\|x\|^2 = x^T x$, i.e. $N = I$. In this case, $\|e^{TL}\|^2$ is given by the maximal eigenvalue of $e^{TL^*} e^{TL}$. We have

$$e^{TL} = e^{-T} \begin{pmatrix} 1 & Tc \\ 0 & 1 \end{pmatrix}$$

$$\text{and thus } e^{TL^*} e^{TL} = e^{-2T} \begin{pmatrix} 1 & Tc \\ Tc & T^2c^2 + 1 \end{pmatrix}.$$

The eigenvalues are

$$\lambda_{1,2} = e^{-2T}\left(1 + \frac{T^2c^2}{2} \pm \sqrt{T^2c^2 + \frac{T^4c^4}{4}}\right).$$

Both eigenvalues are < 1 if and only if

$$|c| < 2\frac{\sinh T}{T} =: c^*(T).$$

Note that $\lim_{T\to 0} c^*(T) = 2$ and $\lim_{T\to\infty} c^*(T) = \infty$. Hence, depending on the finite-time interval $[0, T]$ under consideration, the zero solution is exponentially asymptotically stable, if and only if $|c| < c^*(T)$. As $T \to \infty$, the zero solution is exponentially attractive for all c.

Now we consider the condition that $L^*N + NL$ is Hurwitz of Theorem 7.14, which is sufficient for the exponential attractivity. In this example, $L^*N + NL = L^* + L = \begin{pmatrix} -2 & c \\ c & -2 \end{pmatrix}$. The eigenvalues are $\mu_{1,2} = -2 \pm c$, and they are both negative if $|c| < 2$. Hence, the condition that the zero solution is finite-time attractive **for all** $T > 0$ ($|c| \le 2$) is nearly equivalent to $L^*N + NL$ being Hurwitz ($|c| < 2$).

We can prove the following general lemma.

Lemma 7.4. *Consider the autonomous system*

$$\dot{x} = f(x) \tag{7.27}$$

with $f \in C^1(\mathbb{R}^n, \mathbb{R}^n)$, $f(0) = 0$ and $Df(0) =: A \in \mathbb{R}^{n\times n}$.

The zero solution of this T-periodic system for any $T > 0$ is exponentially stable if and only if the equilibrium 0 is exponentially stable for the autonomous system.

If the zero solution is finite-time exponentially attractive for a $T > 0$, then it is exponentially stable for both the autonomous and periodic system.

Now consider the norm $\|x\|^2 = x^T Nx$ with symmetric positive definite matrix $N \in \mathbb{R}^{n\times n}$. Then we have the following implications

$$(i) \Rightarrow (ii) \Leftrightarrow (iii) \Rightarrow (iv).$$

(i) All eigenvalues λ of $A^TN + NA$ satisfy $\lambda < 0$.
(ii) The zero solution of (7.27) is exponentially stable over the finite-time interval $[0, T]$ for all $T > 0$.
(iii) For all $T > 0$, $\|e^{TA}\| < 1$.
(iv) All eigenvalues λ of $A^TN + NA$ satisfy $\lambda \le 0$.

Proof. The first part follows from the fact that $L = A$ and Theorems 7.3 and 7.7 as well as Theorem 7.15.

For the second part, note that (i)⇒(ii) follows from Theorem 7.14 and (ii)⇔(iii) follows from Theorem 7.9. It is left to show (iii)⇒(iv).

Let us assume that for all $T > 0$, we have $\|e^{TA}\| < 1$ and, in contradiction to the statement, that there is an eigenvalue $\lambda > 0$ and $v \in \mathbb{R}^n \setminus \{0\}$ such that

$$(A^T N + NA)v = \lambda v.$$

Since $\|e^{TA}\| < 1$, we have

$$\begin{aligned}\|v\|^2 &> v^T e^{TA^T} N e^{TA} v \\ &= v^T \left(I + TA^T + \frac{1}{2}T^2(A^T)^2 + \dots\right) N \left(I + TA + \frac{1}{2}T^2A^2 + \dots\right) v \\ &= v^T \left(N + T(A^T N + NA) + \varphi(T)\right) v \\ &= \|v\|^2 + \|v\|_2^2(T\lambda + \varphi(T))\end{aligned}$$

where $\varphi(T) = o(T)$ as $T \to 0$. Hence, there is a $T > 0$ such that $|\varphi(T)| \le T\lambda/2$ and $0 > \|v\|_2^2 T\lambda/2 > 0$ due to $\lambda > 0$, which is a contradiction. □

7.6 Conclusions and Outlook

In this chapter, we have generalised the construction of local Lyapunov functions for general nonlinear systems to periodic-time and finite-time systems. As in the classical autonomous case, we have constructed two types of Lyapunov functions V and W, satisfying

$$\begin{aligned} a_1\|x\|^2 \le V(t,x) \le b_1\|x\|^2, \qquad & V'(t,x) \le -c_1\|x\|^2, \\ a_2\|x\| \le W(t,x) \le b_2\|x\|, \qquad & W'(t,x) \le -c_2\|x\|. \end{aligned}$$

They are global Lyapunov functions for linear systems, and local Lyapunov functions for nonlinear ones.

Although we give explicit formulas for V and W, we are using the Floquet representation of solutions, so that in explicit examples their calculation requires the solution of the first variation equation. The practical use of the results, besides the theoretical existence, is to derive an algorithm for the construction of CPA Lyapunov functions for periodic and finite-time systems, where the results of this paper will be important to close the gap between the local and the global part of the Lyapunov

function. We envisage that, as in the autonomous case [11, 13], where we have used similar results to show the existence and to algorithmically construct a CPA Lyapunov function, we can use the results of this paper for similar algorithms in the time-periodic and finite-time cases.

References

1. Baier, R., Grüne, L., Hafstein, S.: Linear programming based Lyapunov function computation for differential inclusions. Discrete Contin. Dyn. Syst. Ser. B **17**, 33–56 (2012)
2. Berger, A., Doan, T.S., Siegmund, S.: Nonautonomous finite-time dynamics. Discrete Contin. Dyn. Syst. Ser. B **9**, 463–492 (2008)
3. Chicone, C.: Ordinary Differential Equations with Applications. Springer, New York (1999)
4. Culver, W.J.: On the existence and uniqueness of the real logarithm of a matrix. Proc. Am Math. Soc. **17**, 1146–1151 (1966)
5. Doan, T.S., Karrasch, D., Ngyuen, T.Y., Siegmund, S.: A unified approach to finite-time hyperbolicity which extends finite-time Lyapunov exponents. J. Differ. Equ. **252**, 5535–5554 (2012)
6. Doan, T.S., Palmer, K., Siegmund, S.: Transient spectral theory, stable and unstable cones and Gershgorin's theorem for finite-time differential equations. J. Differ. Equ. **250**, 4177–4199 (2011)
7. Giesl, P.: Construction of Global Lyapunov Functions Using Radial Basis Functions. Lecture Notes in Mathematics, vol. 1904. Springer, Berlin (2007)
8. Giesl, P.: Construction of a local and global Lyapunov function using radial basis functions. IMA J. Appl. Math. **73**, 782–802 (2008)
9. Giesl, P.: Construction of a finite-time Lyapunov function by Meshless Collocation. Discrete Contin. Dyn. Syst. Ser. B **17**, 2387–2412 (2012)
10. Giesl, P., Hafstein, S.: Existence of piecewise affine Lyapunov functions in two dimensions. J. Math. Anal. Appl. **371**, 233–248 (2010)
11. Giesl, P., Hafstein, S.: Construction of Lyapunov functions for nonlinear planar systems by linear programming. J. Math. Anal. Appl. **388**, 463–479 (2012)
12. Giesl, P., Hafstein, S.: Existence of piecewise linear Lyapunov functions in arbitrary dimensions. Discrete Contin. Dyn. Syst. **32**, 3539–3565 (2012)
13. Giesl, P., Hafstein, S.: Revised CPA method to compute Lyapunov functions for nonlinear systems (2013, submitted)
14. Giesl, P., Rasmussen, M.: Areas of attraction for nonautonomous differential equations on finite time intervals. J. Math. Anal. Appl. **390**, 27–46 (2012)
15. Grüne, L., Kloeden, P.E., Siegmund, S., Wirth, F.: Lyapunov's second method for nonautonomous differential equations. Discrete Contin. Dyn. Syst. **18**, 375–403 (2007)
16. Hafstein, S.: A constructive converse Lyapunov theorem on exponential stability. Discrete Contin. Dyn. Syst. **10**, 657–678 (2004)
17. Hafstein, S.: An algorithm for constructing Lyapunov functions. Electron. J. Differ. Equ. Monogr. **8**, 100 (2007)
18. Haller, G.: Finding finite-time invariant manifolds in two-dimensional velocity fields. Chaos **10**, 99–108 (2000)
19. Haller, G., Sapsis, T.: Lagrangian coherent structures and the smallest finite-time Lyapunov exponent. Chaos **21**, 1–5 (2011)
20. Haller, G., Yuan, G.: Lagrangian coherent structures and mixing in two-dimensional turbulance. Physica D **147**, 352–370 (2000)
21. Khalil, H.K.: Nonlinear Systems, 3rd edn. Macmillan, New York (2000)

22. Lyapunov, A.M.: The General Problem of the Stability of Motion. Translated by A. T. Fuller, Taylor & Francis (1992)
23. Papachristodoulou, A., Prajna, S.: The construction of Lyapunov functions using the sum of squares decomposition. In: 41th IEEE Conference on Decision and Control, pp. 3482–3487 (2002)
24. Peet, M.: Exponentially stable nonlinear systems have polynomial Lyapunov functions on bounded regions. IEEE Trans. Autom. Control **54**, 979–987 (2009)
25. Peet, M., Papachristodoulou, A.: A converse sum-of-squares Lyapunov result: an existence proof based on the Picard iteration. In: 49th IEEE Conference on Decision and Control, pp. 5949–5954 (2010)
26. Rasmussen, M.: Finite-time attractivity and bifurcation for nonautonomous differential equations. Differ. Equ. Dyn. Syst. **18**, 57–78 (2010)

Chapter 8
Quasi-Steady State: Searching for and Utilizing Small Parameters

Alexandra Goeke and Sebastian Walcher

Dedicated to Jürgen Scheurle on the occasion of his 60th birthday

Abstract We present an outline of quasi-steady state methods (QSS) in ordinary differential equations which model systems of chemical reactions, and its application to reduction of dimension. Special attention is given to the relation between QSS and singular perturbations including, as a new result, a general explicit reduction formula. Moreover, we describe and discuss heuristics which convert a QSS assumption to conditions restricting the parameters of the differential equation.

2010 Mathematics Subject Classification: 92C45, 80A30, 34D05, 34D23.

8.1 Introduction

Quasi-steady state (QSS) reduction is frequently employed to reduce the dimension of differential equations for chemical and biochemical reactions, in particular as a preliminary step in parameter identification problems. While QSS has been used by biologists, chemists, and also by application-oriented mathematicians since the early twentieth century, a precise mathematical description and analysis was achieved only in the late 1980s, and some aspects are still not completely resolved. The issue is complicated by the fact that different groups of scientists (including different groups of mathematicians) have different notions of, and different approaches to, QSS assumptions and reductions. Another critical point concerns the role, the applicability, and the application of singular perturbation theory.

A. Goeke · S. Walcher (✉)
Lehrstuhl A für Mathematik, RWTH Aachen, 52056 Aachen, Germany
e-mail: alexandra.goeke@matha.rwth-aachen.de; walcher@matha.rwth-aachen.de

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_8,

Most of this paper collects some classical and recent results on QSS and QSS reduction. We also present a few new results and aspects, including a general explicit reduction formula for mass action kinetics, given a singular perturbation setting. In Sect. 8.2 we review definitions of QSS (including a working definition we will adopt), give a short historical outline, and describe some problems and applications, including standard examples. Motivated by applications, it seems advisable to distinguish two notions of QSS (which also appear under different names in the literature). On one hand, there is QSS for reactions, where certain (forward–backward) reactions are assumed to reach equilibrium quickly. On the other hand, there is QSS for concentrations of certain chemical species, which goes back to Michaelis and Menten. Analyzing these different QSS assumptions leads to different mathematical problems. Section 8.3 is about reduction of dimension in the classical Tikhonov–Fenichel setting of singular perturbations. We present a general reduction formula, sketch its derivation, and give several examples. Moreover we show that, in the scenario of slow–fast reactions, Tikhonov–Fenichel theory is applicable in rather general circumstances. Section 8.4 is about various heuristics—including scaling methods—for finding "small parameters" from QSS assumptions. While these heuristics provide satisfactory results in many cases, identification of small parameters for QSS—which is closely tied to the chosen definition—still seems unfinished. Most of the examples we give are presented for the purpose of illustration and have been discussed in other publications. One exception is a somewhat larger example to demonstrate the feasibility of the reduction procedure.

8.2 Background and Statement of Problem

8.2.1 Chemical Reactions and ODEs

Systems of chemical reactions are frequently modeled with the help of differential equations. In this paper we will concentrate on systems that can be modeled by ordinary differential equations, which is justified in the following scenario:

- Reactions take place in a closed vessel, and there is no spatial inhomogeneity.
- Thermodynamical parameters such as temperature and pressure are (being kept) constant.
- There are explicit expressions for the reaction rates (usually mass action kinetics).

Given these conditions, there is a standard procedure to transfer a chemical reaction scheme to a system of ordinary differential equations and there is a number of strong theoretical results on the properties of such equations. The procedure was formalized and the class of resulting equations was discussed by several authors in the 1960s and 1970s, with fundamental contributions, in particular with regard to convergence to equilibrium, due to Feinberg [7], and Horn and Jackson [15]. One

important ingredient of this procedure is stoichiometry: Molecules do not vanish into nothing and are not created out of nothing. Thus in a reaction like $A + B \rightleftharpoons C$, for every C-molecule that is created, an A and a B vanish. Hence stoichiometry implies the existence of linear first integrals for the differential equations.

Example 8.1. The Michaelis–Menten system (Michaelis and Menten [20], see also Briggs and Haldane [5] and many textbooks and monographs such as Atkins and de Paula [2]; Berg et al. [3]; Keener and Sneyd [18]; Murray [22]) is a basic model reaction for an enzyme E catalyzing the transformation of substrate S to product P via an intermediate complex C. The reaction scheme

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E + P$$

by way of the above-mentioned procedure with mass action kinetics yields a differential equation system for the concentrations:

$$\begin{aligned}
\dot{e} &= -k_1 es + (k_{-1} + k_2)c - k_{-2} ep,\\
\dot{s} &= -k_1 es + k_{-1} c,\\
\dot{c} &= k_1 es - (k_{-1} + k_2)c + k_{-2} ep,\\
\dot{p} &= k_2 c - k_{-2} ep.
\end{aligned}$$

The relevant initial values are $s(0) = s_0 > 0$, $c(0) = 0$, $e(0) = e_0 > 0$, and $p(0) = 0$. All rate constants k_i are assumed to be > 0, with the possible exception $k_{-2} \geq 0$. In case $k_{-2} = 0$ one speaks of the irreversible Michaelis–Menten reaction, the case $k_{-2} > 0$ is called reversible. The irreversible system is usually presented and investigated in monographs and research articles.

From stoichiometry one obtains the linear first integrals $e+c$ and $s+c+p$, which may be used to reduce the differential equation to dimension two. The standard procedure leads to the following equation:

$$\begin{aligned}
\dot{s} &= -k_1 e_0 s + (k_1 s + k_{-1})c,\\
\dot{c} &= k_1 e_0 s - (k_1 s + k_{-1} + k_2)c + k_{-2}(e_0 - c)(s_0 - s - c).
\end{aligned} \tag{8.1}$$

Example 8.2. A cooperative enzyme-catalyzed reaction is described by the reaction scheme (see e.g. Keener and Sneyd [10, 18, 25]):

$$\begin{aligned}
S + E &\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} C_1 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E + P,\\
S + C_1 &\underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} C_2 \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} C_1 + P.
\end{aligned}$$

Here substrate and enzyme react to form a complex C_1, and moreover substrate and C_1 react to form a complex C_2. In the reversible scenario, enzyme and product may also combine to form C_1 with rate constant $k_{-2} > 0$, and C_1 and P may combine to form C_2 with rate constant $k_{-4} > 0$. Similar to the Michaelis–Menten system one also considers the irreversible case with $k_{-2} = k_{-4} = 0$; all other rate constants are assumed > 0 throughout. From mass action kinetics, stoichiometry, and the initial values $s(0) = s_0 > 0$, $c_1(0) = c_2(0) = 0$, $e(0) = e_0 > 0$, and $p(0) = 0$, one obtains the differential equation

$$\begin{aligned}\dot{s} =& -k_1 e_0 s + (k_{-1} + k_1 s - k_3 s)c_1 + (k_1 s + k_{-3})c_2,\\ \dot{c}_1 =& k_1(e_0 - c_1 - c_2)s - (k_{-1} + k_2)c_1 + k_{-2}(e_0 - c_1 - c_2)(s_0 - s - c_1 - 2c_2)\\ & - k_3 c_1 s + (k_{-3} + k_4)c_2 - k_{-4} c_1(s_0 - s - c_1 - 2c_2),\\ \dot{c}_2 =& k_3 c_1 s - (k_{-3} + k_4)c_2 + k_{-4} c_1(s_0 - s - c_1 - 2c_2).\end{aligned} \tag{8.2}$$

8.2.2 *Quasi-Steady State*

It seems much harder to precisely define QSS, as well as the corresponding QSS assumption, than to illustrate the use of QSS to reduce the dimension of the system. Some authors use a (relatively straightforward) notion of QSS for reactions, which we will consider in Sect. 8.3.3 below. However, the notion of QSS for chemical species, which will be in the focus of this paper, seems more delicate. (The distinction has also been noticed and investigated in detail by Goussis [11]. One also speaks of partial equilibrium instead of QSS for reactions.) It should be emphasized that the choice of a definition for QSS critically influences its translation to mathematical terms, and that various notions exist in the literature. The following characterization (taken from [24]) may be the least common denominator of all definitions:

Working Definition. A reacting system is in *QSS*, or *quasi-stationary*, with respect to certain species, if the rates of change of their concentrations are negligibly small compared to the overall rate of reaction, during some relevant time interval.

A *QSS assumption* amounts to the hypothesis that a reaction is in QSS with respect to certain components.

The source of a QSS assumption generally lies outside mathematics. Usually experimental observations or biological or chemical intuition are invoked. Generally QSS corresponds to restrictions on certain parameters, such as rate constants or initial concentrations.

We give a brief sketch of the history of QSS and mention some contributors to its theory and practice, with no claim to completeness. Michaelis and Menten [20] stated and applied a certain equilibrium assumption, which they did not justify further. Briggs and Haldane [5] seem to be the first who discussed the QSS assumption for the complex C (now sometimes called the *standard QSS*

assumption) in the Michaelis–Menten system (8.1), and moreover they justified this assumption by referring to smallness of certain parameters in the differential equation. Atkins and de Paula's popular introductory text on Physical Chemistry (see [2, p. 812 ff.]) reflects a frequently used notion of QSS in a reacting system: "(. . .) after an initial induction period (. . .), and during the major part of the reaction, the rates of change of concentrations of all reaction intermediates are negligibly small." The biochemistry text by Stryer et al. (see [3]) seems to make direct use of QSS, with no discussion of underlying assumptions. In the contribution by Rubinow and Segel to the collection [31] (see p. 3 ff.), one finds the following description for (irreversible) Michaelis–Menten: under suitable experimental conditions "one expects that after an initial short transient period there will be a balance between the formation of complex by the union of enzyme and substrate and the breaking apart of complex (. . .)." From a mathematical perspective, the (explicit or implicit) involvement of two different time regimes (initial phase vs. major part of the reaction, to paraphrase Atkins et al.) suggests a singular perturbation approach. One of the earliest papers on QSS from the perspective of Tikhonov's theorem is due to Heineken, Tsuchiya, and Aris [12], with "small parameter" e_0/s_0. Segel [29], and Segel and Slemrod [30] performed a careful analysis of QSS and conditions on parameters. These papers seem to be the starting point for time scale arguments in QSS considerations. Among the many who continued and extended this approach, with varying emphasis on mathematical rigor, we mention Ignetik and Deakin [17]; Ignetik et al. [16]; Borghans et al. [4]; Schnell and Maini [28]; and Tzafriri and Edelman [35]. An approach by Schauer and Heinrich [26] to the Michaelis–Menten system, on the other hand, could be seen as emphasizing the slow manifold in a singular perturbation setting, but their reasoning is essentially based on the assumption that the concentration of the complex is almost constant (more precisely, that $\dot{c} \approx 0$).

In Sect. 8.4 we will review some of these arguments and their use in heuristics for finding small parameters.

8.2.3 *The Ad Hoc Reduction from QSS*

The following reduction method (which we call the ad hoc reduction) is directly related to a QSS assumption: In the differential equation, set the negligible rates of change equal to zero, and use the subsequent algebraic relations to obtain a reduced system.

Example 8.3. QSS for the complex C in the Michaelis–Menten system.

In the irreversible case ($k_{-2} = 0$) one has

$$0\,(=\dot{c}) = k_1 e_0 s - (k_1 s + k_{-1} + k_2)c, \quad \text{thus } c = \frac{k_1 e_0 s}{k_1 s + k_{-1} + k_2}.$$

By substitution one obtains the reduced equation

$$\dot{s} = -\frac{k_2 e_0 s}{(s + (k_{-1} + k_2)/k_1)}$$

which can be found in virtually all books and papers which mention Michaelis–Menten. Note that this approach does not explicitly make use of small parameters, although in its justification in the literature (e.g., [4,12,29,30]) small parameters are frequently invoked. We will discuss a different approach in Example 8.5 below.

In the reversible case the ad hoc method leads to a quadratic equation

$$0\,(=\dot{c}) = k_1 e_0 s - (k_1 s + k_{-1} + k_2)c + k_{-2}(e_0 - c)(s_0 - s - c),$$

which yields

$$c = \frac{1}{2k_{-2}}\left(k_1 s + k_{-1} + k_2 + k_{-2}(e_0 + s_0 - s) \pm \sqrt{(\cdots)}\right)$$

and a relatively cumbersome reduced equation, which is not frequently used. (There are discussions, e.g., in Miller and Alberty [21]; Seshadri and Fritsch [32].)

Example 8.4. QSS in the cooperative system.

Consider the system from Eq. (8.2), and assume QSS for both complexes. In the irreversible case ($k_{-2} = k_{-4} = 0$), solving "$\dot{c}_1 = \dot{c}_2 = 0$," which is a linear parameter-dependent system for c_1 and c_2, provides a nice reduced equation for s; see Keener and Sneyd [18]. But the reversible case leads to a system of quadratic equations for c_1 and c_2, which in turn leads to a reduced equation for s which is intractable, for all practical purposes. See [10,25] for more details.

Thus the ad hoc reduction, although conceptually straightforward, may become quite inconvenient even in rather simple settings. And, more fundamentally, there remains the question: How, if at all, can a reduction procedure be justified mathematically?

8.3 Reduction in the Presence of Small Parameters

In this section we consider an analytic ordinary differential equation depending on a "small parameter" $\varepsilon \geq 0$. Thus we have

$$\dot{x} = h(x, \varepsilon) = h^{(0)}(x) + \varepsilon h^{(1)}(x) + \ldots, \quad x \in U \subset \mathbb{R}^{n+m}, \tag{8.3}$$

with both n and m positive integers (to be specified below), and we will be interested in the behavior of the solutions as $\varepsilon \to 0$. Our primary focus is on differential equations modeling chemical reactions, and the small parameter may stem either

from a separation of fast and slow reactions, or (by some yet-to-be-discussed reasoning; see Sect. 8.4) from a QSS assumption. (For the examples introduced in Sect. 8.2, $\varepsilon = e_0$ works.) But once a small parameter is given, the natural starting point is to try singular perturbation theory.

Due to our focus on chemical reactions, we will impose additional conditions on the right-hand side, which go beyond what is necessary from the perspective of singular perturbation theory. Thus we assume that $h^{(0)}$ and $h^{(1)}$ are polynomials. These assumptions are natural in our setting, since we start from polynomial differential equations of mass action kinetics.

We will mostly rely only on the classical results of singular perturbation theory, see Tikhonov [34], Vasil'eva [36], Fenichel [8], and Hoppensteadt [14]. The monograph [37] by Verhulst, in particular Chapter 8, is an appropriate source for most of the relevant material. The principal new result in this section will be an explicit expression for a reduction of (8.3), given the special assumptions on the right-hand side. We obtain a QSS reduction which is both on solid mathematical ground and relatively simple to compute.

8.3.1 Singular Perturbations

The usual scenario for Tikhonov's and Fenichel's theorems starts with a system in what we call *Tikhonov standard form*:

$$\begin{aligned} \dot{y}_1 &= \varepsilon f^{(1)}(y_1, y_2) + \ldots, \quad y_1(0) = y_{1,0}, \\ \dot{y}_2 &= g^{(0)}(y_1, y_2) + \varepsilon g^{(1)}(y_1, y_2) + \ldots, \quad y_2(0) = y_{2,0}, \end{aligned} \tag{8.4}$$

with $(y_1, y_2) \in D \subseteq \mathbb{R}^n \times \mathbb{R}^m$, D open, and (in our setting) analytic right-hand side.

We obtain the system in "slow time" by rescaling $\tau = \varepsilon t$:

$$\begin{aligned} y_1' &= f^{(1)}(y_1, y_2) + \ldots, \quad y_1(0) = y_{1,0}, \\ y_2' &= \varepsilon^{-1} g^{(0)}(y_1, y_2) + g^{(1)}(y_1, y_2) + \ldots, \quad y_2(0) = y_{2,0}. \end{aligned} \tag{8.5}$$

A fundamental result of Tikhonov's theory can be stated as follows. (See Verhulst [37], Theorem 8.1 for a more general theorem under less restrictive hypotheses.)

Theorem 8.1. *Let system (8.5) be given. Assume that:*

(i) The zero set $\tilde{Y}$ of $g^{(0)}$ is nonempty.
(ii) There exist a nonempty relatively open subset $\tilde{M}_0 \subseteq \tilde{Y}$ and $\rho > 0$ such that every eigenvalue of $D_2 g^{(0)}(y_1, y_2)$, with $(y_1, y_2) \in \tilde{M}_0$, has real part $\leq -\rho$.

Then there exists $t_1 > 0$ such that for every $t_0 \in (0, t_1)$ and for every point sufficiently close to $\tilde{M}_0$, the solution of (8.5) with initial condition $(y_{1,0}, y_{2,0})$ approaches the solution of the degenerate system

$$\begin{aligned} y_1' &= f^{(1)}(y_1, y_2, 0), \quad y_1(0) = y_{1,0}\,, \\ 0 &= g^{(0)}(y_1, y_2, 0) \end{aligned}$$

uniformly on $[t_0,\, t_1]$ *as* $\varepsilon \to 0$.

A priori a system (8.3) derived from a chemical reaction network may not be given in Tikhonov standard form, which raises two questions. First, under what conditions does a transformation to standard form exist? Second, assuming the existence of a transformation, how can a reduced equation be computed?

As for existence, one needs a diffeomorphism $\Psi = (\Psi_1,\, \Psi_2)^{\mathrm{tr}}$ (defined on some open subset of D) which sends solutions of (8.3) to solutions of a system (8.4) in standard form. A necessary and sufficient condition is the identity

$$D\Psi(x)\left(h^{(0)}(x) + \varepsilon h^{(1)}(x) + \dots\right) = \begin{pmatrix} \varepsilon \cdot f^{(1)}(x)(\Psi_1(x), \Psi_2(x)) + \dots \\ g^{(0)}(x)(\Psi_1(x), \Psi_2(x)) + \dots \end{pmatrix}.$$

For $\varepsilon = 0$ one obtains

$$D\Psi(x)h^{(0)}(x) = \begin{pmatrix} 0 \\ g^{(0)}(\Psi_1(x), \Psi_2(x)) \end{pmatrix},$$

and this implies the existence of n independent first integrals (viz., the entries of Ψ_1) for $\dot{x} = h^{(0)}(x)$. Recall that the existence of first integrals is not trivial near stationary points. Moreover, $h^{(0)}$ then admits an n-dimensional local manifold M_0 of stationary points. The following result is taken from [25], but it essentially goes back to Fenichel [8].

Proposition 8.1. *Given* $\dot{x} = h(x, \varepsilon)$*, there exists a transformation* Ψ*, defined on some open* $\tilde{U} \subseteq D$*, to Tikhonov standard form with the eigenvalue condition (ii) from Theorem 8.1, if and only if the following hold:*

The zero set Y *of* $h^{(0)}$ *in* $\tilde{U}$ *is nonempty. Moreover there exist a nonempty relatively open* $M_0 \subseteq Y$ *and* $\rho > 0$ *such that for every* $x_0 \in M_0$ *the derivative* $Dh^{(0)}(x_0)$ *admits the eigenvalue* 0 *with algebraic and geometric multiplicity* n*, and the remaining eigenvalues have real part* $\leq -\rho$*. (In particular* M_0 *is a local* n*-dimensional submanifold.)*

The condition given in Proposition 8.1 implies the existence of a direct sum decomposition

$$\mathbb{R}^{n+m} = \operatorname{Ker} Dh^{(0)}(x_0) \oplus \operatorname{Im} Dh^{(0)}(x_0) \tag{8.6}$$

for every $x_0 \in M_0$. Moreover, this condition implies locally the existence of n independent first integrals for $\dot{x} = h^{(0)}(x)$.

8.3.2 Computing a Reduction

First, we discuss the special case when the hypotheses of Proposition 8.1 are satisfied and a transformation Ψ to Tikhonov standard form (as well as its inverse) is explicitly known. Then determining a reduced system in original coordinates is relatively straightforward. Although Ψ cannot be directly applied to the reduced system in the version

$$y_1' = f^{(1)}(y_1, y_2), \quad g^{(0)}(y_1, y_2) = 0,$$

there is an equivalent version on the invariant manifold $\tilde{M}_0$ introduced in Theorem 8.1, viz.

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = p(y) := \begin{pmatrix} f^{(1)}(y_1, y_2) \\ -D_2 g^{(0)}(y_1, y_2)^{-1} D_1 g^{(0)}(y_1, y_2) \cdot f^{(1)}(y_1, y_2) \end{pmatrix}, \quad (8.7)$$

which can be transported back via Ψ, to a differential equation with invariant manifold M_0, see [25].

Example 8.5. The reduced system for reversible Michaelis–Menten.

Here the small parameter is (assumed to be) e_0, and system (8.1) gives the function

$$h^{(0)} = \begin{pmatrix} (k_1 s + k_{-1})c \\ -(k_1 s + k_{-1} + k_2 + k_{-2}(s_0 - s - c))c \end{pmatrix}.$$

The differential equation with right-hand side $h^{(0)}$ is a scaled linear system, and a first integral (and therefore a transformation) can be found explicitly. Carrying out the program outlined above (see [25]), one obtains the reduced equation, in addition to $\dot{c} = 0$,

$$\dot{s} = -e_0 \cdot \frac{s(k_1 k_2 + k_{-1} k_{-2}) - k_{-1} k_{-2} s_0}{k_1 s + k_{-1} + k_2 + k_{-2}(s_0 - s)},$$

which is generally different from the ad hoc-reduced equation, and actually appears less complicated (no square roots). But note that the Tikhonov–Fenichel reduction coincides with the ad hoc reduction when $k_{-2} = 0$.

If an explicit transformation is known, it may provide an additional benefit because Theorem 8.1 in Verhulst [37] characterizes the admissible initial conditions. Moreover, for Michaelis–Menten one can verify Hoppensteadt's [14] criteria for convergence on the interval $[t_0, \infty)$ (notation from Theorem 8.1).

Generally, one cannot hope for an explicit construction of a transformation to Tikhonov standard form, but still it is possible to compute a reduced equation. If the slow manifold M_0 can be explicitly represented as the graph of some function,

Fenichel ([8], Lemma 5.4) and Stiefenhofer ([33], Equation (2.13), with a different and quite short proof) gave expressions for the reduced equation.

One can carry Fenichel's observations further to obtain a reduced system in the general setting. A closer look at Eq. (8.7) shows that one gets $p(y)$ via the kernel-image decomposition of the derivative

$$\begin{pmatrix} 0 & 0 \\ D_1 g^{(0)}(y) & D_2 g^{(0)}(y) \end{pmatrix}$$

(compare Eq. (8.6)) by computing the kernel component of $(f^{(1)}(y), g^{(1)}(y))^{\mathrm{tr}}$. Since the kernel-image decomposition is preserved in coordinate transformations, one obtains (see [25]):

Proposition 8.2. *Given the eigenvalue condition from Proposition 8.1 for transformability to Tikhonov standard form, near a point x_0 with $h^{(0)}(x_0) = 0$, one obtains the reduced system of (8.3) by computing the kernel component of $h^{(1)}(x)$ with respect to the direct kernel-image decomposition of $Dh^{(0)}(x)$.*

As noted in [25], the projection onto the kernel can be found from the minimal polynomial of $Dh^{(0)}(x)$, and for some—relatively small—examples this approach is computationally feasible. For higher dimensions and many parameters, this procedure becomes prohibitively expensive. But in any case the argument shows that for polynomial (or rational) $h^{(0)}$ and $h^{(1)}$ the reduced system will have a rational right-hand side.

A practicable method to compute reduced systems was recently developed in [9]. It is based on an auxiliary result from classical Algebraic Geometry.

Lemma 8.1. *For system (8.3) with polynomials (or rational functions) $h^{(0)}$ and $h^{(1)}$, let x_0 be such that $h^{(0)}(x_0) = 0$ and that the eigenvalue condition from Proposition 8.1 hold for $Dh^{(0)}(x_0)$, with $m = \text{rank } Dh^{(0)}(x_0)$. Then there exist a $(n+m) \times m$ matrix P with rational entries, of rank m, and a vector valued polynomial μ with m entries, such that*

$$h^{(0)}(x) = P(x)\mu(x)$$

in some Zariski neighborhood of x_0. By appropriate choice of the neighborhood, one may assume that $h^{(0)}$ and μ have the same zero sets. The entries of μ may be taken as any m independent entries of $h^{(0)}$.

This Lemma, which is proved in [9], is almost trivial in the local analytic (or differentiable) setting, in view of the Implicit Function Theorem. But the point is that P is rational, and that there are constructive methods to determine P. With this auxiliary result, the reduction is relatively straightforward, as is shown in the next theorem.

Theorem 8.2. *For system (8.3), with assumptions as in Lemma 8.1, let x_0 be such that $h^{(0)}(x_0) = 0$. Then the reduced system, in a Zariski neighborhood M_0 of x_0 in the zero set Y of $h^{(0)}$, is given (in slow time) by*

$$x' = h^{(1)}(x) - P(x)\left(D\mu(x)\,P(x)\right)^{-1} D\mu(x)h^{(1)}(x).$$

Proof. The eigenvalue conditions at x_0 guarantee that Y is locally an n-dimensional manifold. Let M_0 be a relatively open subset of Y such that the eigenvalue conditions hold for all points of M_0. Denote the columns of P by $p_1, \ldots, p_m$. The Jacobian matrix of $h^{(0)}$ equals

$$Dh^{(0)}(x) = \sum_{i=1}^{m} \left(p_i(x)D\mu_i(x) + \mu_i(x)Dp_i(x)\right)$$

in a Zariski neighborhood of x_0, and therefore

$$Dh^{(0)}(x) = P(x) \cdot D\mu(x) \quad \text{for all } x \in M_0. \tag{8.8}$$

Now fix $x \in M_0$. Then

$$\operatorname{Ker} Dh^{(0)}(x) = \operatorname{Ker} D\mu(x), \tag{8.9}$$

due to the rank condition for $P(x)$. The condition $\operatorname{rank} P(x) = m$ also implies

$$\operatorname{Im} Dh^{(0)}(x) = \operatorname{Im} P(x).$$

From our basic hypothesis we have the direct sum decomposition (8.6). Set

$$A(x) := D\mu(x) \cdot P(x).$$

We first show that $A(x)$ is invertible. Thus let $\beta \in \mathbb{R}^m$ be a solution of the equation

$$D\mu(x)P(x)\beta = 0. \tag{8.10}$$

The direct sum decomposition and

$$P(x)\beta \in \operatorname{Ker} Dh^{(0)}(x) \cap \operatorname{Im} Dh^{(0)}(x)$$

show $P(x)\beta = 0$. Since $P(x)$ has full rank, we see $\beta = 0$. Thus Eq. (8.10) admits only the trivial solution, whence $A(x) = D\mu(x)P(x)$ is invertible.

Moreover, due to the direct sum decomposition (8.6), for any $y \in \mathbb{R}^{n+m}$ there exist $z \in \operatorname{Ker} Dh^{(0)}(x) = \operatorname{Ker} D\mu(x)$ and $\alpha \in \mathbb{R}^m$ such that

$$y = z + P(x)\alpha.$$

Since $z = y - P(x)\alpha \in \operatorname{Ker} D\mu(x)$, one finds

$$D\mu(x)\left(y - P(x)\alpha\right) = 0,$$

which implies $\alpha = A(x)^{-1} D\mu(x)y$, and thus

$$z = y - P(x)A(x)^{-1} D\mu(x)y.$$

Apply this to $h^{(1)}(x)$ to obtain the assertion. □

Remark 8.1. (a) It may be appropriate to discuss invertibility of $A(x)$ and the eigenvalue condition in more detail. The zero set Y of $h^{(0)}$ is an algebraic variety in $\mathbb{R}^{n+m}$, and we are actually interested in an n-dimensional component M_0 of this variety. For $x \in Y$, $Dh^{(0)}(x)$ must therefore have eigenvalue 0 with geometric multiplicity n. If the geometric and algebraic multiplicity are equal to n then (and only then) the kernel-image decomposition (8.6) exists, and the latter is equivalent to invertibility of $A(x)$. Thus it is possible to write down the equation in Theorem 8.2. But additional conditions (for instance, all real parts of eigenvalues $\leq -\rho$) are necessary to make this a meaningful reduced system.

(b) The matrix $A(x)$ is of size $m \times m$, hence relatively small. One should also emphasize that inverting this matrix is not actually necessary to determine the reduced system: It suffices to solve a system of linear equations with this matrix.

(c) The eigenvalue condition (ii) from Theorem 8.1 on $Dh^{(0)}(x)$ is satisfied if and only if all eigenvalues of $A(x)$ have real part $\leq -\rho$. Indeed, by virtue of Eq. 8.8 and linear algebra the nonzero eigenvalues of these two matrices are the same.

Example 8.6. Irreversible Michaelis–Menten with slow product formation.
This is an example for a slow–fast reaction separation. One considers the familiar differential equation (8.1), but now with small parameter k_2. The underlying assumption is that product formation is much slower than formation of complex and degrading of complex back to enzyme and substrate. One has (with the arguments s, c suppressed)

$$h^{(0)} = \begin{pmatrix} -\mu \\ \mu \end{pmatrix}, \quad P = \begin{pmatrix} -1 \\ 1 \end{pmatrix}, \quad \mu = k_1 e_0 s - (k_1 s + k_{-1})c, \quad h^{(1)} = \begin{pmatrix} 0 \\ -c \end{pmatrix}.$$

Thus

$$A = (k_1(e_0 - c),\, -(k_1 s + k_{-1})) \begin{pmatrix} -1 \\ 1 \end{pmatrix} = -\left(k_1(e_0 - c) + k_1 s + k_{-1}\right)$$

and the reduction procedure yields the system

$$\begin{pmatrix} \dot{s} \\ \dot{c} \end{pmatrix} = -\frac{k_2}{k_1(e_0 - c) + k_1 s + k_{-1}} \begin{pmatrix} (k_1 s + k_{-1})c \\ k_1(e_0 - c)c \end{pmatrix}$$

on the invariant variety M_0^* defined by $\mu = 0$. Using the parametrization of M_0 one may obtain a reduced equation for substrate alone, viz.

$$\dot{s} = -\frac{k_2 k_1 e_0 s\,(k_1 s + k_{-1})}{k_1 k_{-1} e_0 + (k_1 s + k_{-1})^2}$$

This is different from the standard reduction based on QSS for complex. This example illustrates that different QSS assumptions (QSS for the species C, resp. QSS with slow product formation) will lead to different reductions.

Example 8.7. The cooperative system with small parameter e_0 (see Example 8.2 above) was originally discussed in [25], with the minimal polynomial approach (and serious reliance on a computer algebra system). Theorem 8.2 makes this computation feasible even by hand. With the abbreviations

$$\alpha = -(k_1 + k_3)s - (k_{-1} + k_2) - (k_{-2} + k_{-4})(s_0 - s - c_1 - 2c_2)$$
$$\beta = -k_1 s + k_{-3} + k_4 - k_{-2}(s_0 - s - c_1 - 2c_2)$$

one has

$$h^{(0)} = \begin{pmatrix} (k_{-1} + k_1 s - k_3 s)c_1 + (k_1 s + k_{-3})c_2 \\ c_1\alpha + c_2\beta \\ k_3 c_1 s - (k_{-3} + k_4)c_2 + k_{-4}c_1(s_0 - s - c_1 - 2c_2) \end{pmatrix},$$

$$h^{(1)} = \begin{pmatrix} -k_1 s \\ k_1 s + k_{-2}(s_0 - s - c_1 - 2c_2) \\ 0 \end{pmatrix},$$

and the (relevant component of the) zero set of $h^{(0)}$ is given by $c_1 = c_2 = 0$. A decomposition according to Lemma 8.1 is given by

$$P = \begin{pmatrix} k_{-1} + k_1 s - k_3 s & k_1 s + k_{-3} \\ \alpha & \beta \\ k_3 s + k_{-4}(s_0 - s - c_1 - 2c_2) & -k_{-3} - k_4 \end{pmatrix}$$

and

$$\mu = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}.$$

Note that such a decomposition is not unique. For $A = D\mu \cdot P$ one obtains

$$A(s,\, 0,\, 0) = \begin{pmatrix} a(s) & -k_1 s + k_{-3} + k_4 - k_{-2}(s_0 - s) \\ k_3 s + k_{-4}(s_0 - s) & -k_{-3} - k_4 \end{pmatrix}$$

with abbreviation $a(s) = -(k_1 + k_3)s - (k_{-1} + k_2) - (k_2 + k_{-4})(s_0 - s)$. The matrix A can easily be inverted, and the eigenvalue condition is readily checked. The final result for the reduced equation is, of course, identical to the one given in [25].

Example 8.8. To illustrate the feasibility of the approach via Theorem 8.2, we discuss a somewhat bigger example, from Stiefenhofer [33, Section 3], whose reduction was computed in [33] only for some special parameter choices. This system models communication between slime mold cells such as *Dictyostelium discoideum*. Communication is effected by $cAMP$, denoted by P, furthermore S stands for the substrate ATP, while D and R represent transmembrane receptors, with $\overline{D}$ and $\overline{R}$ representing the corresponding bound states. (See [33] for more details.) The reaction scheme, including production and decomposition of $cAMP$ with constant rates, can be written as follows.

$$\begin{array}{lll} D \underset{\varepsilon k_{-1}}{\overset{\varepsilon k_1}{\rightleftharpoons}} R & S + \overline{R} \overset{\varepsilon k_2}{\rightharpoonup} P + \overline{R} & \overset{\varepsilon k_3}{\rightharpoonup} P \overset{\varepsilon k_{-3}}{\rightharpoonup} \\ \overline{D} \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} D + 2P & \overline{R} \underset{k_{-5}}{\overset{k_5}{\rightleftharpoons}} R + 2P, & \end{array}$$

with parameters $k_i > 0$, and the ε's indicating the slow reactions. We also adopt the further simplification $s(t) = S > 0$ from [33].

We use the first integral $\overline{d} = c - d - r - \overline{r}$, with some constant $c \geq 0$, to obtain the system

$$\begin{aligned} \dot{p} &= -2k_4 dp^2 + 2k_{-4}(c - d - r - \overline{r}) - 2k_5 rp^2 + 2k_{-5}\overline{r} + \varepsilon k_3 - \varepsilon k_{-3} p + \varepsilon k_2 S\overline{r}, \\ \dot{d} &= -k_4 dp^2 + k_{-4}(c - d - r - \overline{r}) - \varepsilon k_1 d + \varepsilon k_{-1} r, \\ \dot{r} &= -k_5 rp^2 + k_{-5}\overline{r} + \varepsilon k_1 d - \varepsilon k_{-1} r, \\ \dot{\overline{r}} &= k_5 rp^2 - k_{-5}\overline{r}. \end{aligned} \tag{8.11}$$

With

$$\mu(p, d, r, \overline{r}) := \begin{pmatrix} -k_5 rp^2 + k_{-5}\overline{r} \\ -k_4 dp^2 + k_{-4}(c - d - r - \overline{r}) \end{pmatrix}, \; P(p, d, r, \overline{r}) := \begin{pmatrix} 2 & 2 \\ 0 & 1 \\ 1 & 0 \\ -1 & 0 \end{pmatrix},$$

and

$$h^{(1)}(p, d, r, \overline{r}) := \begin{pmatrix} k_3 - k_{-3} p + k_2 S\overline{r} \\ -k_1 d + k_{-1} r \\ k_1 d - k_{-1} r \\ 0 \end{pmatrix},$$

Equation (8.11) may be written as

$$\frac{d}{dt}\begin{pmatrix} p \\ d \\ r \\ \overline{r} \end{pmatrix} = P \cdot \mu + \varepsilon h^{(1)},$$

which is a representation according to Lemma 8.1. The variety Y may be taken as the zero set of μ, and one verifies that the choice

$$M_0 := Y \cap \operatorname{int} \mathbb{R}^4_+,$$

is possible. Straightforward computations show that the eigenvalue condition is at least generically satisfied, and the reduced system on the invariant set M_0 is given by

$$\frac{d}{dt}\begin{pmatrix} p \\ d \\ r \\ \overline{r} \end{pmatrix} = \frac{1}{Q}\begin{pmatrix} N_1 \\ N_2 \\ N_3 \\ N_4 \end{pmatrix}$$

on a suitable subset $W \subset \mathbb{R}^2$, which is determined from $M_0 \cap \mathbb{R}^4_+$ by this elimination, with

$$\begin{aligned}
N_1 :=\ & -k_{-3}p^5k_5k_4 + (k_2S\overline{r}\,k_5k_4 + k_3k_5k_4)\,p^4 - k_{-3}\,(k_{-5}k_4 + k_{-4}k_5)\,p^3 \\
& +\big(2\,k_1\,(-k_{-4}k_5 + k_{-5}k_4)\,d - 2\,k_{-1}\,(-k_{-4}k_5 + k_{-5}k_4)\,r \\
& +Sk_2\,(k_{-5}k_4 + k_{-4}k_5)\,\overline{r} + k_3\,(k_{-5}k_4 + k_{-4}k_5)\,\big)p^2 - k_{-3}pk_{-5}k_{-4} + k_3k_{-5}k_{-4} \\
& +k_2S\overline{r}\,k_{-5}k_{-4}, \\
N_2 :=\ & 2\,k_4dp^4k_5k_{-3} + (-2\,k_3k_5k_4 - 2\,k_2S\overline{r}\,k_5k_4)\,dp^3 + \big(\,(2\,k_4k_{-5}k_{-3} - k_5k_{-4}k_1)\,d \\
& +k_5k_{-4}k_{-1}r\big)p^2 + \big(\,-4\,k_1d^2k_4k_{-5} + \big(\,(-4\,k_5k_{-4}k_1 + 4\,k_{-1}k_4k_{-5})\,r \\
& -2\,k_4k_{-5}k_2S\overline{r} - 2\,k_4k_{-5}k_3\big)d + 4\,k_{-1}r^2k_{-4}k_5\big)p - k_1dk_{-5}k_{-4} + k_{-1}rk_{-5}k_{-4}, \\
N_3 :=\ & 2\,k_5rp^4k_4k_{-3} + (-2\,k_3k_5k_4 - 2\,k_2S\overline{r}\,k_5k_4)\,rp^3 + \big(k_{-5}k_4k_1d \\
& +(2\,k_5k_{-4}k_{-3} - k_{-1}k_4k_{-5})\,r\big)p^2 + \big(4\,k_1d^2k_4k_{-5} \\
& +(-4\,k_{-1}k_4k_{-5} + 4\,k_5k_{-4}k_1)\,rd - 4\,k_{-1}r^2k_{-4}k_5 + \big(\,-2\,k_5k_{-4}k_3 \\
& -2\,k_5k_{-4}k_2S\overline{r}\big)r\big)p + k_1dk_{-5}k_{-4} - k_{-1}rk_{-5}k_{-4}, \\
N_4 :=\ & (k_5k_4k_1d - k_5k_4\,(k_{-1} + 2\,k_{-3})\,r)\,p^4 + \big(4\,k_5k_4k_1d^2 + 4\,k_5k_4\,(-k_{-1} + k_1)\,rd \\
& -4\,k_5k_{-1}k_4r^2 + (2\,k_2S\overline{r}\,k_5k_4 + 2\,k_3k_5k_4)\,r\big)p^3 + \big(k_5k_{-4}k_1d - k_5k_{-4}\big(k_{-1} \\
& +2\,k_{-3}\big)r\big)p^2 + (2\,k_5k_{-4}k_2S\overline{r} + 2\,k_5k_{-4}k_3)\,rp
\end{aligned}$$

and

$$\begin{aligned}
Q :=\ & k_5p^4k_4 + (4\,k_4k_5r + 4\,k_4dk_5)\,p^3 + (k_{-5}k_4 + k_{-4}k_5)\,p^2 \\
& + (4\,k_5rk_{-4} + 4\,k_4dk_{-5})\,p + k_{-5}k_{-4}.
\end{aligned}$$

One may eliminate the variables r and $\overline{r}$ via

$$M_0 = \left\{(p,d,r,\overline{r})^{\mathrm{tr}}; r = \frac{k_{-5}(-k_4dp^2 + k_{-4}(c-d))}{k_{-4}(k_{-5} + k_5p^2)}, \overline{r} = \frac{k_5p^2(-k_4dp^2 + k_{-4}(c-d))}{k_{-4}(k_{-5} + k_5p^2)}\right\},$$

and obtain the equivalent version

$$\frac{d}{dt}\begin{pmatrix} p \\ d \end{pmatrix} = \frac{1}{\tilde{Q}}\begin{pmatrix} \tilde{N}_1 \\ \tilde{N}_2 \end{pmatrix}$$

with

$$\begin{aligned}\tilde{N}_1 :=\ & -p^{10}Sdk_5{}^3k_4{}^2k_2 - p^9k_5{}^3k_4\,k_{-4}\,k_{-3} + p^8(cStk_5{}^3k_4\,k_{-4}\,k_2\\ & -2\,Sdk_5{}^3k_4\,k_{-4}\,k_2 - 2\,Sdk_5{}^2k_{-5}\,k_4{}^2k_2k_5{}^3k_4\,k_{-4}\,k_3) + p^7(-k_5{}^3k_{-4}{}^2k_{-3}\\ & -3\,k_5{}^2k_{-5}\,k_4\,k_{-4}\,k_{-3}) + p^6(cSk_5{}^3k_{-4}{}^2k_2 + 2\,cSk_5{}^2k_{-5}\,k_4, k_{-4}\,k_2 - Sdk_5{}^3k_{-4}{}^2k_2\\ & -4\,Sdk_5{}^2k_{-5}\,k_4\,k_{-4}\,k_2 - Sdk_5\,k_{-5}{}^2k_4{}^2k_23\,k_5{}^2k_{-5}\,k_4\,k_{-4}\,k_3 - 2\,dk_5{}^3k_{-4}{}^2k_1\\ & +2\,dk_5{}^2k_{-5}\,k_4\,k_{-4}\,k_1 - 2\,dk_5{}^2k_{-5}\,k_4\,k_{-4}\,k_{-1} + 2\,dk_5\,k_{-5}{}^2k_4{}^2k_{-1} + k_5{}^3k_{-4}{}^2k_3)\\ & +p^5(-3\,k_5{}^2k_{-5}\,k_{-4}{}^2k_{-3} - 3\,k_5\,k_{-5}{}^2k_4\,k_{-4}\,k_{-3}) + p^4(2\,cSk_5{}^2k_{-5}\,k_{-4}{}^2k_2\\ & +cSk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_2 - 2\,Sdk_5{}^2k_{-5}\,k_{-4}{}^2k_2 - 2\,Sdk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_2\\ & +2\,ck_5{}^2k_{-5}\,k_{-4}{}^2k_{-1} - 2\,ck_5\,k_{-5}{}^2k_4\,k_{-4}\,k_{-1} - 4\,dk_5{}^2k_{-5}\,k_{-4}{}^2k_1\\ & -2\,dk_5{}^2k_{-5}\,k_{-4}{}^2k_{-1} + 4\,dk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_1 + 2\,dk_{-5}{}^3k_4{}^2k_{-1}\\ & +3\,k_5\,k_{-5}{}^2k_4\,k_{-4}\,k_3) + p^3(-3\,k_5\,k_{-5}{}^2k_{-4}{}^2k_{-3} - k_{-5}{}^3k_4\,k_{-4}\,k_{-3})\\ & +p^2(cSk_5\,k_{-5}{}^2k_{-4}{}^2k_2 + 3\,k_5{}^2k_{-5}\,k_{-4}{}^2k_3 - Sdk_5\,k_{-5}{}^2k_{-4}{}^2k_2\\ & +2\,ck_5\,k_{-5}{}^2k_{-4}{}^2k_{-1} - 2\,ck_{-5}{}^3k_4\,k_{-4}\,k_{-1} - 2\,dk_5\,k_{-5}{}^2k_{-4}{}^2k_1\\ & -2\,dk_5\,k_{-5}{}^2k_{-4}{}^2k_{-1} + 2\,dk_{-5}{}^3k_4\,k_{-4}\,k_1 + 2\,dk_{-5}{}^3k_4\,k_{-4}\,k_{-1} + 3\,k_5\,k_{-5}{}^2k_{-4}{}^2k_3\\ & +k_{-5}{}^3k_4\,k_{-4}\,k_3) - pk_{-5}{}^3k_{-4}{}^2k_{-3} + k_{-5}{}^3k_{-4}{}^2k_3,\end{aligned}$$

$$\begin{aligned}\tilde{N}_2 :=\ & p^9(2\,Sd^2k_5{}^3k_4{}^2k_2) + p^8(2\,dk_5{}^3k_4\,k_{-4}\,k_{-3}) + p^7(-2\,cSdk_5{}^3k_4\,k_{-4}\,k_2\\ & +2\,Sd^2k_5{}^3k_4\,k_{-4}\,k_2 + 4\,Sd^2k_5{}^2k_{-5}\,k_4{}^2k_2 - 2\,dk_5{}^3k_4\,k_{-4}\,k_3) + p^6(-dk_5{}^3k_{-4}{}^2k_1\\ & +6\,dk_5{}^2k_{-5}\,k_4\,k_{-4}\,k_{-3} - dk_5{}^2k_{-5}\,k_4\,k_{-4}\,k_{-1}) + p^5(-6\,dk_5{}^2k_{-5}\,k_4\,k_{-4}\,k_3\\ & -4\,cSdk_5{}^2k_{-5}\,k_4\,k_{-4}\,k_2 + 4\,Sd^2k_5{}^2k_{-5}\,k_4\,k_{-4}\,k_2 + 2\,Sd^2k_5\,k_{-5}{}^2k_4{}^2k_2)\\ & +p^4(+ck_5{}^2k_{-5}\,k_{-4}{}^2k_{-1} - 3\,dk_5{}^2k_{-5}\,k_{-4}{}^2k_1 - dk_5{}^2k_{-5}\,k_{-4}{}^2k_{-1}\\ & +6\,dk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_{-3} - 2\,dk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_{-1}) + p^3(+4\,d^2k_5{}^2k_{-5}\,k_{-4}{}^2k_1\\ & -4\,d^2k_5\,k_{-5}{}^2k_4\,k_{-4}\,k_1 + 4\,d^2k_5\,k_{-5}{}^2k_4\,k_{-4}\,k_{-1} - 4\,d^2k_{-5}{}^3k_4{}^2k_{-1}\\ & -6\,dk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_3 - 2\,cSdk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_2 + 2\,Sd^2k_5\,k_{-5}{}^2k_4\,k_{-4}\,k_2\\ & -4\,cdk_5^2k_{-5}\,k_{-4}{}^2k_1 - 4\,cdk_5\,k_{-5}{}^2k_4\,k_{-4}\,k_{-1}) + p^2(2\,ck_5\,k_{-5}{}^2k_{-4}{}^2k_{-1}\\ & -3\,dk_5\,k_{-5}{}^2k_{-4}{}^2k_1 - 2\,dk_5\,k_{-5}{}^2k_{-4}{}^2k_{-1} + 2\,dk_{-5}{}^3k_4\,k_{-4}\,k_{-3}\\ & -dk_{-5}{}^3k_4\,k_{-4}\,k_{-1}) + p(4\,c^2k_5\,k_{-5}{}^2k_{-4}{}^2k_{-1} - 4\,cdk_5\,k_{-5}{}^2k_{-4}{}^2k_1\\ & -8\,cdk_5\,k_{-5}{}^2k_{-4}{}^2k_{-1} + 4\,cdk_{-5}{}^3k_4\,k_{-4}\,k_{-1} + 4\,d^2k_5\,k_{-5}{}^2k_{-4}{}^2k_1\\ & +4\,d^2k_5\,k_{-5}{}^2k_{-4}{}^2k_{-1} - 4\,d^2k_{-5}{}^3k_4\,k_{-4}\,k_1 - 4\,d^2k_{-5}{}^3k_4\,k_{-4}\,k_{-1}\\ & -2\,dk_{-5}{}^3k_4\,k_{-4}\,k_3) + ck_{-5}{}^3k_{-4}{}^2k_{-1} - dk_{-5}{}^3k_{-4}{}^2k_1 - dk_{-5}{}^3k_{-4}{}^2k_{-1},\end{aligned}$$

$$\begin{aligned}\tilde{Q} :=\ & \big(p^6k_5{}^2k_4\,k_{-4} + p^5(4\,dk_5{}^2k_4\,k_{-4} - 4\,dk_5\,k_{-5}\,k_4{}^2) + p^4(k_5{}^2k_{-4}{}^2 + 2\,k_5\,k_{-5}\,k_4\,k_{-4})\\ & +p^3(4\,ck_5\,k_{-5}\,k_4\,k_{-4}) + p^2(2\,k_5\,k_{-5}\,k_{-4}{}^2 + k_{-5}{}^2k_4\,k_{-4}) + p(4\,ck_5\,k_{-5}\,k_{-4}{}^2\\ & -4\,dk_5\,k_{-5}\,k_{-4}{}^2 + 4\,dk_{-5}{}^2k_4\,k_{-4}) + k_{-5}{}^2k_{-4}{}^2\big)\big(p^2k_5 + k_{-5}\big).\end{aligned}$$

Up to scaling of time, one therefore obtains a two-dimensional equation with polynomial right-hand side (of degree 11) on W. ($\tilde{Q} > 0$ on W follows from $Q > 0$ on $\mathbb{R}^4_+$.) In particular one has Poincaré–Bendixson theory available in the asymptotic limit.

8.3.3 Slow and Fast Reactions

For slow and fast reactions the reduction program via Tikhonov–Fenichel was carried out by Schauer and Heinrich [27] (who cited Vasil'eva [36] for singular perturbation results), and continued by Stiefenhofer [33]. (A recent paper by Lee and Othmer [19] reproduces several of these results.)

Before discussing slow and fast reactions in some detail, it will be necessary to give a more precise outline of the work by Feinberg [7], Horn and Jackson [15], and others. We mostly follow Feinberg's Lecture Notes [7]; a short overview can be found in Section 2 of Anderson's recent paper [1]. We sketch the formalism for chemical reaction networks and reaction systems with mass action kinetics, using the irreversible Michaelis–Menten system as an illustrating example. One starts with an ordered collection of q chemical *species*, which are identified with the standard basis of $\mathbb{R}^q$. Next one forms *complexes*, which formally speaking are nonnegative (integer) linear combinations of species (appearing as reactants or as reaction products). Then *reactions* are defined as ordered pairs of complexes, usually written in a notation with reaction arrows. (The notion of reversible reaction is obvious). For Michaelis–Menten the species are E, S, C, and P, which will in the following be identified with the standard basis vectors of $\mathbb{R}^4$. Moreover one has complexes $E + S$, C, $E + P$, and reactions

$$E + S \to C; \quad C \to E + S; \quad C \to E + P.$$

Using the identification of species and standard basis vectors, one assigns to each reaction a vector in $\mathbb{R}^q$, counting the reactants with negative sign, calls their span in $\mathbb{R}^q$ the *stoichiometric subspace* S, and collects these column vectors in a matrix Z which is related to the stoichiometric matrix as defined by Feinberg. For the Michaelis–Menten example one has, in the above order, column vectors

$$\begin{pmatrix} -1 \\ -1 \\ 1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 1 \\ -1 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 1 \\ 0 \\ -1 \\ 1 \end{pmatrix}, \text{ thus } \; Z = \begin{pmatrix} -1 & 1 & 1 \\ -1 & 1 & 0 \\ 1 & -1 & -1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The differential equation for the concentrations may now be written in the form

$$\frac{d}{dt} \begin{pmatrix} e \\ s \\ c \\ p \end{pmatrix} = \begin{pmatrix} -1 & 1 & 1 \\ -1 & 1 & 0 \\ 1 & -1 & -1 \\ 0 & 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} k_1 es \\ k_{-1} c \\ k_2 c \end{pmatrix}$$

and generally for mass action kinetics one obtains a differential equation system of the form

$$\dot{x} = Z \cdot \Phi(x) \tag{8.12}$$

for the vector of concentrations. Φ can be characterized more precisely; see [7] for details.

There is a natural assignment of a directed graph to a reaction network: The nodes are the complexes, and there exists a directed edge from one complex to

another if and only if there is a reaction from the former to the latter. The connected components of this graph are called *linkage classes*. The *deficiency* of a network is then defined by

$$\delta := \#\,\text{complexes} - \#\,\text{linkage classes} - \operatorname{rank} Z,$$

and one can show that $\delta \geq 0$. Finally, one calls the reaction network *weakly reversible* if, whenever there is a reaction from one complex to another, there is also a chain of reactions from the latter to the former.

Both $\mathbb{R}^q_+$ and its interior are positively invariant for the system (8.12). The cosets $x^* + S$, with S the stoichiometric subspace and $0 \neq x^* \in \mathbb{R}^q_+$, are called the *stoichiometric compatibility classes*, and are positively invariant sets for the differential equation. Now we can state one fundamental result of the theory.

Deficiency Zero Theorem (Horn and Jackson [15], Feinberg [7]). *Assume that* (8.12) *corresponds to a weakly reversible deficiency zero network. Then the following hold.*

(i) The intersection of every stoichiometric compatibility class with $\operatorname{int}\mathbb{R}^q_+$ *contains exactly one stationary point.*
(ii) This point is locally asymptotically stable within its compatibility class.

Remark 8.2. (a) The proof of part (ii) is based on an ingenious choice of a Lyapunov function. Linear asymptotic stability cannot be deduced from the inequalities in this argument.

(b) In Feinberg's Notes [7] a stronger claim is made, viz. global asymptotic stability within the intersection of the stoichiometric compatibility class and $\operatorname{int}\mathbb{R}^q_+$. Later a problem in the global stability argument was pointed out; and generally global stability is still a conjecture. Only recently Anderson [1] succeeded with a proof in the case of a single linkage class.

Now we turn to slow–fast systems of chemical reactions. These are usually described by

$$\dot{x} = h^{(0)}(x) + \varepsilon h^{(1)}(x), \quad x \in \mathbb{R}^q \tag{8.13}$$

with the fast subsystem $h^{(0)}$ (large rate constants) and the slow subsystem $\varepsilon h^{(1)}$ (small rate constants, symbolized by the factor ε). Thus both $h^{(0)}$ and $h^{(1)}$ admit a representation of the form (8.12). A substantial part of the following result is due to Schauer and Heinrich [27]. The transfer from Schauer and Heinrich's condition to weakly reversible deficiency zero systems in the statement and proof of the following Proposition is a first step, in ongoing work [9], toward a more comprehensive theorem.

Proposition 8.3. *Assume that the fast subsystem of* (8.13) *has deficiency zero and is weakly reversible. Assume moreover that every stationary point in* $\operatorname{int}\mathbb{R}^q_+$ *is linearly asymptotically stable for* $h^{(0)}$ *within its stoichiometric compatibility class. Then the following hold.*

(a) The eigenvalue conditions from Proposition 8.1 are satisfied for $Dh^{(0)}(x_0)$ at all zeros $x_0 \in \operatorname{int} \mathbb{R}^q_+$ of $h^{(0)}$.
(b) There exists a linear transformation of the system to Tikhonov standard form.

Proof. We write

$$h^{(0)}(x) = Z \cdot \Phi(x)$$

and let $x_0 \in \operatorname{int} \mathbb{R}^q_+$ be a stationary point. Let s be the dimension of the stoichiometric subspace, thus $\operatorname{rank} Z = s$. Then

$$Dh^{(0)}(x_0) = Z\, D\Phi(x_0)$$

admits the eigenvalue 0 with multiplicity $\geq q - s$ (due to the rank of Z) and has s eigenvalues with negative real part, due to the linear stability requirement. In particular

$$\operatorname{rank} Z = \operatorname{rank} Dh^{(0)}(x_0). \tag{8.14}$$

Therefore the zero set of $h^{(0)}$ is locally a manifold of dimension s. Moreover there are independent linear forms $\lambda_1, \ldots, \lambda_{q-s}$ such that $\lambda_i \circ Z = 0$, $1 \leq i \leq q - s$, and these are also first integrals of $h^{(0)}$. Completing these by suitable linear forms to a basis of the dual of $\mathbb{R}^q$ will produce the desired transformation to Tikhonov standard form. □

Remark 8.3. The importance of the rank condition (8.14) for the existence of a linear transformation to standard form was first noted by Schauer and Heinrich [27]. They also stated (with only a partial justification for some special cases, it seems; see [27, Section 4]) that the rank condition holds when every fast reaction is reversible with fast reverse reaction. It seems that linear stability conditions did not play a role in [27].

8.3.4 Why Does the Ad Hoc Method Persist?

As noted earlier, the ad hoc reduction produces the same result as Tikhonov–Fenichel in some relevant cases, but not in general. In [10] we provide a detailed investigation for several basic reaction schemes in biochemistry (including Michaelis–Menten), with the result that ad hoc and Tikhonov–Fenichel reduction coincide when certain product-forming reactions are irreversible, but differ in the fully reversible case. Since such reductions are actually used to determine rate constants and other reaction parameters, it is very likely that serious discrepancies between ad hoc reduction and reality would have been noticed in experimental verification. To explain this apparent lack of serious discrepancy, we note two possible reasons for good approximation by the irreversible reduced system.

First, some reversible reactions may be almost irreversible (for instance, k_{-2} may be very small in the Michaelis–Menten example). Since the reduced system in the irreversible case is the limit of the reversible case, the discrepancy may be hardly noticeable.

Second, continuous removal of product may be responsible, as noted in Heinrich and Schuster [13], Keener and Sneyd [18]. A thorough justification of this argument was also given in [10] for reversible Michaelis–Menten with product removal (rate $\alpha > 0$). Indeed, a Tikhonov–Fenichel reduction of the system

$$\begin{aligned}\dot{s} &= -k_1 e_0 + (k_1 s + k_{-1})c,\\ \dot{c} &= k_1 e_0 s - (k_1 s + k_{-1} + k_2)c + k_{-2}(e_0 - c)p,\\ \dot{p} &= k_2 c - k_{-2} p(e_0 - c) - \alpha p\end{aligned}$$

with "small parameter" e_0 yields the familiar "irreversible reduced equation" (with α vanishing along the way). See Examples 8.3 and 8.5 with $k_{-2} = 0$.

8.4 Finding Small Parameters

While the results in the previous section are based on a well-defined mathematical scenario, there is another facet of QSS which, in the present stage, is not so amenable to rigorous mathematics. The underlying problem is that the translation of a model assumption to mathematical terms is rarely straightforward, and it may depend on seemingly small details. Here we are concerned with translating certain assumptions on chemical reacting systems—in particular QSS assumptions—to mathematical terms.

8.4.1 Underlying Assumptions: QSS vs. Slow–Fast

Frequently QSS assumptions—directly or indirectly—amount to slow–fast hypotheses, and we briefly review some of these.

A direct slow–fast assumption (small and large rate constants) underlies the discussion of slow and fast reactions, as in Eq. (8.13). As noticed above, this is different from a QSS assumption for chemical species, which we discuss now. An indirect slow–fast assumption for species (based on the fact that the linearization of a system (8.4) in Tikhonov standard form necessarily has some very small eigenvalues near the slow manifold) works by seeking conditions to ensure a very small ratio of absolutely smallest to absolutely largest eigenvalue of the linearization (near some submanifold). This was used, for instance, by Duchêne and Rouchon [6], but the method frequently has to rely on numerical calculations, and general parameter conditions seem to be hard to derive. A different indirect slow–fast assumption

proposed by Segel [29], Segel and Slemrod [30] starts from the observation that in singular perturbation scenarios there is a fast initial phase (a "boundary layer in time"; see Verhulst [37]), followed by a slower time regime. From time scale estimates for the initial and the later phase, and the requirement that their ratio should be very small, Segel and Slemrod [30] obtain conditions on the parameters in the Michaelis–Menten system. The approach by Schauer and Heinrich [26] may also be justified by singular perturbation arguments, but the line of reasoning is concerned not with time scales but rather with the presumed slow manifold, and derives parameter conditions from requiring closeness of a solution trajectory to this manifold.

Generally, all these approaches are (at least partly) of heuristic nature, and validity of QSS will have to be checked a posteriori. A potentially erroneous conclusion from the time scale comparisons in Segel and Slemrod [30] for so-called reverse QSS (QSS for substrate) is discussed in [10, Section 4]. Moreover it is easy to construct examples which satisfy the condition proposed by Schauer and Heinrich [26] but do not satisfy any initial phase requirement (as stated in Atkins and de Paula [2]; see quote in Sect. 8.2): consider systems with a first integral. We emphasize that, while QSS hypotheses frequently lead to singularly perturbed systems (with the benefit of a solid reduction theory), this does not seem to be the case in every relevant scenario. Again, much depends on the exact notion of QSS that is used.

8.4.2 The Role of Scaling

In the context of this chapter, a scaling transformation for an ordinary differential equation consists of multiplying the independent variable (time) and each dependent variable by positive numbers. In most mathematically oriented texts and research papers (see in particular Murray [22], Segel and Slemrod [30], Heineken et al. [12]) scaling is used, and frequently employed to find "small parameters." While there is no doubt that scaling is highly relevant for an appropriate analysis of differential equations modeling a real-life situation, in particular for concrete estimates, there may be some danger in the "lumping together" of several model parameters into one "small parameter" for asymptotic arguments.

We will briefly discuss the necessity, benefits, and limitations of scaling for irreversible Michaelis–Menten and the "small parameter" $\varepsilon^* = \frac{e_0}{s_0}$ from Heineken et al. [12]. (For the Segel–Slemrod "small parameter" $\varepsilon = \frac{k_1 e_0}{s_0 + k_{-1} + k_2}$—see [30]—similar remarks apply.) Note that ε^* tends to zero when $e_0 \to 0$, and this case has been resolved above in a satisfactory manner. But ε^* also tends to zero when $s_0 \to \infty$, and to properly analyze the latter case one should keep in mind that the relevant domain for the Michaelis–Menten system (8.1) is defined by the inequalities $0 \leq c \leq e_0$ and $0 \leq s + c \leq s_0$. Hence $s_0 \to \infty$ will blow up the region of interest. Since Tikhonov's theory applies to differential equations on fixed domains, scaling is necessary here. We scale (following Heineken et al. [12] in part, but not completely) by setting $\sigma = s/s_0$ and $\gamma = c/e_0$, and $\varepsilon = 1/s_0$, obtaining the system

$$\begin{aligned}\dot{\sigma} &= -k_1 e_0 \sigma(1-\gamma) + \varepsilon \cdot e_0 k_{-1}\gamma \\ \dot{\gamma} &= \varepsilon^{-1} k_1 \sigma\,(1-\gamma) - (k_{-1}+k_2)\gamma\end{aligned} \tag{8.15}$$

on the domain defined, e.g., by the inequalities $0 \leq \sigma + e_0\gamma \leq 1$, $0 \leq \gamma \leq 1$.

With the usual notation we have

$$h^{(0)} = \begin{pmatrix} 0 \\ k_1\sigma(1-\gamma) \end{pmatrix}, \quad h^{(1)} = \begin{pmatrix} -k_1 e_0 \sigma(1-\gamma) \\ -(k_{-1}+k_2)\gamma \end{pmatrix}.$$

The zero set M_0 of $h^{(0)}$ has two components; the one defined by $\gamma = 1$ corresponds to the standard QSS assumption. The conditions for Tikhonov–Fenichel are satisfied, and a straightforward computation shows that the reduced equation is given by

$$\dot{\sigma} = 0, \quad \gamma = 1.$$

In other words, Tikhonov–Fenichel applies but it yields a degenerate reduced system. Including higher-order terms in ε (thus passing to a O'Malley-Vasil'eva expansion, see Verhulst [37]) one formally obtains the familiar reduced equation. The approach in [12, Equations (10) to (13)] encounters the same problem in the case $s_0 \to \infty$, because some of the scaled parameters approach zero in this limiting case. Taking this into account, the lowest-order reduction in [12] will also be trivial.

The point we want to emphasize here is the necessity to consider all possible ways in which a "small parameter" may approach zero. This also may be of some practical relevance, since $e_0 \to 0$ ("very little enzyme") and $s_0 \to \infty$ ("very high substrate concentration") represent different experimental settings. These cases require individual consideration, with one case not amenable to standard singular perturbation methods. However, other lines of reasoning, such as the phase plane arguments in [23], show that a QSS assumption is indeed justified for this scenario.

Finally, we note that the other component of M_0 for Eq. (8.15) is given by $\sigma = 0$, which would correspond to the reverse QSS assumption (with s approaching its equilibrium 0 very fast). In this case the Tikhonov–Fenichel reduction formalism is not applicable, due to a nilpotent Jacobian. (One may question whether reverse QSS is chemically sensible for very high s_0.)

8.4.3 Near-Invariance Heuristics

In [24] a proposal was made to generalize Schauer and Heinrich's [26] heuristics from Michaelis–Menten to general systems. We will present the heuristics here in a somewhat informal manner, referring for details to [24]. Thus we start with a parameter-dependent differential equation

$$\dot{x} = q(x,\, p) \tag{8.16}$$

with $q : U \times V \to \mathbb{R}^n$, where $U \subset \mathbb{R}^n$ is a neighborhood of a compact set K^* and V is some subset of $\mathbb{R}^d$ ($d \geq 1$).

(i) We assume that certain functions $\psi_1(x,p), \ldots, \psi_r(x,p)$, defined for all $x \in K^*$, are in QSS (according to the Working Definition) for some relevant solution. Thus their rates of change, given by the Lie derivatives

$$\phi_i(x,\, p) := L_q(\psi_i)(x,\, p) = \langle \operatorname{grad} \psi_i(x,\, p),\, q(x,\, p) \rangle\,, \quad 1 \leq i \leq r,$$

are small along this solution. (All derivatives are to be understood with respect to x only.) In applications, the ψ_i are frequently coordinate functions.

(ii) For $p \in V$ let

$$K = K_p = \{ y \in K^* : \phi_1(y) = \cdots = \phi_r(y) = 0 \}\,.$$

Then, due to continuity and compactness arguments, the maximum of the terms $|\phi_1(x)|, \ldots, |\phi_r(x)|$ tends to zero if and only if $\operatorname{dist}(x,\, K)$ tends to zero. Thus requiring QSS with higher and higher accuracy, one obtains invariance of the set K for the differential equation in the limiting case. This is one motivation for the following definition.

(iii) *Near-invariance* (see [24]): Let K^*, $\phi_1, \ldots, \phi_r$ and K be as above, let the ϕ_i be sufficiently differentiable, and assume that the rank of $(D\phi_1, \ldots, D\phi_r)$ on K is equal to r. Given $0 \leq \delta \leq 1$ we say that K is δ-nearly invariant for $\dot{x} = q(x,\, p)$ with respect to $\phi_1, \ldots, \phi_r$ if for all $x \in K$ and $1 \leq j \leq r$ one has the inequality

$$|\, \langle \operatorname{grad} \phi_j(x,\, p),\, q(x,\, p) \rangle\, | \leq \delta \cdot \|\operatorname{grad} \phi_j(x,\, p)\| \cdot \|q(x,\, p)\|.$$

The inequality always holds for $\delta = 1$, due to Cauchy–Schwarz, and for $\delta = 0$ the condition implies invariance of K. Thus one may expect solutions to stay close to K when $\delta \ll 1$.

It should be emphasized that this is another heuristic approach, replacing slow–fast heuristics by "invariant set-heuristics." Also, the notion does not only depend on the set K but also on the defining functions.

(iv) Some properties of near-invariance.

- Locally the desired property from (ii) is a consequence of near-invariance (see [24]): Let K be δ-nearly invariant. Then locally in time ($|t| \leq \rho$), solutions starting on K remain $(C \cdot \delta)$-close to K, with C and ρ independent of the starting point and of δ.
- In the limiting case $\delta \to 0$ one obtains an invariant set. Since one has a parameter-dependent system, and K may change with parameters, one has to take care that no degeneracies occur, as in scaling procedures, and one should avoid blowing up K^*.
- The notion is compatible (up to an error of order ε) with Tikhonov–Fenichel: While the asymptotic slow manifold M_0 in the singular

perturbation setting is not necessarily $(C \cdot \varepsilon)$-nearly invariant, by an order ε correction in the defining equations one will obtain order ε near-invariance (see [24]).

(v) Use in practice: Let a parameter-dependent system (8.16) be given on K^*. Let $\phi_1, \cdots, \phi_r$ define a desired or suspected nearly invariant set K. The near-invariance property cannot be expected to hold generally, but only for certain parameter combinations. Thus evaluation of the near-invariance condition in (iii) will produce (necessary) conditions for the parameters. Determine (or estimate)

$$\delta(p) := \max\left\{ \frac{|\,\langle (\operatorname{grad}\phi_j)(x,\,p),\, q(x,\,p)\rangle\,|}{\|\operatorname{grad}\phi_j(x,\,p)\| \cdot \|q(x,\,p)\|};\quad x \in K,\quad 1 \le j \le r \right\}$$

as a function of the parameters. Requiring $\delta(p)$ to be small provides conditions on the parameter set p. Asymptotic conditions are obtained from the limiting case $\delta(p) \to 0$.
Again we emphasize that further analysis and verification is necessary.

Example 8.9. Reversible Michaelis–Menten.
Consider the reversible Michaelis–Menten reaction (8.1), which we restate as

$$\begin{aligned} \dot{s} &= -\ \phi(s,c)\ - k_2 c + k_{-2}(e_0 - c)(s_0 - s - c), \\ \dot{c} &= \quad \phi(s,c), \end{aligned} \tag{8.17}$$

(the right-hand side will be called q) with QSS for complex $\psi(s,\,c) = c$, and its Lie derivative

$$\phi(s,c) = k_1 e_0 s - (k_1 s + k_{-1} + k_2)c + k_{-2}(e_0 - c)(s_0 - s - c)$$

on the set $K^* \subseteq \mathbb{R}^2$ given by $0 \le c \le e_0^*$, $s \ge 0$, and $s + c \le s_0^*$. (Here e_0^* and s_0^* are upper bounds for the initial concentrations.)
This system was discussed in detail in [24], with attention to the range for which QSS is assumed to hold. (For instance, requiring QSS only when sufficient substrate is still present would amount to a different choice of K^*.) Here we focus on QSS for the whole course of the reaction (after some initial phase), and look at the asymptotic scenario. To obtain QSS conditions for $\psi = c$, evaluate

$$L_q(\phi)(s,\,c) = ((k_1 - k_{-2})(e_0 - c),\, *) \begin{pmatrix} -k_2 c + k_{-2}(e_0 - c)(s_0 - s - c) \\ 0 \end{pmatrix}$$

for $(s,\,c) \in K$ (taking into account $\phi = 0$). One has

$$\frac{|L_q(\phi)(s,\,c)|}{||q(s,\,c)||} = |k_1 - k_{-2}| \cdot (e_0 - c),$$

and a rough estimate yields

$$\frac{|L_q(\phi)(s,\, c)|}{||q(s,\, c)|| \cdot ||\mathrm{grad}\, \phi(s,c)||} \leq \frac{|k_1 - k_{-2}| \cdot e_0}{k_{-1} + k_2} =: \delta^*$$

for all $(s,\, c) \in K$.

The particular case $k_1 = k_{-2}$ actually yields an invariant set (regardless of other parameters), as noted by Miller and Alberty [21]. For the irreversible case $k_{-2} = 0$, the expression for δ^* is equal to the one introduced by Seshadri and Fritzsch [32]; compare the discussion in [23].

It may be appropriate to clarify what has actually been gained. By design of the procedure, one is assured of an invariant set in the limiting case $e_0 \to 0$. This may be taken as a motivation for choosing the small parameter e_0 in the reversible Michaelis–Menten differential equation, which we did throughout this paper. One then verifies that the hypotheses for Tikhonov–Fenichel are satisfied, and one obtains a reduced system with a mathematically solid foundation. Finally (see [10]) one can check a posteriori that QSS does indeed hold for complex under the assumption of small e_0. Thus the circle closes.

Near-invariance heuristics like all the proposed heuristics leading from a QSS assumption (to a precisely stated QSS assumption) to finding small parameters should be seen as work in progress, but there seems to be more potential in this particular approach. One advantage is that the implementation of the procedure (see (v) above) is in principle straightforward.

References

1. Anderson, D.F.: A proof of the global attractor conjecture in the single linkage class case. SIAM J. Appl. Math. **71**(4), 1487–1508 (2011)
2. Atkins, P., de Paula, J.: Physical Chemistry, 8th edn. Oxford University Press, Oxford (2006)
3. Berg, J.M., Tymovzko, J.L., Stryer, L.: Biochemistry: International Edition, 6th edn. Palgrave Macmillan, New York (2006)
4. Borghans, J.A.M., de Boer, R.J., Segel, L.A.: Extending the quasi-steady state approximation by changing variables. Bull. Math. Biol. **58**, 43–63 (1996)
5. Briggs, G.E., Haldane, J.B.S.: A note on the kinetics of enzyme action. Biochem. J. **19**, 338–339 (1925)
6. Duchêne, P., Rouchon, P.: Kinetic scheme reduction via geometric singular perturbation techniques. Chem. Eng. Sci. **51**, 4461–4472 (1996)
7. Feinberg, M.: Lectures on chemical reaction networks. Lecture Notes. URL: http://www.che.eng.ohio-state.edu/feinberg/LecturesOnReactionNetworks/ (1979)
8. Fenichel, N.: Geometric singular perturbation theory for ordinary differential equations. J. Differ. Equ. **31**, 53–98 (1979)
9. Goeke, A.: Reduktion und asymptotische Reduktion von Differentialgleichungen für chemische Reaktionen. Doctoral Thesis. RWTH Aachen (2013, in preparation)
10. Goeke, A., Schilli, C., Walcher, S., Zerz, E.: Computing quasi-steady state reductions. J. Math. Chem. **50**, 1495–1513 (2012)

11. Goussis, D.A.: Quasi steady state and partial equilibrium approximations: their relation and their validity. Combust. Theory Model. **16**, 869–926 (2012)
12. Heineken, F.G., Tsuchiya, H.M., Aris, R.: On the mathematical status of the pseudo-steady state hypothesis of biochemical kinetics. Math. Biosci. **1**, 95–113 (1967)
13. Heinrich, R., Schuster, S.: The Regulation of Cellular Systems. Chapman and Hall, New York (1996)
14. Hoppensteadt, F.: Stability in systems with parameter. J. Math. Anal. Appl. **18**, 129–134 (1967)
15. Horn, F., Jackson, R.: General mass action kinetics. Arch. Ration. Mech. Anal. **47**, 81–116 (1972)
16. Ignetik, R., Deakin, M.A.B., Fandry, C.: Phase plane analyses of the Michaelis–Menten reaction equations. Bull. Math. Biol. **43**, 361–370 (1981)
17. Ignetik, R., Deakin, M.A.B.: Asymptotic analysis of the Michaelis–Menten reaction equations. Bull. Math. Biol. **43**, 375–388 (1981)
18. Keener, J., Sneyd, J.: Mathematical Physiology. I: Cellular Physiology. Springer, New York (2009)
19. Lee, C.H., Othmer, H.G.: A multi-scale analysis of chemical reaction networks: I. Deterministic systems. J. Math. Biol. **60**, 387–450 (2010)
20. Michaelis, L., Menten, M.L.: Die Kinetik der Invertinwirkung. Biochem. Z. **49**, 333–369 (1913)
21. Miller, W.G., Alberty, R.A.: Kinetics of the reversible Michaelis–Menten mechanism and the applicability of the steady state approximation. J. Am. Chem. Soc. **80**, 5146–5151 (1958)
22. Murray, J.D.: Mathematical Biology. I. An Introduction, 3rd edn. Springer, New York (2002)
23. Noethen, L., Walcher, S.: Quasi-steady state in the Michaelis–Menten system. Nonlinear Anal. Real World Appl. **8**, 1512–1535 (2007)
24. Noethen, L., Walcher, S.: Quasi-steady state and nearly invariant sets. SIAM J. Appl. Math. **70**, 1341–1363 (2009)
25. Noethen, L., Walcher, S.: Tihkonov's theorem and quasi-steady state. Discrete Contin. Dyn. Syst. Ser. B **16**(3), 945–961 (2011)
26. Schauer, M., Heinrich, R.: Analysis of the quasi-steady-state approximation for an enzymatic one-substrate reaction. J. Theor. Biol. **79**, 425–442 (1979)
27. Schauer, M., Heinrich, R.: Quasi-steady-state approximation in the mathematical modeling of biochemical reaction networks. Math. Biosci. **65**, 155–170 (1983)
28. Schnell, S., Maini, C.: Enzyme kinetics at high enzyme concentration. Bull. Math. Biol. **62**, 483–499 (2000)
29. Segel, L.A.: On the validity of the steady state assumption of enzyme kinetics. Bull. Math. Biol. **50**, 579–593 (1988)
30. Segel, L.A., Slemrod, M.: The quasi-steady-state assumption: a case study in perturbation. SIAM Rev. **31**, 446–477 (1989)
31. Segel, L.A. (ed.): Biological Kinetics. Cambridge University Press, Cambridge (1991)
32. Seshadri, M., Fritzsch, G.: Analytical solutions of a simple enzyme kinetic problem by a perturbative procedure. Biophys. Struct. Mech. **6**, 111–123 (1980)
33. Stiefenhofer, R.: Quasi-steady-state approximation for chemical reaction networks. J. Math. Biol. **36**, 593–609 (1998)
34. Tikhonov, A.N.: Systems of differential equations containing a small parameter multiplying the derivative. Mat. Sb. **31**(73), 575–586 (1952, in Russian).
35. Tzafriri, A.R., Edelman, E.R.: The total quasi-steady-state approximation is valid for reversible enzyme kinetics. J. Theoret. Biol. **226**(3), 303–313 (2004)
36. Vasil'eva, A.B.: Asymptotic behavior of solutions to certain problems involving nonlinear differential equations containing a small parameter multiplying the highest derivatives. Russ. Math. Surveys **18**, 13–84 (1963)
37. Verhulst, F.: Methods and Applications of Singular Perturbations. Springer, Berlin (2005)

Chapter 9
On a Global Uniform Pullback Attractor of a Class of PDEs with Degenerate Diffusion and Chemotaxis in One Dimension

Messoud Efendiev and Anna Zhigun

Abstract In this chapter, we deal with a class of nonautonomous degenerate parabolic systems that encompasses two different effects: porous medium and chemotaxis. Such classes of equations arise in the mesoscale level modeling of biomass spreading mechanisms via chemotaxis. Under certain "balance" conditions on the order of the porous medium degeneracy and the growth of the chemotactic function, we establish the existence of a strong uniform pull back attractor for the case of one spatial dimension, thus improving our previous study, where a weak attractor was constructed.

Keywords Attractor • Biofilms • Chemotaxis • Dissipative estimate • Nonautonomous equation • Porous-medium

2010 Mathematics Subject Classification: 35A01, 35A02, 35B41, 35B45, 35D30, 35K65

M. Efendiev (✉)
Helmholtz Center Munich, Institute of Computational Biology, 85764 Neuherberg, Germany
e-mail: messoud.efendiyev@helmholtz-muenchen.de

A. Zhigun
Centre for Mathematical Sciences, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany
e-mail: zhigun@ma.tum.de

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_9,

9.1 Introduction

In this chapter, we consider the following model:

$$\partial_t M = \partial_x \left(|M|^\alpha \partial_x M\right) - \partial_x \left(|M|^\gamma \partial_x \rho\right) + f(t, M, \rho) \text{ in } (\tau, \infty) \times (a, b), \quad (9.1)$$

$$\partial_t \rho = \partial_{xx} \rho - g(t, M, \rho) \qquad \text{in } (\tau, \infty) \times (a, b), \quad (9.2)$$

$$M(t, a) = M(t, b) = 0, \quad \rho(t, a) = \rho(t, b) = 1 \qquad \text{for } t \in (\tau, \infty), \quad (9.3)$$

$$M(\tau, x) = M_\tau, \quad \rho(\tau, x) = \rho_\tau \qquad \text{for } x \in (a, b), \quad (9.4)$$

where $\tau, a, b \in \mathbb{R}$, $a < b$, and α and γ are given positive constants satisfying

$$\frac{\alpha}{2} + 1 \leq \gamma < \alpha. \quad (9.5)$$

Remark 9.1.

1. We call conditions (9.5) "balance" conditions since they establish a balance between the diffusion and transport terms, that is, between the porous medium and chemotaxis effects.
2. It is clear from (9.5) that $\alpha, \gamma > 2$ should hold.

This system of partial differential equations, models, for example, a population described in terms of its density M which grows in dependence of a substrate with concentration ρ. The governing equation, Eq. (9.1), describes the evolution of the population. The spatial movement of the population is caused by two different effects. First of all, Eq. (9.1) includes a density-dependent diffusion term. This nonlinear diffusion effect becomes stronger as the population grows larger locally, following a power law as in case of the porous medium equation. Secondly, the population moves toward regions with increased substrate availability, i.e., follows a chemical signal with concentration ρ. This effect is also controlled by the population density, and its intensity increases as the local population density grows. Both effects of population mobility increase/diminish with the population, each following a power law. Thus, Eq. (9.1) degenerates wherever $M = 0$. Finally, Eq. (9.1) includes a "source" term: a nonlinear reaction interaction term f. As usual, it stays for the sink/source density (net number of particles created per unit time and unit volume). The evolution equation (9.2) for the substrate includes the standard linear diffusion term and a nonlinear reaction interaction term g.

The study of the model (9.1)–(9.4) expands the previous work of the authors (see [7], where the autonomous case was studied). In our research, we generalize the model to include the case of nonautonomous reaction interaction terms f and g. We make the following assumptions: for all $t, M, \rho \in \mathbb{R}$ let

$$|f(t,M,\rho)| \le f_1(t)(1+|M|^{\xi})^{\frac{1}{2}},\ f_1 \in L^2_b(\mathbb{R}),\ 0 \le \xi < \alpha - \gamma + 2, \tag{9.6}$$

$$f(t,M,\rho)M \le -F_2M^2 + f_3(t)|M|,\ f_3 \in L^{\kappa}_b(\mathbb{R}),\ \kappa > 1, \tag{9.7}$$

$$g(t,M,\rho) = g_1(t)\rho + g_2(t,\rho)M,\ g_1 \in C^1(\mathbb{R}):\ \partial_t g_1 \le 0,\ g_1(\pm\infty) < \infty, \tag{9.8}$$

$$|g_2(t,\rho)| \le g_3(t),\ g_3 \in L^{\eta}_b(\mathbb{R}),\ \eta > 4, \tag{9.9}$$

and, in order to ensure the uniqueness and the non-negativity of solutions for nonnegative initial data,

$$\tilde{f}(t,M,\rho) := f\left(t, M|M|^{\frac{2}{2+\alpha}-1}, \rho\right) - F_4 M|M|^{\frac{2}{2+\alpha}-1},\ \frac{\partial \tilde{f}}{\partial M} \in L^{\infty}_{loc}(\mathbb{R}^3), \tag{9.10}$$

$$Df \in L^{\infty}_{loc}(\mathbb{R}^3),\ Dg_2 \in L^{\infty}_{loc}(\mathbb{R}^2),\ f(t,0,\rho) = 0,\ g_2(t,0) = 0, \tag{9.11}$$

where f_1, f_3, g_1, g_3 are nonnegative functions and F_2, F_4 are some constants, F_2 is strictly positive and for $p \in [1,\infty]$, $Q \subset \mathbb{R}^m$

$$L^p_{loc}(Q) = \{u : Q \to \mathbb{R} : u \in L^p(K) \text{ for all compact sets } K \subset Q\},$$

$$L^p_b(Q) = \left\{u \in L^p_{loc}(Q) : ||u||_{L^p_b(Q)} := \sup_{x_0 \in \mathbb{R}^m} ||u||_{L^p\left(Q \cap B^1_{x_0}\right)} < \infty\right\}$$

where $B^1_{x_0}$ is a ball of unit radius centered at x_0. We would like to point out that conditions on f and g are forced by the analysis of well-posedness of the system (9.1)–(9.4). The admissible reaction-interaction terms that enjoy (9.6)–(9.11) are not those commonly used in most population studies. The following example of functions f and g satisfies the conditions (9.6)–(9.11):

Example 9.1.

$$f(t,M,\rho) = -M + \frac{M_+^{\frac{2+\alpha}{2}}}{M_+^{\frac{2+\alpha}{2}} + 1}\sin(t),$$

$$g(t,M,\rho) = \frac{1}{1+t}\rho + M\frac{\rho_+}{\rho_+ + 1}\cos(t),$$

where $M_+ = \max\{M, 0\}$. For the initial data, we assume that

$$M_\tau \in L^{\infty}(a,b),\ \rho_\tau \in W^{1,\infty}(a,b).$$

In this paper, we treat weak solutions of the system (9.1)–(9.4). The definition is as follows:

Definition 9.1 (Weak Solution). A pair of functions (M, ρ) defined in $[\tau, \infty) \times [a, b]$ is said to be a weak solution of (9.1)–(9.4) for $M_\tau \in L^\infty(a, b)$, $\rho_\tau \in W^{1,\infty}(a, b)$, if for all $T > \tau$

(i) $M \in L^\infty((\tau, T) \times \Omega)$, $|M|^\alpha M \in L^2\left((\tau, T); H_0^1(a, b)\right)$, $\partial_t M \in L^2\left((\tau, T); H^{-1}(a, b)\right)$;
(ii) $\rho - 1 \in C((\tau, T); W_0^{1,\infty}(a, b))$;
(iii) (M, ρ) satisfies

$$\int_\tau^T (M, v)\, \partial_t \varphi - \left(|M|^\alpha \partial_x M - |M|^\gamma \partial_x \rho, \partial_x v\right) \varphi + \left(f(s, M, \rho), v\right) \varphi \, ds = 0$$

for any $v \in H_0^1(a, b)$, $\varphi \in C_0^\infty[\tau, T]$, $M(\tau) = M_\tau$ in $C_w((\tau, T); L^2(a, b))$-sense and

$$\rho(t) - 1 = e^{(t-\tau)\Delta}(\rho_\tau - 1) - \int_\tau^t e^{(t-s)\Delta} g(s, M(s), \rho(s))\, ds$$

in $W_0^{1,\infty}(a, b)$.

Notation 1.
$||\cdot||$ stands for $||\cdot||_{L^2(a,b)}$-norm and (u, v) for $\int_a^b u(x)v(x)\, dx$ or, more generally (in the case of distributional derivatives for instance), for $\langle u, v\rangle$.

Remark 9.2. From $M \in L^\infty([\tau, T]; L^2(a, b))$ and $\partial_t M \in L^2\left([\tau, T]; H^{-1}(a, b)\right)$ it follows that $M \in C_w([\tau, T]; L^2(a, b))$ (see [2]), therefore the initial condition for M does make sense.

The main focus of this study is on proving a dissipative estimate for the problem (9.1)–(9.4). We use this intermediate result to establish the existence of a global uniform pullback attractor for our system in $L^\infty(a, b) \times W^{1,\infty}(a, b)$. We emphasize on the fact that the analysis of long-time behavior for the models that contain either porous-medium or chemotaxis-type terms is quite difficult. Interested reader is referred to the results on dynamics of the porous medium equation [3], the classical Keller–Segel model for chemotaxis [8, 15], and the Keller–Segel model with growth in both autonomous and non-autonomous cases [5]. We also mention interesting works [9, 11, 12, 14] (see also references therein) on well-posedness and asymptotic behavior of solutions for a model without growth which includes both of the nonlinear motion effects as in (1)–(4) but which is stated in $\Omega = R^d$. With all three nonlinear effects present in our model, we face significant difficulties. In order to overcome these difficulties, we impose the "balance" conditions between the order of porous-medium degeneracy and the growth order of the chemotaxis

function: $\frac{\alpha}{2}+1 \leq \gamma < \alpha$. We showed in [6] that our model is a well-posed one and that it exhibits no singular behavior. For each pair of starting values the solution is uniformly bounded in time and space.

The condition $\alpha > \gamma$ reads: the density-dependent diffusion coefficient "dominates" the intensity of response to the chemical signal as the population density grows. This, as we showed in [6], results in the uniform boundedness of M and ρ.

On the other hand, we also showed in [6] that even in the areas with low population density, the porous medium effect is due to $\frac{\alpha}{2}+1 \leq \gamma$ strong enough to keep the population spreading without vanishing locally, which means that the support of $M(t,\cdot)$, the set $\overline{\{x \in (a,b) | M(t,x) > 0\}}$, is not shrinking in t.

We proved in [6] the time-global existence and boundedness of solution to our system. The main result of [6] can be summarized as follows:

Theorem 9.1. *Let the functions f and g satisfy the assumptions (9.6)–(9.11) and let the given constants α and γ satisfy $\gamma \in \left[\frac{\alpha}{2}+1, \alpha\right)$. Then the initial boundary-value problem (9.1)–(9.4) has at most one nonnegative solution (in the sense of Definition 9.1) for each pair of starting values $(M_\tau, \rho_\tau) \in L^\infty(a,b) \times W^{1,\infty}(a,b)$. The solution is uniformly bounded in time in the phase space $L^\infty(a,b) \times W^{1,\infty}(a,b)$.*

However, the estimates derived there were not sufficient to show the existence of the attractor. In this paper, we use the condition $\alpha > \gamma$ to establish a dissipative estimate for our model, which will be necessary to show the existence of attractor.

Our main result can be summarized as follows:

Theorem 9.2. *Let the functions f and g satisfy the assumptions (9.6)–(9.11) and let the given constants α and γ satisfy $\gamma \in \left[\frac{\alpha}{2}+1, \alpha\right)$. Then the following dissipative estimate holds for the initial boundary-value problem (9.1)–(9.4):*

$$||M(t)||_{L^\infty(a,b)} + ||\rho(t)||_{W^{1,\infty}(a,b)} \leq C_\infty \left(||M_\tau||_{L^\infty(a,b)} + ||\rho_\tau||_{W^{1,\infty}(a,b)}\right)^{r_\infty} \cdot e^{-\omega_\infty (t-\tau)} + D_\infty \ \forall t \geq \tau, \tag{9.12}$$

where the positive constants $C_\infty, r_\infty, \omega_\infty, D_\infty$ depend only on α, γ, f and g and are independent of M_τ, ρ_τ or t.

Remark 9.3. As will become clear from the proof, we do not actually need the condition $\gamma \geq \frac{\alpha}{2}+1$ for the dissipative estimate that we want to obtain, but it is crucial for well-posedness (see [6]).

We will prove this theorem in *Sect. 9.2*. As a consequence of *Theorems 9.1* and *9.2*, we obtain the existence of a global uniform pullback attractor for (9.1)–(9.4): we prove in *Sect. 9.3* the following

Theorem 9.3. *Let the functions f and g satisfy the assumptions (9.6)–(9.11) and let the given constants α and γ satisfy $\gamma \in \left[\frac{\alpha}{2}+1, \alpha\right)$. Further let the solution of the problem (9.1)–(9.4) be described by the process $\{U(t,\tau)\}_{t \geq \tau}$. Then there exists a family $\{\mathscr{A}(t)\}_{t \in \mathbb{R}}$ of sets with the properties:*

1. $\{\mathscr{A}(t)\}_{t\in\mathbb{R}}$ *is a global uniform pullback attractor for the process* $\{U(t,\tau)\}_{t\geq\tau}$ *in* $L^\infty(a,b)\times W^{1,\infty}(a,b)$;
2. $\bigcup\limits_{t\in\mathbb{R}}\mathscr{A}(t)$ *is relative compact in* $L^\infty(a,b)\times W^{1,\infty}(a,b)$.

9.2 Dissipative Estimates (Proof of *Theorem 9.2*)

In this section, we derive a collection of coupled dissipative estimates for M and $\nabla\rho$ in various L^δ norms, with $\delta<\infty$ for the M component, and then apply a bootstrap argument in order to obtain the desired dissipative estimate in the L^∞ norm for both components.

We start with rewriting equation (9.1) in the following way:

$$\partial_t M=\partial_x\left(|M|^\gamma\partial_x\left(\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right)\right)+f(t,M,\rho)$$

In order to derive our first *a priori* estimate, we multiply this equation by $\left(\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right)$ and integrate over (a,b) to obtain that

$$\begin{aligned}
&\left(\partial_t M,\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right)\\
&=-\left(|M|^\gamma,\left|\partial_x\left(\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right)\right|^2\right)\\
&\quad+\left(f(t,M,\rho),\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right)\\
&\leq\left(f(t,M,\rho),\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right).\\
\Leftrightarrow\quad&\frac{d}{dt}\left(\frac{1}{(\alpha-\gamma+1)(\alpha-\gamma+2)}\left\||M|^{\frac{\alpha-\gamma+2}{2}}\right\|^2-(M,\rho)\right)\\
&\leq\left(f(t,M,\rho),\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right)-(\partial_t\rho,M)
\end{aligned}\tag{9.13}$$

and we multiply equation (9.2) by $(\partial_t\rho+\rho-1)$ in the same sense as above to obtain that

$$\begin{aligned}
&\|\partial_t\rho\|^2+\frac{1}{2}\frac{d}{dt}||\rho-1||^2=-\frac{1}{2}\frac{d}{dt}\|\partial_x\rho\|^2-\|\partial_x\rho\|^2-(g(t,M,\rho),\partial_t\rho+\rho-1)\Leftrightarrow\\
&\frac{1}{2}\frac{d}{dt}\left(\|\partial_x\rho\|^2+||\rho-1||^2\right)=-\|\partial_x\rho\|^2-\|\partial_t\rho\|^2-(g(t,M,\rho),\partial_t\rho+\rho-1).
\end{aligned}\tag{9.14}$$

Adding inequalities (9.13) and (9.14) together, we obtain that

$$\begin{aligned}
&\frac{d}{dt}\left(\frac{1}{(\alpha-\gamma+1)(\alpha-\gamma+2)}\left\||M|^{\frac{\alpha-\gamma+2}{2}}\right\|^2-(M,\rho)+\frac{1}{2}\left\|\partial_x\rho\right\|^2+\frac{1}{2}||\rho-1||^2\right)\\
&\leq\left(f(t,M,\rho),\frac{1}{\alpha-\gamma+1}M|M|^{(\alpha-\gamma+1)-1}-\rho\right)-\left\|\partial_x\rho\right\|^2-(\partial_t\rho,M)-\left\|\partial_t\rho\right\|^2\\
&\quad-\left(g(t,M,\rho),\partial_t\rho+\rho-1\right).
\end{aligned}\tag{9.15}$$

We consider first the term containing $g(t,M,\rho)=g_1(t)\rho+g_2(t,\rho)M$. It holds:

$$\begin{aligned}
-\left(g_1\rho,\partial_t\rho+\rho-1\right)=&-\frac{1}{2}\frac{d}{dt}\left(g_1||\rho||^2\right)+\frac{1}{2}\frac{d}{dt}g_1||\rho||^2-g_1\left(||\rho||^2-(1,\rho)\right)\\
\underset{(9.8)}{\leq}&-\frac{1}{2}\frac{d}{dt}\left(g_1||\rho||^2\right)-g_1\left(||\rho||^2-(1,\rho)\right)\\
\leq&-\frac{1}{2}\frac{d}{dt}\left(g_1||\rho||^2\right)-(1-\varepsilon)g_1||\rho||^2+\frac{1}{4\varepsilon}g_1\\
\underset{(9.8)}{\leq}&-\frac{1}{2}\frac{d}{dt}\left(g_1||\rho||^2\right)-(1-\varepsilon)g_1||\rho||^2+\frac{1}{4\varepsilon}g_1(\tau)
\end{aligned}\tag{9.16}$$

and

$$\begin{aligned}
-\left(g_2(t,\rho)M,\partial_t\rho+\rho-1\right)\leq&\ \varepsilon\left\|\partial_t\rho\right\|^2+\varepsilon||\rho-1||^2+\frac{1}{2\varepsilon}||g_2(t,\rho)M||^2\\
\underset{(9.9)}{\leq}&\ \varepsilon\left\|\partial_t\rho\right\|^2+\varepsilon||\rho-1||^2+\frac{1}{2\varepsilon}g_3^2||M||^2.
\end{aligned}\tag{9.17}$$

By combining (9.16) and (9.17) with inequality

$$-(\partial_t\rho,M)-\left\|\partial_t\rho\right\|^2\leq\frac{1}{2}||M||^2-\frac{1}{2}\left\|\partial_t\rho\right\|^2\tag{9.18}$$

and by choosing $\varepsilon\leq\frac{1}{2}$, we have that

$$\begin{aligned}
&-(\partial_t\rho,M)-\left\|\partial_t\rho\right\|^2-\left(g(t,M,\rho),\partial_t\rho+\rho-1\right)\\
\leq&-\frac{1}{2}\frac{d}{dt}\left(g_1||\rho||^2\right)-(1-\varepsilon)g_1||\rho||^2+\varepsilon||\rho-1||^2+\frac{1}{4\varepsilon}g_1(\tau)-\left(\frac{1}{2}-\varepsilon\right)\left\|\partial_t\rho\right\|^2\\
&+\left(\frac{1}{2}+\frac{1}{2\varepsilon}g_3^2\right)||M||^2
\end{aligned}$$

$$\underset{\varepsilon\leq\frac{1}{2}}{\leq} -\frac{1}{2}\frac{d}{dt}\left(g_1||\rho||^2\right)-(1-\varepsilon)g_1||\rho||^2+\varepsilon||\rho-1||^2+\frac{1}{4\varepsilon}g_1(\tau)$$
$$+\left(\frac{1}{2}+\frac{1}{2\varepsilon}g_3^2\right)||M||^2. \tag{9.19}$$

Further, we can estimate the terms with f from (9.15) in the following way:

$$\left(f(t,M,\rho),M|M|^{(\alpha-\gamma+1)-1}\right)\underset{(9.7)}{\leq}\left(-F_2M^2+f_3|M|,|M|^{(\alpha-\gamma+1)-1}\right)$$
$$=-F_2\left\||M|^{\frac{\alpha-\gamma+2}{2}}\right\|^2+f_3\left\||M|^{\frac{\alpha-\gamma+1}{2}}\right\|^2, \tag{9.20}$$

$$-\left(f(t,M,\rho),\rho\right)\underset{(9.6)}{\leq}\varepsilon||\rho||^2+\frac{1}{4\varepsilon}f_1^2\left(1+\left\||M|^{\frac{\xi}{2}}\right\|^2\right)$$
$$\leq 2\varepsilon||\rho-1||^2+2\varepsilon+\frac{1}{4\varepsilon}f_1^2+\frac{1}{4\varepsilon}f_1^2\left\||M|^{\frac{\xi}{2}}\right\|^2. \tag{9.21}$$

Using inequalities (9.19)–(9.21), we conclude from (9.15) that

$$\frac{d}{dt}\left(\frac{1}{(\alpha-\gamma+1)(\alpha-\gamma+2)}\left\||M|^{\frac{\alpha-\gamma+2}{2}}\right\|^2-(M,\rho)\right.$$
$$\left.+\frac{1}{2}\left\|\partial_x\rho\right\|^2+\frac{1}{2}||\rho-1||^2+\frac{1}{2}g_1||\rho||^2\right)$$
$$\leq -F_2\left\||M|^{\frac{\alpha-\gamma+2}{2}}\right\|^2+f_3\left\||M|^{\frac{\alpha-\gamma+1}{2}}\right\|^2+\frac{1}{4\varepsilon}f_1^2\left\||M|^{\frac{\xi}{2}}\right\|^2+\left(\frac{1}{2}+\frac{1}{2\varepsilon}g_3^2\right)||M||^2$$
$$-\left\|\partial_x\rho\right\|^2-(1-\varepsilon)g_1||\rho||^2+3\varepsilon||\rho-1||^2+2\varepsilon+\frac{1}{4\varepsilon}g_1(\tau)+\frac{1}{4\varepsilon}f_1^2. \tag{9.22}$$

In order to shorten the formulas, we introduce a new variable:

$$\varphi:=\frac{1}{(\alpha-\gamma+1)(\alpha-\gamma+2)}\left\||M|^{\frac{\alpha-\gamma+2}{2}}\right\|^2-(M,\rho)$$
$$+\frac{1}{2}\left\|\partial_x\rho\right\|^2+\frac{1}{2}||\rho-1||^2+\frac{1}{2}g_1||\rho||^2+C \tag{9.23}$$

where the constant C can be chosen in such a way that $\varphi\geq 1$ holds. Indeed, $|M|^{\frac{\alpha-\gamma+2}{2}}$ is the leading M-power present in the expression (9.23) due to the assumptions made on α,γ and ξ, and we also have the estimate

$$(M,\rho)\leq\varepsilon||\rho||^2+\frac{1}{4\varepsilon}||M||^2 \tag{9.24}$$

valued for all $\varepsilon > 0$. Moreover, applying the Poincaré and the Hölder inequalities and adjusting the constants C and ε, we can deduce from (9.22) inequality

$$\frac{d}{dt}\varphi \leq -A_1\varphi + a_2\varphi^{\theta} \tag{9.25}$$

for some $A_1 \in \mathbb{R}_+$ and $a_2 \in L^1_b(\mathbb{R})$, $a_2 \geq 0$ and

$$\theta := \frac{\max\left\{\frac{\alpha-\gamma+1}{2}, \frac{\xi}{2}\right\}}{\frac{\alpha-\gamma+2}{2}} \in (0,1).$$

A simple calculation shows that any solution φ of inequality (9.25) satisfies inequality

$$\varphi(t) \leq \left(\varphi^{1-\theta}(\tau)e^{-A_1(1-\theta)(t-\tau)} + (1-\theta)\int_\tau^t e^{-A_1(1-\theta)(t-s)}a_2(s)\,ds\right)^{\frac{1}{1-\theta}}. \tag{9.26}$$

Applying *Lemma 9.1* from the *Appendix* to inequality (9.26) and tacking into account that $a_2 \in L^1_b(\mathbb{R})$ and inequality (9.24) holds, we finally obtain our first dissipative estimate. Set for short

$$\begin{aligned} y_{\delta_0} &:= ||M||^{\delta_0}_{\delta_0} + 1 + \|\partial_x\rho\|^2, \\ \delta_0 &:= \alpha - \gamma + 2 > 2, \end{aligned} \tag{9.27}$$

it holds then:

$$y_{\delta_0}(t) \leq C_{y_{\delta_0}}\, y_{\delta_0}(\tau)e^{-\omega_{y_{\delta_0}}(t-\tau)} + D_{y_{\delta_0}}. \tag{9.28}$$

for some $C_{y_{\delta_0}}, \omega_{y_{\delta_0}}, D_{y_{\delta_0}}$ that dependent only upon the parameters of the problem.

Notation 2. *For the sake of convenience, we assume that the constants B_i (appear below) for all indices i are only dependent upon the parameters of the problem (9.6)–(9.11), that is, upon the constants a, b,α, γ, η, κ, ξ, $||f_1||_{L^2_b(\mathbb{R})}$, F_2, $||f_3||_{L^\kappa_b}$, $||g_1||_\infty$, $||g_3||_{L^\eta_b(\mathbb{R})}$, and **not** upon the initial data M_τ, ρ_τ or the time variables τ and t, or (unless stated otherwise) any other parameters.*

In what follows, we use (9.28) in order to obtain several intermediate dissipative estimates for M and ρ, which in turn lead to an L^∞-dissipative estimate. The following observation, which is an implication from the theory of abstract parabolic evolution equations (see [15]), will be helpful further.

Having a $\delta \in (2, \infty)$ fixed, consider the unbounded operator

$$\partial_{xx} : L^\delta(a,b) \to L^\delta(a,b)$$

equipped with the domain

$$D\left(\partial_{xx}\right) := \left\{ u \in W_0^{1,\delta}(a,b) : \ \partial_{xx} u \in L^{\delta}(a,b) \right\}.$$

It is known (see [15]) that this operator generates an analytic semigroup $e^{t\partial_{xx}}$ and its spectrum lies entirely in $\{\lambda \in \mathbb{R} : \lambda \leq -\beta\}$ for some $\beta > 0$. As such it has the following properties:

$$\left(-\partial_{xx}\rho\right)^{\mu} e^{t\partial_{xx}} = e^{t\partial_{xx}} \left(-\partial_{xx}\rho\right)^{\mu}, \tag{9.29}$$

$$\left\| e^{t\partial_{xx}} \left(-\partial_{xx}\rho\right)^{\mu} \right\|_{\delta} \leq A^{\mu,\delta} e^{-\beta t} t^{-\mu} \tag{9.30}$$

for all $t > 0$ and $\mu > 0$ for some constants $A^{\mu,\delta}$. Now, Eq. (9.2) can be rewritten in the following way:

$$\partial_t(\rho - 1) = \partial_{xx}(\rho - 1) - g(t, M, \rho)$$

and can thus be regarded as an abstract parabolic evolution equation with respect to $\rho - 1$. Therefore for all $t > 0$ holds:

$$\rho(t) - 1 = e^{(t-\tau)\partial_{xx}}(\rho(\tau) - 1) - \int_{\tau}^{t} e^{(t-s)\partial_{xx}} g(s, M(s), \rho(s))\, ds. \tag{9.31}$$

and applying the operator ∂_x to both sides of (9.31) and making use of the property (9.29), we obtain that

$$\partial_x \rho(t) = e^{(t-\tau)\partial_{xx}} \partial_x \rho(\tau) - \int_{\tau}^{t} \partial_x \left(e^{(t-s)\partial_{xx}} g(s, M(s), \rho(s)) \right) ds. \tag{9.32}$$

We want to estimate $\|\partial_x \rho\|_{\infty}$ using (9.32). The initial value $\rho(\tau)$ is assumed to be sufficiently smooth, so that holds

$$\|\partial_x \rho(\tau)\|_{\infty} < \infty. \tag{9.33}$$

What remains is to estimate the L^{∞}-norm of the integral from (9.32) with the help of (9.30) and the assumptions on g. Choosing $\mu = \frac{3}{4}$ and recalling that $W^{\frac{3}{2},\hat{\delta}}(a,b) \hookrightarrow W^{1,\infty}(a,b)$ for $\hat{\delta} > 2$, we arrive at the estimate

$$\begin{aligned}
&\left\| \int_{\tau}^{t} \partial_x \left(e^{(t-s)\partial_{xx}} g(s, M(s), \rho(s)) \right) ds \right\|_{\infty} \\
\leq &\int_{\tau}^{t} \left\| \left(-\partial_{xx}\right)^{\frac{3}{4}} \left(e^{(t-s)\partial_{xx}} g(s, M(s), \rho(s)) \right) \right\|_{\hat{\delta}} ds \\
\leq &A^{\frac{3}{4},\hat{\delta}} \int_{\tau}^{t} e^{-\beta(t-s)} (t-s)^{-\frac{3}{4}} \left(|g_1(s)| \|\rho(s)\|_{\hat{\delta}} + g_3(s) \|M(s)\|_{\hat{\delta}} \right) ds.
\end{aligned} \tag{9.34}$$

Altogether, we obtain from (9.32)–(9.34) the following estimate:

$$\begin{aligned}\|\partial_x\rho(t)\|_\infty \leq & e^{-\beta t}\|\partial_x\rho(\tau)\|_\infty + A^{\frac{3}{4},\delta}\cdot\\ & \cdot\int_\tau^t e^{-\beta(t-s)}(t-s)^{-\frac{3}{4}}(|g_1(s)|+g_3(s))\left(\|\rho(s)\|_\delta + \|M(s)\|_\delta\right)ds.\end{aligned} \tag{9.35}$$

Leaving this result for a moment and returning to Eq. (9.1), we multiply this equation by $M|M|^{\delta-1}$ for an arbitrary $\delta \geq \alpha-\gamma+1$, so that all occurring powers remain nonnegative, and (formally) integrate over (a,b):

$$\begin{aligned}\left(\partial_t M, M|M|^{\delta-1}\right) = & \left(\partial_x\left(|M|^\alpha\partial_x M\right) - \partial_x\left(|M|^\gamma\partial_x\rho\right)\right.\\ & \left.+f(t,M,\rho), M|M|^{\delta-1}\right).\end{aligned}$$

It follows:

$$\begin{aligned}\frac{1}{\delta+1}\frac{d}{dt}\left\||M|^{\frac{\delta+1}{2}}\right\|^2 = & -\frac{4\delta}{(\alpha+\delta+1)^2}\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|^2\\ & +\frac{2\delta}{\alpha+\delta+1}\left(\partial_x|M|^{\frac{\alpha+\delta+1}{2}}, |M|^{\gamma-\frac{\alpha}{2}+\frac{\delta-1}{2}}\partial_x\rho\right)\\ & +(f(t,M,\rho), M|M|^{\delta-1}).\end{aligned} \tag{9.36}$$

Denote $\vartheta(\delta) := \frac{\gamma-\frac{\alpha}{2}+\frac{\delta-1}{2}}{\frac{\alpha+\delta+1}{2}}$. Then $\vartheta(\delta) < 1$ holds due to the assumption $\alpha > \gamma$. Applying Hölder's inequality, we obtain that

$$\begin{aligned}\left(\partial_x|M|^{\frac{\alpha+\delta+1}{2}}, |M|^{\gamma-\frac{\alpha}{2}+\frac{\delta-1}{2}}\partial_x\rho\right) &= \left(\partial_x|M|^{\frac{\alpha+\delta+1}{2}}, |M|^{\vartheta(\delta)\frac{\alpha+\delta+1}{2}}\partial_x\rho\right)\\ &\leq \|1\|_{\frac{2}{1-\vartheta(\delta)}}\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|\left\||M|^{\frac{\alpha+\delta+1}{2}}\right\|_2^{\vartheta(\delta)}\|\partial_x\rho\|_\infty\\ &\leq B_1\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|^{1+\vartheta(\delta)}\|\partial_x\rho\|_\infty.\end{aligned} \tag{9.37}$$

For the last inequality the Poincaré inequality has been used.

Further, we use once more the Hölder inequality and the assumptions on the function f and write:

$$\left(f(t,M,\rho), M|M|^{\delta-1}\right) \leq -F_2\left\||M|^{\frac{\delta+1}{2}}\right\|^2 + f_3\left\||M|^{\frac{\delta}{2}}\right\|^2 \tag{9.38}$$

$$\leq -F_2\left\||M|^{\frac{\delta+1}{2}}\right\|^2 + f_3\|1\|_{\delta+1}\left(\left\||M|^{\frac{\delta+1}{2}}\right\|^2\right)^{\frac{\delta}{\delta+1}}. \tag{9.39}$$

We can conclude from (9.36) using (9.37) and (9.39) that:

$$\begin{aligned}\frac{1}{\delta+1}\frac{d}{dt}\left\||M|^{\frac{\delta+1}{2}}\right\|^2 \leq & -\frac{4\delta}{(\alpha+\delta+1)^2}\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|^2 \\ & +\frac{2\delta}{\alpha+\delta+1}B_1\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|^{1+\vartheta(\delta)}\|\partial_x\rho\|_\infty \\ & -F_2\left\||M|^{\frac{\delta+1}{2}}\right\|^2+f_3\|1\|_{\delta+1}\left(\left\||M|^{\frac{\delta+1}{2}}\right\|^2\right)^{\frac{\delta}{\delta+1}}.\end{aligned}$$

Since $1+\vartheta(\delta)<2$ it follows with the Young inequality:

$$\begin{aligned}\frac{1}{\delta+1}\frac{d}{dt}\left\||M|^{\frac{\delta+1}{2}}\right\|^2 \leq & -F_2\left\||M|^{\frac{\delta+1}{2}}\right\|^2+f_3\|1\|_{\delta+1}\left(\left\||M|^{\frac{\delta+1}{2}}\right\|^2\right)^{\frac{\delta}{\delta+1}} \\ & +B_2(\delta)\|\partial_x\rho\|_\infty^{\frac{2}{1-\vartheta(\delta)}},\end{aligned} \tag{9.40}$$

where $B_2(\delta)=\frac{1-\vartheta(\delta)}{2}\left(\frac{2\delta}{\alpha+\delta+1}B_1\right)^{\frac{2}{1-\vartheta(\delta)}}\left(\frac{4\delta}{(\alpha+\delta+1)^2}\frac{2}{1+\vartheta(\delta)}\right)^{-\frac{1+\vartheta(\delta)}{1-\vartheta(\delta)}}$, therefore this constant depends only on δ and the parameters of the problem.

Next, we return to equality (9.36) to repeat the whole procedure once more but this time being more precise about the estimates being made and using the regularity achieved up to this point. First, due to (9.38) and two obvious inequalities, we have

$$\begin{aligned}\frac{d}{dt}\left\||M|^{\frac{\delta+1}{2}}\right\|^2 = & -\frac{4\delta(\delta+1)}{(\alpha+\delta+1)^2}\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|^2 \\ & +\frac{2\delta(\delta+1)}{\alpha+\delta+1}\left(\partial_x|M|^{\frac{\alpha+\delta+1}{2}},|M|^{\gamma-\frac{\alpha}{2}+\frac{\delta-1}{2}}\partial_x\rho\right) \\ & +(\delta+1)(f(t,M,\rho),M|M|^{\delta-1}). \\ \leq & -B_3\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|^2 \\ & +(\delta+1)B_4\|\partial_x\rho\|_\infty\left\|\partial_x|M|^{\frac{\alpha+\delta+1}{2}}\right\|\left\||M|^{\frac{\alpha+\delta+1}{2}}\right\|^{\vartheta(\delta)} \\ & -(\delta+1)F_2\left\||M|^{\frac{\delta+1}{2}}\right\|^2+(\delta+1)B_5f_3\left\||M|^{\frac{\alpha+\delta+1}{2}}\right\|^{2\zeta}\end{aligned} \tag{9.41}$$

for $\delta\geq\alpha-\gamma+1$ with $\zeta=\frac{\delta}{\alpha+\delta+1}$.

Recall that $f_3\in L_b^\kappa(\mathbb{R})$ and $\kappa>1$. Taking into account a special case of the interpolation inequality for Sobolev spaces (see [1]):

$$||v||\leq I^{\kappa,p}\|\partial_x v\|^{1-\frac{1}{\kappa}}||v||_p^{\frac{1}{\kappa}},\quad p=\frac{6}{1+2\kappa},$$

we obtain with the help of the Young inequality that

$$\begin{aligned}
&(\delta+1)\left\|\partial_x v\right\|\left|\left|v\right|\right|^{\vartheta(\delta)} \\
\leq&(\delta+1)\left(I^{\kappa,p}\right)^{\vartheta(\delta)}\left\|\partial_x v\right\|^{1+\vartheta(\delta)\left(1-\frac{1}{\kappa}\right)}\left|\left|v\right|\right|_p^{\vartheta(\delta)\frac{1}{\kappa}} \\
\leq&\left(I^{\kappa,p}\right)^{\vartheta(\delta)}\left(\varepsilon\left\|\partial_x v\right\|^2+B_6(\varepsilon)(\delta+1)^{\frac{2}{1-\vartheta(\delta)\left(1-\frac{1}{\kappa}\right)}}\left|\left|v\right|\right|_p^{\frac{2\vartheta(\delta)\frac{1}{\kappa}}{1-\vartheta(\delta)\left(1-\frac{1}{\kappa}\right)}}\right)
\end{aligned} \tag{9.42}$$

and

$$\begin{aligned}
(\delta+1)f_3||v||^{2\zeta} &\leq (\delta+1)f_3\left(I^{\kappa,p}\right)^{2\zeta}\left\|\partial_x v\right\|^{2\zeta\left(1-\frac{1}{\kappa}\right)}\left|\left|v\right|\right|_p^{2\zeta\frac{1}{\kappa}} \\
&\leq \left(I^{\kappa,p}\right)^{2\zeta}\left(\varepsilon\left\|\partial_x v\right\|^2+B_7(\varepsilon)\left(f_3(\delta+1)\right)^{\frac{1}{1-\zeta\left(1-\frac{1}{\kappa}\right)}}\left|\left|v\right|\right|_p^{\frac{2\zeta\frac{1}{\kappa}}{1-\zeta\left(1-\frac{1}{\kappa}\right)}}\right),
\end{aligned} \tag{9.43}$$

where $B_6(\varepsilon)$ and $B_7(\varepsilon)$ depend only on ε and the parameters of the problem. With the Hölder inequality, we also have

$$\left\|\left|M\right|^{\frac{\alpha+\delta+1}{2}}\right\|_p \leq \left\|\left|M\right|^{\frac{\alpha}{2}}\right\|_{\frac{qp}{q-p}}\left\|\left|M\right|^{\frac{\delta+1}{2}}\right\|_q, \quad q>\frac{2}{1+2\kappa}. \tag{9.44}$$

Since $\kappa>1$ is, $\frac{2}{1+2\kappa}<2$ holds, we can assume that $q<2$ and that it is independent from δ. Combining (9.42)–(9.43) for $v:=|M|^{\frac{\alpha+\delta+1}{2}}$ with (9.44) and choosing ε small enough depending only on $I^{\kappa,p}$ and B_3 (thus it depends only on the parameters of the problem), we can conclude from (9.41):

$$\begin{aligned}
\frac{d}{dt}\left\|\left|M\right|^{\frac{\delta+1}{2}}\right\|^2 \leq\ & B_8\left(\left\|\partial_x\rho\right\|_\infty(\delta+1)\right)^{\frac{2}{1-\vartheta(\delta)\left(1-\frac{1}{\kappa}\right)}}\left(\left\|\left|M\right|^{\frac{\alpha}{2}}\right\|_{\frac{qp}{q-p}}\left\|\left|M\right|^{\frac{\delta+1}{2}}\right\|_q\right)^{\frac{2\vartheta(\delta)\frac{1}{\kappa}}{1-\vartheta(\delta)\left(1-\frac{1}{\kappa}\right)}} \\
&+B_8\left(f_3(\delta+1)\right)^{\frac{1}{1-\zeta\left(1-\frac{1}{\kappa}\right)}}\left(\left\|\left|M\right|^{\frac{\alpha}{2}}\right\|_{\frac{qp}{q-p}}\left\|\left|M\right|^{\frac{\delta+1}{2}}\right\|_q\right)^{\frac{2\zeta\frac{1}{\kappa}}{1-\zeta\left(1-\frac{1}{\kappa}\right)}} \\
&-F_2(\delta+1)\left\|\left|M\right|^{\frac{\delta+1}{2}}\right\|^2
\end{aligned}$$

for $\delta\geq\alpha-\gamma+1$. Since $\vartheta(\delta),\zeta\in(0,1)$ it follows for all $\delta\geq\alpha-\gamma+2$:

$$\begin{aligned}
\frac{d}{dt}\left(\left\|M\right\|_\delta^\delta+1\right) \leq& B_8\delta^{2\kappa}\left(\left\|\partial_x\rho\right\|_\infty^{2\kappa}+f_3^\kappa+1\right)\left(\left\|M\right\|_{\frac{\alpha}{2}\frac{qp}{q-p}}^\alpha+1\right)\left(\left\|M\right\|_{q\delta/2}^{q\delta/2}+1\right)^{\frac{2}{q}} \\
&-F_2\delta\left(\left\|M\right\|_\delta^\delta+1\right)
\end{aligned}$$

and once more, we get an integral inequality for $\|M(t)\|_\delta^\delta + 1$:

$$\|M(t)\|_\delta^\delta + 1 \leq B_8 \delta^{2\kappa} \int_\tau^t e^{-\delta F_2(t-s)} \left(\|\partial_x \rho(s)\|_\infty^{2\kappa} + f_3^\kappa(s) + 1\right) \left(\|M(s)\|_{\frac{\alpha}{2}\frac{qp}{q-p}}^{\alpha} + 1\right) \cdot \left(\|M(s)\|_{q\delta/2}^{q\delta/2} + 1\right)^{\frac{2}{q}} ds + e^{-\delta F_2(t-\tau)} \left(\|M(\tau)\|_\delta^\delta + 1\right). \tag{9.45}$$

We are now ready to derive more dissipative estimates for the problem (9.1)–(9.4). We will extensively use the following

Lemma 9.1. *Let $z_1, z_2, z_3 : [\tau, +\infty) \to [0, +\infty)$ be such functions that*

$$\begin{aligned} &z_1(t) \leq \psi_1(z_1(\tau)) e^{-\omega_1 t} + D_1, \\ &z_2(t) \leq \psi_2(z_2(\tau)) e^{-\omega_2 t} + D_2, \\ &z_3(t) \leq z_3(\tau) e^{-\omega_3 t} + \int_\tau^t e^{-\omega_3(t-s)} d_3(t,s) z_1(s)\, ds, \\ &z_1(\tau), z_2(\tau), z_3(\tau) \geq 1, \end{aligned} \tag{9.46}$$

for some constants $\omega_1, \omega_2, \omega_3 > 0$ and $D_1, D_2 \geq 1$, some non-decreasing functions $\psi_1, \psi_2 : [1, +\infty) \to [1, +\infty)$ and some $d_3 \in L^\infty(\mathbb{R}_\tau^+, L_b^1(\mathbb{R}_\tau^+))$. Then it holds that

1. $(z_1 + z_2)(t) \leq (\psi_1 + \psi_2)((z_1 + z_2)(\tau)) e^{-\min\{\omega_1, \omega_2\} t} + D_1 + D_2$;
2. $z_1 z_2(t) \leq 3 D_1 D_2 \psi_1 \psi_2 (z_1 z_2(\tau)) e^{-\min\{\omega_1, \omega_2\} t} + D_1 D_2$;
3. $z_1^\sigma(t) \leq \max\left\{1, 2^{\sigma-1}\right\} (\psi_1^\sigma(z_1(\tau)) e^{-\sigma\omega_1 t} + D_1^\sigma) \quad \forall \sigma > 0$;
4. *For $\omega_1 \neq \omega_3$*

$$z_3(t) \leq \left(\psi_1(z_1(\tau)) \frac{1}{1 - e^{-|\omega_1 - \omega_3|}} e^{-\min\{\omega_1, \omega_3\} t} + D_1 \frac{1}{1 - e^{-\omega_3}}\right) \cdot ||d_3||_{L^\infty(\mathbb{R}_\tau^+, L_b^1(\mathbb{R}_\tau^+))} + z_3(\tau) e^{-\omega_3 t} \tag{9.47}$$

and for $\omega_3 = \omega_1$

$$z_3(t) \leq \left(\psi_1(z_1(\tau)) \lceil t \rceil e^{-\omega_1 t} + D_1 \frac{1}{1 - e^{-\omega_1}}\right) ||d_3||_{L^\infty(\mathbb{R}_\tau^+, L_b^1(\mathbb{R}_\tau^+))} + z_3(\tau) e^{-\omega_1 t}.$$

For $\omega_1 < \omega_3$, we also have that

$$z_3(t) \leq z_3(\tau) e^{-\omega_3 t} + z_1(t) \int_\tau^t e^{-(\omega_3 - \omega_1)(t-s)} d_3(t,s)\, ds.$$

(See *Appendix 9.3* for some details regarding the proof of this lemma.)

Lemma 9.1 is very useful in our situation. It shows actually that the "dissipative property" is preserved under standard operations (addition, multiplication, raising to a power and integration).

To shorten the formulas let us set

$$h := \|\partial_x \rho\|_\infty + 1,$$
$$u_\delta := ||M||_\delta^\delta + 1, \quad \delta \in [1, \infty).$$

Observe that particular powers of y_{δ_0} and h, h and u_δ (for sufficiently large δ) can be connected with one another by the inequalities of the type (9.46) in the same manner as z_1 and z_3 from *Lemma 9.1* are. From the *Lemma 9.1*, we can conclude that all of them dissipate exponentially with t:

$$h(t) \le C_h (h + y_{\delta_0})^{r_h}(\tau) e^{-\omega_h (t-\tau)} + D_h, \tag{9.48}$$

$$u_\delta(t) \le U \left(u_\delta(\tau) + C_{u_\delta} (h + y_{\delta_0})^{r\delta}(\tau) \right) e^{-\frac{F_2}{2}\delta(t-\tau)} + D_{u_\delta} =: \tilde{u}_\delta(t), \tag{9.49}$$

where the appearing coefficients depend on the parameters of the problem, and only the coefficients C_{u_δ} and D_{u_δ} depend on δ as well. We especially emphasize that r is independent from δ (it will be crucial for the existence of the uniform dissipative estimate). Indeed, from (9.35) and the definition of y_{δ_0} ($y_{\delta_0} > 1$, see (9.27)), we obtain for $\hat{\delta} := \min\{\alpha - \gamma + 2, 3\} > 2$ that

$$\begin{aligned}
\|\partial_x \rho(t)\|_\infty \le & e^{-\beta t} \|\partial_x \rho(\tau)\|_\infty + A^{\frac{3}{4},\hat{\delta}} \cdot \\
& \cdot \int_\tau^t e^{-\beta(t-s)} (t-s)^{-\frac{3}{4}} (|g_1(s)| + g_3(s)) \left(||\rho(s)||_{\hat{\delta}} + ||M(s)||_{\hat{\delta}} \right) ds \\
\le & e^{-\beta t} \|\partial_x \rho(\tau)\|_\infty + C^{\hat{\delta}} A^{\frac{3}{4},\hat{\delta}} \cdot \\
& \cdot \int_\tau^t e^{-\beta(t-s)} (t-s)^{-\frac{3}{4}} (|g_1(s)| + g_3(s)) y_{\delta_0}(s)\, ds
\end{aligned} \tag{9.50}$$

since $W^{\frac{3}{2},\hat{\delta}} \subset W^{1,\infty}$ and $W^{1,2}(a,b) \subset L^{\hat{\delta}}(a,b)$ (with the embedding constant $C^{\hat{\delta}}$). The estimate for h now follows with (9.50) and *Lemma 9.1* due to the fact that for the function $d(t,s) := (t-s)_+^{-\frac{3}{4}} (|g_1(s)| + g_3(s))$ the condition $\sup\limits_{t>0} ||d(t,\cdot)||_{L_b^1(\mathbb{R})} < \infty$ is satisfied (recall that we assumed that $g_1 \in L^\infty(\mathbb{R})$ and $g_3 \in L_b^\eta(\mathbb{R})$, $\eta > 4$).

Let us now check the dissipative estimate (9.49). The estimate (9.40) reads:

$$\frac{1}{\delta} \frac{d}{dt} u_\delta \le -F_2 u_\delta + (b-a) f_3 u_\delta^{\frac{\delta-1}{\delta}} + B_2(\delta) h^{\frac{2}{1-\vartheta(\delta)}}. \tag{9.51}$$

Recall that $\vartheta(\delta) = \frac{\gamma - \frac{\alpha}{2} + \frac{\delta-2}{2}}{\frac{\alpha+\delta}{2}}$ and consequently $\frac{2}{1-\vartheta(\delta)} = \frac{\alpha+\delta}{\alpha-\gamma+1} \leq B_9\delta$ for some B_9 and $\delta \geq \delta_*$ sufficiently large. Now, the Young inequality yields:

$$u_\delta^{\frac{\delta-1}{\delta}} = (\varepsilon u_\delta)^{\frac{\delta-1}{\delta}} \varepsilon^{-\frac{\delta-1}{\delta}} \leq \frac{\delta-1}{\delta}\varepsilon u_\delta + \frac{1}{\delta}\varepsilon^{-(\delta-1)},$$

therefore it follows from (9.51)

$$\frac{d}{dt}u_\delta \leq -\delta\left(F_2 - \varepsilon(b-a)f_3\frac{\delta-1}{\delta}\right)u_\delta + \varepsilon^{-(\delta-1)}(b-a)f_3 + \delta B_2(\delta)h^{B_9\delta}.$$

Gronwall's lemma yields then

$$\begin{aligned} u_\delta(t) \leq & \int_\tau^t e^{-\delta\int_s^t F_2 - \varepsilon(b-a)f_3(s)\frac{\delta-1}{\delta}\,ds}\left(\varepsilon^{-(\delta-1)}(b-a)f_3(s) + \delta B_2(\delta)h^{B_9\delta}(s)\right)\,ds \\ & + e^{-\delta\int_\tau^t F_2 - \varepsilon(b-a)f_3(s)\frac{\delta-1}{\delta}\,ds}u_\delta(\tau). \end{aligned} \tag{9.52}$$

Observe that it holds

$$\begin{aligned} \int_\tau^t F_2 - \varepsilon(b-a)f_3(s)\,ds \geq & F_2(t-\tau) - \varepsilon(b-a)\int_{\lfloor\tau\rfloor}^{\lceil t\rceil} f_3(s)\,ds \\ \geq & F_2(t-\tau) - \varepsilon(b-a)||f_3||_{L_b^1(\mathbb{R})}(\lceil t\rceil - \lfloor\tau\rfloor) \\ \geq & \left(F_2 - \varepsilon(b-a)||f_3||_{L_b^1(\mathbb{R})}\right)(t-\tau) - 2\varepsilon(b-a)||f_3||_{L_b^1(\mathbb{R})}. \end{aligned} \tag{9.53}$$

For $\varepsilon := \frac{F_2}{2(b-a)||f_3||_{L_b^1(\mathbb{R})}}$ it follows with (9.52) and (9.53)

$$\begin{aligned} u_\delta(t) \leq & e^{F_2}\left(\int_\tau^t e^{-(t-s)\delta\frac{F_2}{2}}\left(\varepsilon^{-(\delta-1)}(b-a)f_3(s) + \delta B_2(\delta)h^{B_9\delta}(s)\right)\,ds\right. \\ & \left. + e^{-(t-\tau)\delta\frac{F_2}{2}}u_\delta(\tau)\right). \end{aligned}$$

The dissipate estimate (9.49) follows now with the estimate (9.47) of *Lemma 9.1* and the dissipate estimate (9.48) for h.

Now, we can conclude from inequality (9.45) that

$$u_\delta(t) \leq e^{-\delta F_2(t-\tau)}u_\delta(\tau) + B_8\delta^{2\kappa}\int_\tau^t e^{-\delta F_2(t-s)}H_1(s)\tilde{u}_{\frac{q}{2}\delta}^{\frac{2}{q}}(s)\,ds, \tag{9.54}$$

where

$$H_1(t) := \left(h^{2\kappa}(t) + f_3^{\kappa}(t) + 1\right) \tilde{u}_{\frac{\alpha}{2}\frac{qp}{q-p}}^{\frac{2(q-p)}{qp}}(t).$$

Taking into account that $u_{\frac{q}{2}\delta}^{\frac{2}{q}}$ dissipates with $e^{-\delta\frac{F_2}{2}(t-\tau)}$ and that H_1 dissipates with an exponent independent of δ, we consecutively apply (9.47) to (9.54) and obtain that

$$\begin{aligned} u_\delta(t) \leq & e^{-\delta\frac{F_2}{2}(t-\tau)}u_\delta(\tau) + B_8\delta^{2\kappa}\tilde{u}_{\frac{q}{2}\delta}^{\frac{2}{q}}(t)\int_\tau^t e^{-\delta\frac{F_2}{2}(t-s)}H_1(s)\,ds \\ \leq & e^{-\delta\frac{F_2}{2}(t-\tau)}u_\delta(\tau) + B_{10}\delta^{2\kappa-1}H_2(t)\tilde{u}_{\frac{q}{2}\delta}^{\frac{2}{q}}(t), \end{aligned}$$

where

$$H_2(t) := \left(h^{2\kappa}(t) + ||f_3||_{L_b^\kappa(\mathbb{R})}^\kappa + 1\right)\tilde{u}_{\frac{\alpha}{2}\frac{qp}{q-p}}^{\frac{2(q-p)}{qp}}(t)$$

and $\delta \geq \delta_*$ is sufficiently large. The bound δ_* depends only on the parameters of the problem. Therefore, we may assume that

$$\tilde{u}(t) = e^{-\delta\frac{F_2}{2}(t-\tau)}u_\delta(\tau) + B_{10}\delta^{2\kappa-1}H_2(t)\tilde{u}_{\frac{q}{2}\delta}^{\frac{2}{q}}(t), \tag{9.55}$$

Since

$$u_\delta(\tau) = ||M(\tau)||_\delta^\delta + 1 \leq ||M(\tau)||_\infty(b-a) + 1,$$

we conclude from (9.55) that for

$$A_\delta(t) := \tilde{u}_\delta(t)\left(\frac{e^{\frac{F_2}{2}(t-\tau)}}{||M(\tau)||_\infty + 1}\right)^\delta + 1 \tag{9.56}$$

it holds

$$A_\delta(t) \leq B_{11}H_2(t)\delta^{2\kappa-1}A_{\frac{q}{2}\delta}^{\frac{2}{q}}(t).$$

One can show by induction then that

$$\begin{aligned} A_{\left(\frac{q}{2}\right)^n\delta_*}^{\left(\frac{q}{2}\right)^n}(t) \leq & \left(B_{11}H_2(t)\delta_*^{2\kappa-1}\right)^{\sum_{k=1}^n\left(\frac{q}{2}\right)^k}\left(\frac{q}{2}\right)^{(2\kappa-1)\sum_{k=1}^n k\left(\frac{q}{2}\right)^k}A_{\delta_*}(t) \\ \underset{n\to\infty}{\to} & \left(B_{11}H_2(t)\delta_*^{2\kappa-1}\right)^{\frac{\frac{q}{2}}{1-\frac{q}{2}}}\left(\frac{q}{2}\right)^{(2\kappa-1)\frac{q}{2}\left(\frac{1}{1-\frac{q}{2}}\right)^2}A_{\delta_*}(t) \\ =: & H^{\delta_*}(t)A_{\delta_*}(t). \end{aligned}$$

Therefore, we obtain that

$$\limsup_{\delta\to\infty} A_{\delta}^{\frac{1}{\delta}}(t) \le H(t) A_{\delta_*}^{\frac{1}{\delta_*}}(t). \tag{9.57}$$

Combining (9.57) with (9.56), we finally arrive at an estimate for $||M(t)||_\infty$:

$$\begin{aligned} ||M(t)||_\infty + 1 &= \lim_{\delta\to\infty} u_{\delta}^{\frac{1}{\delta}}(t) \\ &\le \limsup_{\delta\to\infty} \tilde{u}_{\delta}^{\frac{1}{\delta}}(t) \\ &\le H(t)\left(\tilde{u}_{\delta_*}^{\frac{1}{\delta_*}}(t) + (||M_\tau||_\infty + 1)\, e^{-\frac{F_2}{2}(t-\tau)}\right). \end{aligned} \tag{9.58}$$

Now, since the functions H and $\tilde{u}_{\delta_*}$ dissipate exponentially (recall (9.48) and (9.49) and the definition of H and H_2), we apply *Lemma 9.1* to (9.58) and conclude that $||M||_\infty$ dissipates exponentially as well. Moreover, it follows from our proof that there exists a dissipative estimate for $||M||_\infty$ of the form given in (9.12). The dissipative estimate for $\|\partial_x \rho\|_\infty + 1 = h$ is given in (9.48) and the *Theorem 9.2* is thus proved.

9.3 Global Uniform Pullback Attractor (Proof of *Theorem 9.3*)

It is generally known that the long time behavior of a non-autonomous dynamical system can often be described in terms of its uniform pull-back attractor. Let us recall several facts from the general theory of pullback attractors in Banach spaces (for details we refer to [10, 13] and, for further development, to [4, 5] and the references therein). We start with an abstract nonautonomous initial problem

$$\begin{cases} \dfrac{du}{dt} = F(t,u),\ t > \tau, \\ u(\tau) = u_\tau, \tau \in \mathbb{R}, \end{cases} \tag{9.59}$$

in a Banach space $\mathscr{T}$. If the initial problem (9.59) is well-posed for $u_\tau \in \mathscr{T}$, it generates a family of solving operators $U(t,\tau)$ that acts on $\mathscr{T}$ mapping the initial data at time τ onto the solution at time t:

$$U(t,\tau)(u_\tau) := u(t),\ t \ge \tau,\ u_\tau \in \mathscr{T}.$$

This family of operators satisfies the properties

$$U(\tau,\tau) = id_{\mathscr{T}} \ \forall \tau \in \mathbb{R},$$
$$U(t,\tau) = U(t,s) \circ U(s,\tau) \ \forall t \geq s \geq \tau, \ \tau, s, t \in \mathbb{R},$$

where $id_{\mathscr{T}}$ denotes the identity operator. We say that it forms a process on the phase space $\mathscr{T}$.

Definition 9.2. A family $\mathscr{A}(t), \ t \in \mathbb{R}$ is called global uniform pullback attractor in $\mathscr{T}$ if

(i) The sets $\mathscr{A}(t)$ are compact in $\mathscr{T}$ for all $t \in \mathbb{R}$.
(ii) The invariance property holds:

$$U(t,\tau)\mathscr{A}(\tau) = \mathscr{A}(t) \ \forall t \in \mathbb{R}.$$

(iii) The uniform pullback attracting property holds: for every bounded set $B \subseteq \mathscr{T}$

$$\lim_{s \to +\infty} \sup_{t \in \mathbb{R}} \operatorname{dist}_{\mathscr{T}}(U(t,t-s)B, \mathscr{A}(t)) = 0.$$

Here $\operatorname{dist}_{\mathscr{T}}$ denotes the non-symmetric Hausdorff distance between subsets of $\mathscr{T}$:

$$\operatorname{dist}_{\mathscr{T}}(X,Y) := \sup_{x \in \mathscr{T}} \inf_{y \in \mathscr{T}} ||x - y||.$$

Recall now a general criteria for the existence of a global uniform pullback attractor:

Theorem 9.4. *Let $U(t,\tau)$ be a process in a Banach space $\mathscr{T}$, $U(t,\tau) \in C(\mathscr{T})$ for all $t \geq \tau$, and having a compact uniformly absorbing set $K \subseteq \mathscr{T}$. Then the process $U(t,\tau)$ has a global uniform pullback attractor $\mathscr{A}(t), \ t \in \mathbb{R}$ and it holds: $\bigcup_{t \in \mathbb{R}} \mathscr{A}(t) \subseteq K$.*

Remark 9.4. The attraction property (iii) from *Definition 9.2* has the following interpretation: for each time $t \in \mathbb{R}$, $\mathscr{A}(t)$ attracts bounded sets of initial data coming from the past (i.e., from $-\infty$). Forward convergence is not true in general. However, under the assumptions of *Theorem 9.4*, the absorbing set K is uniformly both pullback and forward absorbing. In this case, each global uniform pullback attractor $\mathscr{A}(t)$ is at the same time a global uniform forward-attractor for the process $U(t,\tau)$, i.e., the following forward attracting property (compare with the property (iii) from the *Definition 9.2*) holds: for every bounded set $B \subseteq \mathscr{T}$

$$\lim_{s \to +\infty} \sup_{t \in \mathbb{R}} \operatorname{dist}_{\mathscr{T}}(U(t+s,t)B, \mathscr{A}(t+s)) = 0.$$

Our goal is now to apply the general theory to the problem (9.1)–(9.4). We showed in [6] that the problem (9.1)–(9.4) if considered as an equation with respect to (M,ρ) in the Banach space $L^{\infty}(a,b) \times W^{1,\infty}(a,b)$ is well posed: for each pair

of initial values $(M_\tau, \rho_\tau) \in L^\infty(a,b) \times W^{1,\infty}(a,b)$ there exist a unique solution $(M(t), \rho(t))$, $t \in [\tau, +\infty)$ in terms of *Definition 9.1*. We define the solving process $U(t,\tau)$ as follows:

$$U(t,\tau)(M_\tau, \rho_\tau) := (M(t), \rho(t)) \text{ for all } t \geq \tau.$$

Therefore, it is sufficient to show the existence of a compact uniformly absorbing set in B_* and the continuity of the process operators $U(t,\tau)$ for all $t > \tau$. The general criteria *Theorem 9.4* would be then applicable to $U(t,\tau)$.

Let us first show the existence of a compact uniformly absorbing set. We multiply equation (9.1) with $(\alpha+1)\partial_t |M|^\alpha M$ and integrate over (a,b):

$$(\alpha+1)\left(\partial_t M, \partial_t |M|^\alpha M\right) = \left(\partial_{xx}|M|^\alpha M + (\alpha+1)\hat{f}(t,M,\rho), \partial_t |M|^\alpha M\right).$$

Here:

$$\hat{f}(t,M,\rho) = -\partial_x\left(|M|^\gamma \partial_x \rho\right) + f(t,M,\rho).$$

After integration by parts, we obtain that

$$\left(\frac{\alpha+1}{\frac{\alpha}{2}+1}\right)^2 \left\|\partial_t |M|^{\frac{\alpha}{2}+1}\right\|^2 = -\frac{1}{2}\frac{d}{dt}\left\|\partial_x |M|^{\alpha+1}\right\|^2 + (\alpha+1)\left(\hat{f}(t,M,\rho), \partial_t |M|^\alpha M\right).$$

It follows with multiplying by $t-\tau$:

$$\begin{aligned}
&\frac{1}{2}\frac{d}{dt}\left((t-\tau)\left\|\partial_x |M|^{\alpha+1}\right\|^2\right) \\
=&\frac{1}{2}\left\|\partial_x |M|^{\alpha+1}\right\|^2 - \left(\frac{\alpha+1}{\frac{\alpha}{2}+1}\right)^2 \left(\sqrt{t-\tau}\left\|\partial_t |M|^{\frac{\alpha}{2}+1}\right\|\right)^2 \\
&+ \frac{(\alpha+1)^2}{\frac{\alpha}{2}+1}\left(|M|^{\frac{\alpha}{2}}\hat{f}(t,M,\rho)\sqrt{t-\tau}, \sqrt{t-\tau}\partial_t |M|^{\frac{\alpha}{2}} M\right) \\
\leq&\frac{1}{2}\left\|\partial_x |M|^{\alpha+1}\right\|^2 + \frac{1}{2}C_1(t-\tau)\left\||M|^{\frac{\alpha}{2}}\hat{f}(t,M,\rho)\right\|^2.
\end{aligned}$$

Integrating over $[\tau, t]$, we obtain that

$$\begin{aligned}
(t-\tau)\left\|\partial_x |M(t)|^{\alpha+1}\right\|^2 \leq &\int_\tau^t \left\|\partial_x |M(s)|^{\alpha+1}\right\|^2 \\
&+ C_1(s-\tau)\left\||M(s)|^{\frac{\alpha}{2}}\hat{f}(s,M(s),\rho(s))\right\|^2 ds.
\end{aligned} \tag{9.60}$$

It remains therefore to estimate the integral on the right side of (9.60). We have

$$\begin{aligned}\left\||M|^{\frac{\alpha}{2}}\hat{f}(s,M,\rho)\right\| =& \left\|\frac{\gamma}{\frac{\alpha}{2}+\gamma}\partial_x|M|^{\frac{\alpha}{2}+\gamma}\cdot\partial_x\rho+|M|^{\frac{\alpha}{2}+\gamma}\partial_{xx}\rho+|M|^{\frac{\alpha}{2}}f(s,M,\rho)\right\|\\ \leq& \frac{\gamma}{\frac{\alpha}{2}+\gamma}\left\|\partial_x\rho\right\|_\infty\left\|\partial_x|M|^{\frac{\alpha}{2}+\gamma}\right\|+\left\|\partial_{xx}\rho\right\|_2\left\||M|^{\frac{\alpha}{2}+\gamma}\right\|_\infty\\ &+\left\||M|^{\frac{\alpha}{2}}f(s,M,\rho)\right\|.\end{aligned}$$

From the derivation of dissipative estimates for $||M||_\delta^\delta$ in *Sect. 9.2*, we conclude that there exist nonnegative functions $\Phi_i(s,x,y)$, which are nondecreasing with respect to s, x and y, independent of M_τ and ρ_τ and such that it holds:

$$\int_\tau^t\left\|\partial_x|M|^{\alpha+1}(s)\right\|^2ds\leq\Phi_1\left(t-\tau,||M_\tau||_{L^\infty(a,b)},\left\|\partial_x\rho_\tau\right\|_\infty\right), \tag{9.61}$$

$$\int_\tau^t\left\|\partial_x|M|^{\frac{\alpha}{2}+\gamma}(s)\right\|^2ds\leq\Phi_2\left(t-\tau,||M_\tau||_{L^\infty(a,b)},\left\|\partial_x\rho_\tau\right\|_\infty\right). \tag{9.62}$$

Recall that due to the classical energy estimate

$$\begin{aligned}\int_\tau^t\left\|\partial_{xx}\rho(s)\right\|^2ds\leq&\left\|\partial_x\rho(\tau)\right\|^2+\int_\tau^t\left\|g(s,M(s),\rho(s))\,ds\right\|^2\\ \leq&\Phi_3\left(t-\tau,||M_\tau||_{L^\infty(a,b)},\left\|\partial_x\rho_\tau\right\|_\infty\right).\end{aligned} \tag{9.63}$$

Combining (9.61)–(9.63) with (9.60), we get the following smoothing estimate for M:

$$\sqrt{t-\tau}\left\|\partial_x|M(t)|^{\alpha+1}\right\|\leq\Phi_M\left(t-\tau,||M_\tau||_{L^\infty(a,b)},\left\|\partial_x\rho_\tau\right\|_\infty\right) \tag{9.64}$$

for some nonnegative function $\Phi_M(s,x,y)$, which is, again, nondecreasing with respect to s,x and y and independent of M_τ and ρ_τ. Next, since

$$\begin{aligned}(-\partial_{xx})^{\frac{11}{12}}\rho(t)=&(-\partial_{xx})^{\frac{11}{12}}e^{(t-\tau)\partial_{xx}}\rho_\tau\\ &-\int_\tau^t(-\partial_{xx})^{\frac{11}{12}}e^{(t-\omega)\partial_{xx}}g(\omega,M(\omega),\rho(\omega))\,d\omega,\end{aligned}$$

we obtain that

$$\begin{aligned}&\left\|(-\partial_{xx})^{\frac{11}{12}}\partial_{xx}\rho(t)\right\|_2\\ \leq&\left\|(-\partial_{xx})^{\frac{11}{12}}e^{(t-\tau)\partial_{xx}}\rho_\tau-\int_\tau^t(-\partial_{xx})^{\frac{11}{12}}e^{(t-\omega)\partial_{xx}}g(\omega,M(\omega),\rho(\omega))\,d\omega\right\|_2\\ \leq&C_2(t-\tau)^{-\frac{5}{12}}\left\|\partial_x\rho_\tau\right\|_2+C_2\int_\tau^t(t-\omega)^{-\frac{11}{12}}\left\|g(\omega,M(\omega),\rho(\omega))\,d\omega\right\|_2.\end{aligned}$$

Thus there exists a nonnegative function $\Phi_\rho(t,x,y)$ which is nondecreasing with respect to t,x and y, independent of M_τ and ρ_τ and such that the following smoothing estimate for ρ holds:

$$(t-\tau)^{\frac{5}{12}} \left\| (-\partial_{xx})^{\frac{11}{12}} \partial_{xx}\rho(t) \right\|_2 \leq \Phi_\rho \left(t-\tau, ||M_\tau||_{\hat{\delta}}, \|\partial_x \rho_\tau\|_\infty \right). \tag{9.65}$$

Due to the smoothing properties (9.64)–(9.65) and the compact embeddings

$$H^1(a,b) \subset\subset L^\infty(a,b),$$
$$W^{\frac{11}{6},2}(a,b) \subset\subset W^{1,\infty}(a,b),$$

we obtain that $U(t,\tau)$ maps $L^\infty(a,b)\times W^{1,\infty}(a,b)$-balls onto relative compact sets of $L^\infty(a,b)\times W^{1,\infty}(a,b)$ for all $t\geq\tau$. The dissipative estimate (9.12) provides the existence of a ball $B_* := B(0,2D_\infty)$ centered in 0 with radius $2D_\infty$ which uniformly absorbs all bounded sets of the space $L^\infty(a,b)\times W^{1,\infty}(a,b)$.

Consequently, the set $V_* := \underset{t\in\mathbb{R}}{\cup} U(t,t-T_*)B_*$ for $T_* := T(B_*)$ (recall that T_* is such that $U(t,t-T_*)B_* \subseteq B_*$, T_* exists due to the fact that B_* is an absorbing set) is a relatively compact uniformly absorbing set for the process $U(t,\tau)$.

In [6], we derived local Lipschitz continuity for the solutions of (9.1)–(9.4) in the following sense: for all $T,R>0$ it holds

$$\begin{aligned} &\left\| U(t,\tau)\left(M_\tau^{(1)},\rho_\tau^{(1)}\right) - U(t,\tau)\left(M_\tau^{(2)},\rho_\tau^{(2)}\right) \right\|_{H^{-1}(a,b)\times L^2(a,b)} \\ \leq &L(t,R) \left\| \left(M_\tau^{(1)},\rho_\tau^{(1)}\right) - \left(M_\tau^{(2)},\rho_\tau^{(2)}\right) \right\|_{H^{-1}(a,b)\times L^2(a,b)} \end{aligned} \tag{9.66}$$

for all $\left\| \left(M_\tau^{(i)},\rho_\tau^{(i)}\right) \right\|_{L^\infty(a,b)\times W^{1,\infty}(a,b)} \leq R$, $i=1,2$ and some nonnegative nondecreasing in both t and R function $L(t,R)$ independent of M,ρ and τ.

Recall that due to embedding theorems for Sobolev spaces, we have

$$L^\infty(a,b)\times W^{1,\infty}(a,b) \subset H^{-1}(a,b)\times L^2(a,b). \tag{9.67}$$

Let $\left(M_\tau^{(n)},\rho_\tau^{(n)}\right)$ be a sequence convergent to some (M_τ,ρ_τ). Due to the continuous embedding (9.67) it converges in $H^{-1}(a,b)\times L^2(a,b)$ to the same limit (M_τ,ρ_τ). From the property (9.66), we deduce that the sequences $\left(U(t,\tau)\left(M_\tau^{(n)},\rho_\tau^{(n)}\right)\right)$ converge to $U(t,\tau)\left(M_\tau,\rho_\tau\right)$ in $H^{-1}(a,b)\times L^2(a,b)$ for all $t\geq\tau$. Let us further assume that for some $t\geq\tau$ the sequence $\left(U(t,\tau)\left(M_\tau^{(n)},\rho_\tau^{(n)}\right)\right)$ is convergent in $L^\infty(a,b)\times W^{1,\infty}(a,b)$. Due to (9.67) the limit is $U(t,\tau)\left(M_\tau,\rho_\tau\right)$. Therefore, we can conclude that the operators $U(t,\tau)$ are closed. Since any (nonlinear) closed compact operator is completely

continuous (i.e., continuous and compact), we get the continuity of the operators $U(t,\tau)$ in $L^\infty(a,b)\times W^{1,\infty}(a,b)$.

Now let us denote by $\overline{V}_*$ the closure of the set V_* in of $L^\infty(a,b)\times W^{1,\infty}(a,b)$. The set $\overline{V}_*$ is then a compact uniformly absorbing set for the process $\{U(t,\tau)\}_{t\geq\tau}$ in B_* equipped with $L^\infty(a,b)\times W^{1,\infty}(a,b)$-topology and, as we have just shown, the operators $U(t,\tau)$ are continuous in this space. Applying *Theorem 9.4*, we deduce the existence of a global uniform pullback attractor $\mathscr{A}(t)$ for the process $\{U(t,\tau)\}_{t\geq\tau}$ in of $L^\infty(a,b)\times W^{1,\infty}(a,b)$. It holds: $\bigcup\limits_{t\in\mathbb{R}}\mathscr{A}(t)\subseteq\overline{V}_*$.

Appendix (Proof of the Auxiliary *Lemma 9.1*)

Consider first the differential inequality

$$\frac{d}{dt}y\leq-\omega_y y+d_y y^{\zeta_y}$$

assuming that $y\geq 1$, $\zeta_y\in(0,1)$, $d_y\in L^1_b(\mathbb{R})$ so that with some computation the estimate

$$(y(t))^{1-\zeta_y}\leq(y(\tau))^{1-\zeta_y}e^{-\omega_y(1-\zeta_y)t}+(1-\zeta_y)\int_\tau^t e^{-\omega_y(1-\zeta_y)(t-s)}d_y(s)\,ds$$

follows.

Lemma 9.2. *Let $z_1,z_2,z_3:[\tau,+\infty)\to[0,+\infty)$ be such functions that*

$$\begin{aligned}
&z_1(t)\leq\psi_1(z_1(\tau))e^{-\omega_1 t}+D_1,\\
&z_2(t)\leq\psi_2(z_2(\tau))e^{-\omega_2 t}+D_2,\\
&z_3(t)\leq z_3(\tau)e^{-\omega_3 t}+\int_\tau^t e^{-\omega_3(t-s)}d_3(t,s)z_1(s)\,ds,\\
&z_1(\tau),z_2(\tau),z_3(\tau)\geq 1,
\end{aligned}\tag{9.68}$$

for some constants $\omega_1,\omega_2,\omega_3>0$ and $D_1,D_2\geq 1$, some non-decreasing functions $\psi_1,\psi_2:[1,+\infty)\to[1,+\infty)$ and some $d_3\in L^\infty(\mathbb{R}^+_\tau,L^1_b(\mathbb{R}^+_\tau))$. Then it holds that

1. $(z_1+z_2)(t)\leq(\psi_1+\psi_2)((z_1+z_2)(\tau))e^{-\min\{\omega_1,\omega_2\}t}+D_1+D_2$;
2. $z_1z_2(t)\leq 3D_1D_2\psi_1\psi_2(z_1z_2(\tau))e^{-\min\{\omega_1,\omega_2\}t}+D_1D_2$;
3. $z_1^\sigma(t)\leq\max\left\{1,2^{\sigma-1}\right\}(\psi_1^\sigma(z_1(\tau))e^{-\sigma\omega_1 t}+D_1^\sigma)\ \ \forall\sigma>0$;
4. *For* $\omega_1\neq\omega_3$

$$z_3(t) \le \left(\psi_1(z_1(\tau))\frac{1}{1-e^{-|\omega_1-\omega_3|}}e^{-\min\{\omega_1,\omega_3\}t} + D_1\frac{1}{1-e^{-\omega_3}}\right) \cdot ||d_3||_{L^\infty(\mathbb{R}^+_\tau, L^1_b(\mathbb{R}^+_\tau))} + z_3(\tau)e^{-\omega_3 t} \tag{9.69}$$

and for $\omega_3 = \omega_1$

$$z_3(t) \le \left(\psi_1(z_1(\tau))\,\lceil t\rceil\, e^{-\omega_1 t} + D_1\frac{1}{1-e^{-\omega_1}}\right)||d_3||_{L^\infty(\mathbb{R}^+_\tau, L^1_b(\mathbb{R}^+_\tau))} + z_3(\tau)e^{-\omega_1 t}.$$

For $\omega_1 < \omega_3$*, we also have*

$$z_3(t) \le z_3(\tau)e^{-\omega_3 t} + z_1(t)\int_\tau^t e^{-(\omega_3-\omega_1)(t-s)}d_3(t,s)\,ds.$$

Proof. We only check the property (9.69). Since

$$\begin{aligned}
&\int_\tau^t e^{-\omega_3(t-s)}e^{-\omega_1 s}d_3(t,s)\,ds\\
=&e^{-\min\{\omega_1,\omega_3\}t}\begin{cases}\int_\tau^t e^{-|\omega_1-\omega_3|(t-s)}d_3(t,s)\,ds & \text{if } \omega_1<\omega_3\\ \int_\tau^t e^{-|\omega_1-\omega_3|s}d_3(t,s)\,ds & \text{if } \omega_1>\omega_3\end{cases}\\
\le&\frac{1}{1-e^{-|\omega_1-\omega_3|}}e^{-\min\{\omega_1,\omega_3\}t}||d_3||_{L^\infty(\mathbb{R}^+_\tau, L^1_b(\mathbb{R}^+_\tau))},
\end{aligned}$$

we conclude from (9.68) that

$$\begin{aligned}
&\int_\tau^t e^{-\omega_3(t-s)}d_3(t,s)z_1(s)\,ds\\
\le&\int_\tau^t e^{-\omega_3(t-s)}d_3(t,s)\left(\psi_1(z_1(s))e^{-\omega_1(t-s)} + D_1\right)ds\\
\le&\left(\psi_1(z_1(\tau))\frac{1}{1-e^{-|\omega_1-\omega_3|}}e^{-\min\{\omega_1,\omega_3\}t} + D_1\frac{1}{1-e^{-\omega_3}}\right)\\
&\cdot||d_3||_{L^\infty(\mathbb{R}^+_\tau, L^1_b(\mathbb{R}^+_\tau))},
\end{aligned}$$

and (9.69) follows. □

Acknowledgements The authors express their thanks to T. Senba for many stimulating discussions.

Anna Zhigun is sponsored by The Elite Network of Bavaria.

References

1. Adams, R.A.: Sobolev Spaces. Academic, New York (1975)
2. Babin, A.V., Vishik, M.I.: Attractors of Evolution Equations. North-Holland, Amsterdam (1992)
3. Efendiev, M.A.: Attractors of Degenerate Parabolic Type Equations. AMS (2013, in press)
4. Efendiev, M.A.: Finite and Infinite Dimensional Attractors for Evolution Equations of Mathematical Physics, vol. 33. Gakkotoscho International Series, Tokyo (2010)
5. Efendiev, M.A., Yamamoto, Y., Yagi, A.: Exponential attractors for non-autonomous dissipative system. J. Math. Soc. Japan **63**(2), 647–673 (2011)
6. Efendiev, M.A., Zhigun, A.: On a 'balance' condition for a class of PDEs including porous medium and chemotaxis effect: nonautonomous case. Adv. Math. Sci. Appl. **21**, 285–304 (2011)
7. Efendiev, M.A., Zhigun, A., Senba, T.: On a weak attractor of a class of PDE with degenerate diffusion and chemotaxis (2013, in press)
8. Horstmann, D.: From 1970 until present: the Keller–Segel model in chemotaxis and its consequences, Part I. Jahresbericht der DMV **105**(3), 103–165 (2003)
9. Ishida, S., Yokota, T.: Global existence of weak solutions to quasilinear degenerate Keller–Segel systems of parabolic-parabolic type. J. Differ. Equ. **252**(2), 1421–1440 (2012)
10. Kloeden, P.: Pullback attractors of nonautonomous semidynamical systems. Stoch. Dyn. **3**(1), 101–112 (2003)
11. Luckhaus, S., Sugiyama, Y.: Asymptotic profile with the optimal convergence rate for a parabolic equation of chemotaxis in super-critical cases. Indiana Univ. Math. J. **56**(3), 1279–1297 (2007)
12. Luckhaus, S., Sugiyama, Y.: Large time behavior of solutions in super-critical cases to degenerate Keller–Segel systems. ESAIM, Math. Model. Numer. Anal. **40**(3), 597–621 (2006)
13. Schmalfuss, B.: Attractors for the nonautonomous dynamical systems. In: Gröger, K., Fiedler, B., Sprekels, J. (eds.) Proceedings of EQUADIFF99. World Scientific (2000)
14. Sugiyama, Y.: Global existence in sub-critical cases and finite time blow-up in super-critical cases to degenerate Keller–Segel system. Differ. Integral Equ. **20**, 841–876 (2006)
15. Yagi, A.: Abstract Parabolic Evolution Equations and their Applications. Springer Monographs in Mathematics. Springer, Berlin (2010)

Chapter 10
A Guided Sequential Monte Carlo Method for the Assimilation of Data into Stochastic Dynamical Systems

Sebastian Reich

Dedicated to Jürgen Scheurle on the occasion of his 60th birthday

Abstract Assimilation of measurements into stochastic dynamical systems is challenging due to the generally non-Gaussian behavior of the underlying probability density functions. While sequential Monte Carlo methods have emerged as a methodology for tackling assimilation problems under rather general circumstances, those methods suffer from the curse of dimensionality. At the same time ensemble transform filters, such as the ensemble Kalman filter, have emerged as attractive alternatives to sequential Monte Carlo methods since they also work for high dimensional problems. Typical ensemble transform filters are however based on rather crude approximations to the involved probability density functions and are therefore of limited accuracy. For that reason there have been a number of recent attempts to combine sequential Monte Carlo methods with ensemble transform techniques, so-called guided sequential Monte Carlo (GSMC) methods. In this paper, we first put ensemble transform filters in the context of coupling and optimal transportation and secondly propose a new GSMC method based on combining approximate couplings with importance sampling. The effect of various filtering strategies is demonstrated for a simple Brownian dynamics model.

S. Reich (✉)
Universität Potsdam, Institut für Mathematik, Am Neuen Palais 10, 14469 Potsdam, Germany
e-mail: sreich@math.uni-potsdam.de

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_10,

10.1 Introduction

We consider random dynamical systems [10] induced by the iteration

$$X_{n+1} = \Psi(X_n) + \Xi_n, \qquad n \geq 0, \tag{10.1}$$

under the assumption that $X_0 : \Omega \to \mathbb{R}^N$ is a multivariate random variable over some sample space Ω with given probability density function (PDF) $\pi_0(x)$, $\Xi_n : \Omega \to \mathbb{R}^N$ are independent and identically distributed Gaussian random variables with mean zero and covariance matrix $Q \in \mathbb{R}^{N\times N}$, i.e., $\Xi_n \sim \mathrm{N}(0, Q)$, and $\Psi : \mathbb{R}^N \to \mathbb{R}^N$ is an appropriate map. The associated Chapman–Kolmogorov equation [10] for the marginal PDFs $\pi_n(x)$, i.e., $X_n \sim \pi_n$, is given by

$$\begin{aligned}\pi_{n+1}(x) &= \int_{\mathbb{R}^N} \frac{1}{(2\pi)^{N/2}|Q|^{1/2}} \exp\left(-\frac{1}{2}(x-\Psi(x'))^T Q^{-1}(x-\Psi(x'))\right) \pi_n(x')\, dx' \\ &= \int_{\mathbb{R}^N} \pi(x|x')\, \pi_n(x')\, dx' \end{aligned} \tag{10.2}$$

with Markov transition kernel

$$\pi(x|x') = \frac{1}{(2\pi)^{N/2}|Q|^{1/2}} \exp\left(-\frac{1}{2}(x-\Psi(x'))^T Q^{-1}(x-\Psi(x'))\right).$$

Here $|Q|$ denotes the determinant of Q.

In this paper, we will assume that (10.1) arises from the discretization of an underlying stochastic differential equation (SDE) by the Euler–Maruyama method [13] with step-size $\Delta t > 0$, i.e.,

$$X_{n+1} = X_n + \Delta t f(X_n) + \sqrt{2\Delta t} Z_n, \qquad t_{n+1} = t_n + \Delta t,$$

and the marginal PDFs will be denoted by $\pi_X(x, t_n) = \pi_n(x)$. To avoid confusion we will also use the notation $X(t_n) = X_n$ from now on. Hence $\Psi(X(t_n)) = X(t_n) + \Delta t f(X(t_n))$ is the drift term and $\Xi(t_n) = \sqrt{2\Delta t} Z_n$ describes diffusion with $Z_n \sim \mathrm{N}(0, \gamma I)$ and diffusion coefficient $\gamma > 0$. The zero diffusion limit $\gamma = 0$, i.e.

$$X(t_{n+1}) = \Psi(X(t_n)) = X(t_n) + \Delta t f(X(t_n)),$$

gives formally rise to the Chapman–Kolmogorov equation

$$\pi_X(x, t_{n+1}) = \int_{\mathbb{R}^N} \delta(x - \Psi(x'))\, \pi_X(x', t_n)\, dx',$$

which is equivalent to

$$\pi_X(\Psi(x), t_{n+1})\, |D\Psi(x)| = \pi_X(x, t_n), \tag{10.3}$$

where $D\Psi(x) \in \mathbb{R}^{N\times N}$ denotes the Jacobian matrix of partial derivatives of $\Psi(x)$ and $\delta(\cdot)$ the Dirac delta function. We emphasize that we are not interested in the $\Delta t \to 0$ limit in this paper and will consider $\Delta t > 0$ as a given, fixed quantity. We emphasize furthermore that the algorithms considered in this paper do not depend on this assumption and are applicable to general intermittent data assimilation problems.

We assume the availability of partial observations $y_{\rm obs}(j\Delta t_{\rm obs})$ of the stochastic process generated by (10.1) in discrete time intervals $\Delta t_{\rm obs}$ and for $j \geq 1$. We also assume that $\Delta t_{\rm obs} = L\Delta t$, i.e., measurements are taken every $L \geq 1$ time-steps. The forward model for the observational process is assumed to be linear, i.e.,

$$Y = HX + \Theta \tag{10.4}$$

with forward operator $H : \mathbb{R}^N \to \mathbb{R}^K$, measurement error $\Theta \sim \mathrm{N}(0, R)$, and measurement error covariance matrix $R \in \mathbb{R}^{K\times K}$. It is assumed that measurement errors at different instances in time are independent and identically distributed. The associated likelihood function is denoted by $\pi_Y(y|x)$.

The task of intermittent data assimilation is to determine the conditional PDFs $\pi_X(x, t_n|\mathrm{Y}_k)$ of the random variable $X(t_n)$ at $t_n = n\Delta t$ given collected measurements

$$\mathrm{Y}_k = (y_{\rm obs}(\mathrm{t}_1)^T, y_{\rm obs}(\mathrm{t}_2)^T, \cdots, y_{\rm obs}(\mathrm{t}_k)^T)^T$$

at observation times $\mathrm{t}_j = j\Delta t_{\rm obs}$, $j = 1, \ldots, k$. We will consider the two cases $t_n = \mathrm{t}_k$ (filtering) and $t_n > \mathrm{t}_k$ (prediction). See [12] for an excellent introduction to stochastic processes and filtering.

Let us first consider the pure prediction problem with no observations ($k = 0$). A Monte Carlo simulation of (10.1) would proceed as follows. First, one finds M independent realizations $x_i(t_0)$ from the initial PDF $\pi_X(x, t_0)$. Second, each realization is updated recursively and independently according to

$$x_i(t_{n+1}) = \Psi(x_i(t_n)) + \xi_i(t_n), \qquad n \geq 0, \tag{10.5}$$

where $\xi_i(t_n)$ are independent realizations from the normal distribution $\mathrm{N}(0, Q)$ with $Q = 2\gamma\Delta t I$. Under appropriate conditions on the marginal PDFs $\pi_X(x, t_n)$ it can be shown that the empirical distribution

$$\pi_X^{\rm em}(x, t_n) = \frac{1}{M}\sum_{i=1}^{M} \delta(x - x_i(t_n))$$

converges weakly to $\pi_X(x, t_n)$ as $M \to \infty$ for any fixed $n > 0$.

Monte Carlo simulation approaches for the pure prediction problem have been extended to the combined filtering-prediction problem and have given rise to a broad

range of sequential Monte Carlo methods [2,8]. The essential idea is to augment the realizations $x_i(t_n)$ by weights $w_i(t_n) > 0$ subject to

$$\sum_i w_i(t_n) = 1$$

and to adjust the weights such that they reflect the importance of the samples $x_i(t_n)$ relative to the available measurements Y_k while the samples $x_i(t_n)$ continue to follow the stochastic dynamics (10.1). A common problem of this basic importance sampling approach is a degeneracy of weights which requires resampling techniques. Straightforward resampling can be achieved by eliminating samples with small weights and duplication of those with large weights. Practical experience shows that the just described combined importance sampling–resampling approach of sequential Monte Carlo methods does not work well for high dimensional problems unless the number of samples M is increased at a rate which scales exponentially in the phase space dimension N [3]. At the same time, the ensemble Kalman filter (EnKF) [9] has emerged as a robust alternative to sequential Monte Carlo methods applicable to high dimensional problems. The EnKF relies on Gaussian approximations to the marginal PDFs $\pi_X(x, t_n|\mathrm{Y}_k)$ and can be shown to be statistically inconsistent in the limit $M \to 0$ (contrary to sequential Monte Carlo methods) [16]. Hence the EnKF is only applicable to problems with an unimodal and nearly Gaussian behavior of the underlying PDFs.

Recently, increased efforts have been made to turn sequential Monte Carlo methods viable for large-scale problems. All these efforts have in common that one tries to dynamically steer samples $x_i(t_n)$ to regions of high probability in $\pi_X(x, t_n|\mathrm{Y}_k)$ and hence to maintain nearly uniform weights $w_i(t_n)$ without the need for frequent resampling. We will call these methods guided sequential Monte Carlo (GSMC) methods. Particular instances of GSMC methods have been discussed, for example, by [4, 5, 15, 18]. An alternative line of research is focused on improvements of the EnKF. We mention the rank histogram filter (RHF) of [1] and the moment corrected EnKFs of [16].

In this paper, we combine the coupling/transportation perspective on filtering with sequential Monte Carlo methods in order to propose a novel GSMC method. The outline of this paper is as follows. We will first review Bayes' theorem and its connection to coupling of random variables in Sect. 10.2. This will provide us with an abstract Monte Carlo methodology for the combined filtering-prediction problem. Finding exact couplings is impossible in most practical cases which suggest to combine available couplings, such as EnKFs [9], with sequential Monte Carlo methods. Algorithmic details will be given in Sect. 10.3 while a numerical demonstration is provided in Sect. 10.4.

10.2 Bayes' Theorem, Filtering, and Coupling of Random Variables

Recall that we have assumed that observations are taken in intervals of Δt_{obs} which satisfy $\Delta t_{\text{obs}} = \Delta t L$ for an appropriate integer $L \geq 1$. In this context it is helpful to generalize the Chapman–Kolmogorov equation (10.2) to its L-fold recursive application, i.e.,

$$\begin{aligned}\pi_X(x, t_n + \Delta t_{\text{obs}}) &= \int_{\mathbb{R}^N} \cdots \int_{\mathbb{R}^N} \pi(x|x')\, \pi(x'|x'') \cdots \\ &\qquad \pi(x^{(L-1)}|x^{(L)})\, \pi_X(x^{(L)}, t_n)\, dx' \cdots dx^{(L)} \\ &= \int_{\mathbb{R}^N} \pi_L(x|\tilde{x})\, \pi_X(\tilde{x}, t_n)\, d\tilde{x}.\end{aligned}$$

At the level of PDFs, the sequential data assimilation problem can now be stated as follows: For $j = 0, 1, \ldots$ alternate between

(i) Prediction:

$$\pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_j) = \int_{\mathbb{R}^N} \pi_L(x|x')\, \pi_X(x', \mathrm{t}_j|\mathrm{Y}_j)\, dx', \tag{10.6}$$

(ii) Filtering:

$$\pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_{j+1}) = \frac{\pi_Y(y_{\text{obs}}(\mathrm{t}_{j+1})|x)\, \pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_j)}{\int_{\mathbb{R}^N} \pi_Y(y_{\text{obs}}(\mathrm{t}_{j+1})|x)\, \pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_j)\, dx} \tag{10.7}$$

with likelihood function

$$\pi_Y(y|x) = \frac{1}{(2\pi)^{K/2}|R|^{1/2}} \exp\left(-\frac{1}{2}(y - Hx)^T R^{-1}(y - Hx)\right)$$

from the forward model (10.4).

Recall that $\pi_X(x, 0|\mathrm{Y}_0)$ is equal to the given PDF $\pi_X(x, t_0)$ of the initial random variable $X(t_0)$.

We now summarize a few Monte Carlo approaches for sequential data assimilation. We have already discussed a Monte Carlo approach to simulating the Chapman–Kolmogorov equation (10.2). We formally extend this approach to the data assimilation problem (10.6)–(10.7). We introduce the notation $(x_i^f(\mathrm{t}_j), w_i^f(\mathrm{t}_j))$, $i = 1, \ldots, M$, to denote M weighted samples from the forecast (or predicted) distribution $\pi_X(x, \mathrm{t}_j|\mathrm{Y}_{j-1})$ and, correspondingly, $(x_i^a(\mathrm{t}_j), w_i^a(\mathrm{t}_j))$, $i = 1, \ldots, M$, to denote weighted samples from the analyzed (or filtered) distribution $\pi_X(x, \mathrm{t}_j|\mathrm{Y}_j)$ at time $\mathrm{t}_j = j\Delta t_{\text{obs}}$. It follows that expectation values $\bar{g}$ of a function $g : \mathbb{R}^N \to \mathbb{R}$ can be approximated according to

$$\bar{g}_M^f = \sum_{i=1}^M w_i^f(\mathrm{t}_j)\, g(x_i^f(\mathrm{t}_j)) \approx \int_{\mathbb{R}^N} g(x)\, \pi_X(x, \mathrm{t}_j|\mathrm{Y}_{j-1})\, dx$$

and

$$\bar{g}_M^a = \sum_{i=1}^M w_i^a(\mathrm{t}_j)\, g(x_i^a(\mathrm{t}_j)) \approx \int_{\mathbb{R}^N} g(x)\, \pi_X(x, \mathrm{t}_j|\mathrm{Y}_j)\, dx,$$

respectively.

The basic sequential Monte Carlo method is based on the following importance sampling approach [2, 8]: For $j = 0, 1, \ldots$ alternate between

(i) Prediction:

$$x_i^f(\mathrm{t}_{j+1}) \sim \pi_L(\cdot|x_i^a(\mathrm{t}_j)), \qquad w_i^f(\mathrm{t}_{j+1}) = w_i^a(\mathrm{t}_j), \tag{10.8}$$

(ii) Filtering:

$$x_i^a(\mathrm{t}_{j+1}) = x_i^f(\mathrm{t}_{j+1}), \quad w_i^a(\mathrm{t}_{j+1}) \propto w_i^f(\mathrm{t}_{j+1})\, \pi_Y(y_{\mathrm{obs}}(\mathrm{t}_{j+1})|x_i^f(\mathrm{t}_{j+1})), \tag{10.9}$$

where the constant of proportionality is chosen such that

$$\sum_{i=1}^M w_i^a(\mathrm{t}_{j+1}) = 1.$$

Due to a possible degeneracy of weights, it is necessary to perform resampling either after each filtering step or whenever an appropriate criterion on the distribution of weights is satisfied. Residual resampling is one of the popular resampling methods [2, 14].

We now summarize an alternative approach which leads to constant weights $w_i^f = w_i^a = 1/M$. The basic idea is that of coupling the prior and posterior distributions [6, 19, 22]. In order to explain this idea in more detail, we simplify the notation in (10.7) and use the shorthands

$$\pi_X^{\mathrm{prior}}(x) = \pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_j), \qquad \pi_X^{\mathrm{post}}(x) = \pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_{j+1})$$

for the prior and posterior distributions at t_{j+1}, respectively. A coupling between these two distributions is defined by a joint PDF $\pi_{XZ}(x, z)$ (or more generally by a joint measure μ_{XZ} on $\mathbb{R}^N \times \mathbb{R}^N$) such that

$$\pi_X^{\mathrm{prior}}(x) = \int_{\mathbb{R}^N} \pi_{XZ}(x, z)\, dz$$

and

$$\pi_X^{\text{post}}(z) = \int_{\mathbb{R}^N} \pi_{XZ}(x, z)\, dx,$$

respectively. Given a coupling, one can replace (10.9) by the following Monte Carlo approach:

(ii) Filtering:

$$x_i^a(\mathrm{t}_{j+1}) \sim \pi_Z(\cdot|x_i^f(\mathrm{t}_{j+1})) := \frac{\pi_{XZ}(x_i^f(\mathrm{t}_{j+1}), \cdot)}{\pi_X^{\text{prior}}(x_i^f(\mathrm{t}_{j+1}))}, \quad w_i^a(\mathrm{t}_{j+1}) = w_i^f(\mathrm{t}_{j+1}). \tag{10.10}$$

Hence the filtering step has now the same structural form as (10.8). The important difference is that the Markov transition kernel $\pi_L(x|x')$ is determined explicitly by the model (10.1) while such a transition kernel needs to be constructed using a coupling in case of the Bayesian filtering step (10.7) and depends on the observed $y_{\text{obs}}(\mathrm{t}_{j+1})$. In the literature, this is sometimes called the McKean approach to filtering [17].

There is a further difficulty in that the prior and posterior PDFs are not explicitly available in the context of sequential Monte Carlo methods and that approximations $\bar{\pi}_X^{\text{prior}}$ and $\bar{\pi}_X^{\text{post}}$, respectively, need to be estimated from the available samples x_i^f, their weights w_i^f and likelihoods $\pi_Y(y_{\text{obs}}|x_i^f)$ either via parametric or nonparametric statistics [11, 26].

We now continue with a simple demonstration of the concept of coupling by means of two univariate Gaussian random variables $X \sim \mathrm{N}(\bar{x}, \sigma_{xx}^2)$ and $Z \sim \mathrm{N}(\bar{z}, \sigma_{zz}^2)$. Since the marginal are given a priori, a joint Gaussian has to be of the form $\mathrm{N}(m, P)$ with mean $m = (\bar{x}, \bar{z})^T \in \mathbb{R}^2$ and covariance matrix

$$P = \begin{pmatrix} \sigma_{xx}^2 & \sigma_{xz}^2 \\ \sigma_{zx}^2 & \sigma_{zz}^2 \end{pmatrix} \in \mathbb{R}^{2\times 2},$$

where the only free parameter $\sigma_{xz}^2 = \sigma_{zx}^2$ has to satisfy

$$\sigma_{xx}^2\sigma_{zz}^2 - \sigma_{xz}^4 > 0$$

to make P symmetric positive definite. Setting $\sigma_{xz} = 0$ implies independence of X and Z and is equivalent to defining a coupling via the product PDF

$$\pi_{XZ}(x, z) = \pi_X^{\text{prior}}(x)\pi_X^{\text{post}}(z)$$

in the general case. We now consider $\sigma_{xz}^2 > 0$ and recall that (10.10) requires the conditional PDF $\pi_Z(z|x)$ which in this example is characterized by the conditional mean

$$\hat{z} = \bar{z} + \frac{\sigma_{xz}^2}{\sigma_{xx}^2}\,(x - \bar{x})$$

and the variance

$$\sigma^2 = \sigma_{zz}^2 - \sigma_{zx}^2 \sigma_{xx}^{-2} \sigma_{xz}^2.$$

If one sets $\sigma_{xz}^2 = \sqrt{\sigma_{xx}^2 \sigma_{zz}^2}$, then σ^2 becomes zero and one obtains a deterministic coupling of the two random variables via

$$Z = \bar{z} + \frac{\sigma_{zz}}{\sigma_{xx}}(X - \bar{x}),$$

which amounts to the well-known transformation of univariate Gaussians under a linear function.

Inspired by this example we will search for general deterministic couplings $Z = T(X)$ with associated joint probability measure

$$\mu_{XZ}(dx, dz) = \delta(z - T(x))\, \pi_X^{\rm prior}(x)\, dx\, dz$$

from which it follows via marginalization that T has to satisfy

$$\pi_X^{\rm post}(T(x))\, |DT(x)| = \pi_X^{\rm prior}(x) \tag{10.11}$$

(compare (10.3)). Once a deterministic coupling has been found, (10.10) can be replaced by

(ii) Filtering:

$$x_i^a(\mathrm{t}_{j+1}) = T_{j+1}(x_i^f(\mathrm{t}_{j+1})), \qquad w_i^a(\mathrm{t}_{j+1}) = w_i^f(\mathrm{t}_{j+1}) \tag{10.12}$$

where T_{j+1} is a transport map, satisfying (10.11) at t_{j+1}, given the prior $\pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_j)$ and the measurement $y_{\rm obs}(\mathrm{t}_{j+1})$.

Optimality in the sense of Monge–Kantorovitch [25] is defined by

$$\mu_{XZ}^* = \arg \inf_{\mu_{XZ} \in \Pi} \int_{\mathbb{R}^N \times \mathbb{R}^N} \|x - z\|^2 \mu_{XZ}(dx, dz),$$

where the infimum runs over the set of all couplings μ_{XZ}, denoted by Π, with marginals $\pi_X^{\rm prior}$ and $\pi_X^{\rm post}$. We first note that finding the optimal coupling between empirical measures

$$\mu_X^{\rm prior}(dx) = \frac{1}{M} \sum_{i=1}^{M} \delta(x - x_i^f)\, dx$$

and

$$\mu_X^{\text{post}}(dz) = \sum_{i=1}^{M} w_i \delta(z - x_i^f)\, dz$$

leads to a linear programming problem [6, 24]. Another key result of optimal transportation states that the optimal coupling is induced by a transport map $T(x)$ for sufficiently regular prior PDFs π_X^{prior} [25]. Furthermore, the transport map satisfies $T(x) = \nabla_x \psi(x)$ for an appropriate potential $\psi : \mathbb{R}^N \to \mathbb{R}$. It follows from (10.11) that the potential ψ has to satisfy the nonlinear elliptic PDE

$$\pi_X^{\text{post}}(\nabla_x \psi(x))\, |D\nabla_x \psi(x)| = \pi_X^{\text{prior}}(x). \tag{10.13}$$

For univariate prior and posterior random variables with cumulative probability distribution functions

$$F_{\text{prior}}(x) = \int_{-\infty}^{x} \pi_X^{\text{prior}}(x')\, dx'$$

and $F_{\text{post}}(x)$, respectively, a transport map is easily found via

$$z = T(x) = F_{\text{post}}^{-1}(F_{\text{prior}}(x)). \tag{10.14}$$

It becomes however computationally infeasible to solve (10.13) for ψ in case the dimension N of phase space is large and/or π_X^{prior} is non-Gaussian. Being forced to give up the idea of strict optimality, one can resort to an idea of [20] (see also [25]) to find *a* deterministic coupling. Moser suggested to utilize a dynamic embedding of the form

$$\frac{dx}{ds} = -\frac{1}{\pi_s(x)} \nabla_x \phi(x), \tag{10.15}$$

with linearly interpolated PDFs

$$\pi_s = (1 - s)\pi_X^{\text{prior}} + s\pi_X^{\text{post}}, \qquad s \in [0, 1],$$

and the potential ϕ determined by the Poisson equation

$$\nabla_x \cdot (\nabla_x \phi) = -\pi_X^{\text{prior}} + \pi_X^{\text{post}}.$$

The desired transport map T is defined as the time-one flow map of the ODE (10.15). The embedding technique of Moser has been applied and refined for the Bayesian filtering step in sequential data assimilation by [22]. It has been demonstrated by [23] that explicit solutions to the embedding technique (10.15) can be found in case π_X^{prior} is a multivariate Gaussian or a mixture of multivariate Gaussians.

Furthermore, the popular family of EnKFs [9] can be viewed as providing couplings under the assumption that the prior distribution is approximated at t_{j+1} by a multivariate Gaussian with mean $\bar{x}^f$ and covariance matrix P^f. For example,

the EnKF with perturbed observations leads to a non-deterministic (non-optimal) coupling, which gives rise to

$$x_i^a(\mathrm{t}_{j+1}) \sim \pi_Z(\cdot|x_i^f(\mathrm{t}_{j+1})) = \mathrm{N}(x_i^f(\mathrm{t}_{j+1}) - K(Hx_i^f(\mathrm{t}_{j+1}) - \mathrm{y}_{\rm obs}), KRK^T)$$

in (10.10) with Kalman gain matrix

$$K = P^f H^T (HP^f H^T + R)^{-1}.$$

See also the discussion in [6] on an optimal coupling for EnKFs based on the work of [21].

We will now utilize available couplings for Bayesian inference in order to derive GSMC methods for more general classes of prior and posterior distributions.

10.3 A GSMC Method

We assume that, at initial time $\mathrm{t}_0 = 0$, an ensemble of M independent samples $x_i(0) \in \mathbb{R}^N$ from the given PDF $\pi_X(x,0)$ is being generated. Each sample is given an initial weight of $w_i^f(0) = 1/M$.

In between observations, the ensemble is propagated under the dynamical model (10.5). The initial conditions for each simulation interval are provided by the analyzed ensemble members $x_i^a(\mathrm{t}_j)$ from the most current data assimilation step. The model predictions at the next observation point t_{j+1} are denoted by $x_i^f(\mathrm{t}_{j+1})$. The weights $w_i(\mathrm{t}_j)$ do not change during model simulations. Observed values $y(\mathrm{t}_j) \in \mathbb{R}^K$ are assimilated in time intervals of $\Delta t_{\rm obs}$ using the forward model (10.4).

An essential ingredient of the proposed GSMC method is to find an appropriate estimate $\bar{\pi}_X^{\rm prior}(x, \mathrm{t}_{j+1})$ of the prior PDF from the weighted samples $(x_i^f(\mathrm{t}_{j+1}), w_i^f(\mathrm{t}_{j+1}))$, $i = 1, \ldots, M$. For example, the prior distribution at time t_{j+1} can be approximated using a Gaussian mixture or a Gaussian kernel density estimator [26] of the form

$$\bar{\pi}_X^{\rm prior}(x, \mathrm{t}_{j+1}) = \sum_{i=1}^M \frac{w_i^f}{(2\pi h)^{N/2}|P^f|^{1/2}} \exp\left(-\frac{1}{2h}(x - x_i^f)^T (P^f)^{-1}(x - x_i^f)\right),$$

where $x_i^f = x_i^f(\mathrm{t}_{j+1})$, $w_i^f = w_i^f(\mathrm{t}_{j+1})$, P^f denotes the empirical covariance matrix of the forecast ensemble, and $1 \geq h > 0$ is the bandwidth of the estimator. From now on we will drop the time argument and assume that all relevant quantities are computed for $t = \mathrm{t}_{j+1}$ unless indicated otherwise.

Our GSMC approach relies on an appropriate transport map $\hat{T} : \mathbb{R}^N \to \mathbb{R}^N$, which will depend on both the forecast ensemble $\{(x_i^f, w_i^f)\}_{i=1}^M$ and the measured

y_{obs}. It should be chosen such that the transformed posterior distribution $\tilde{\pi}_X^{\text{post}}$, defined according to (10.11) by

$$\tilde{\pi}_X^{\text{post}}(\hat{T}(x))|D\hat{T}(x)| = \bar{\pi}_X^{\text{prior}}(x),$$

is close to the desired posterior distribution

$$\bar{\pi}_X^{\text{post}}(x) := \pi_X(x|y_{\text{obs}}) \propto \pi_Y(y_{\text{obs}}|x)\bar{\pi}_X^{\text{prior}}(x).$$

Following the idea of importance sampling, we now define the analyzed ensemble by $x_i^a = \hat{T}(x_i^f)$ with updated weights

$$\begin{aligned} w_i^a &\propto w_i^f \frac{\pi_Y(y_{\text{obs}}|x_i^a)\,\bar{\pi}_X^{\text{prior}}(x_i^a)}{\tilde{\pi}_X^{\text{post}}(x_i^a)} \\ &= w_i^f\, \pi_Y(y_{\text{obs}}|x_i^a)\,|D\hat{T}(x_i^f)|\,\frac{\bar{\pi}_X^{\text{prior}}(x_i^a)}{\bar{\pi}_X^{\text{prior}}(x_i^f)}, \end{aligned}$$

where the normalization constant is chosen such that $\sum_i w_i^a = 1$.

Furthermore, we will assume that the transport map $\hat{T}$ couples a prior $\hat{\pi}_X^{\text{prior}}$ and its associated posterior PDF $\hat{\pi}_X^{\text{post}}$ exactly. Such transport maps exists for Gaussian prior PDFs as well as Gaussian mixture prior PDFs. The coupling property implies that

$$|D\hat{T}(x)| = \frac{\hat{\pi}_X^{\text{prior}}(x)}{\hat{\pi}_X^{\text{post}}(\hat{T}(x))}$$

and, furthermore, since

$$\hat{\pi}_X^{\text{post}}(\hat{T}(x)) \propto \pi_Y(y_{\text{obs}}|\hat{T}(x))\,\hat{\pi}_X^{\text{prior}}(\hat{T}(x))$$

we may conclude that

$$|D\hat{T}(x)| \propto \frac{\hat{\pi}_X^{\text{prior}}(x)}{\pi_Y(y_{\text{obs}}|\hat{T}(x))\,\hat{\pi}_X^{\text{prior}}(\hat{T}(x))}.$$

Combining all these results leads to the modified filtering step:

(ii) Filtering:

$$x_i^a = \hat{T}(x_i^f), \qquad w_i^a \propto w_i^f\, \frac{\hat{\pi}_X^{\text{prior}}(x_i^f)}{\hat{\pi}_X^{\text{prior}}(x_i^a)}\, \frac{\bar{\pi}_X^{\text{prior}}(x_i^a)}{\bar{\pi}_X^{\text{prior}}(x_i^f)}, \tag{10.16}$$

where the time argument t_{j+1} has been dropped for notational convenience. Here $\bar{\pi}_X^{\rm prior}(x)$ denotes an estimate of the true underlying prior and $\hat{\pi}_X^{\rm prior}(x)$ denotes a PDF which allows for the computation of a transport map $\hat{T}(x)$.

The PDF $\hat{\pi}_X^{\rm prior}(x)$ should be chosen such that particle weights remain nearly uniform. If the accumulated weights become however strongly nonuniform, particles can be resampled by the same techniques as being employed for traditional sequential Monte Carlo methods.

We mention that the proposed GSMC algorithm can be extended to the case that an exact coupling cannot be found for the given likelihood $\pi_Y(y|x)$ and a simplified likelihood $\hat{\pi}_Y(y|x)$ needs be employed. In that case one would use

$$w_i^a \propto w_i^f \, \frac{\pi_Y(y_{\rm obs}|x_i^a)}{\hat{\pi}_Y(y_{\rm obs}|x_i^a)} \, \frac{\hat{\pi}_X^{\rm prior}(x_i^f)}{\hat{\pi}_X^{\rm prior}(x_i^a)} \, \frac{\bar{\pi}_X^{\rm prior}(x_i^a)}{\bar{\pi}_X^{\rm prior}(x_i^f)}$$

in (10.16).

We next provide a simple numerical demonstration of the proposed GSMC method using one-dimensional Brownian dynamics.

10.4 Brownian Dynamics Under a Double Well Potential

We consider one-dimensional Brownian dynamics

$$dx = -V'(x)\, dt + dW(t)$$

with potential

$$V(x) = \cos(x) + \frac{3}{4}\left(\frac{x}{6}\right)^4$$

and standard Brownian motion $W(t)$ as a test example. The stochastic equations are solved numerically by the Euler–Maruyama method, i.e.,

$$x_{n+1} = x_n - \Delta t V'(x_n) + \sqrt{\Delta t} Z_n$$

with step-size $\Delta t = 0.1$ and $Z_n \sim \mathrm{N}(0,1)$.

The measurement equation is

$$Y = X + \sqrt{R}\Xi$$

with $\Xi \sim \mathrm{N}(0,1)$ and $R = 36$ is the variance of the measurement error. Actual measurements are obtained from a reference trajectory (see Fig. 10.1) of our Brownian dynamics model and added measurement noise. Ensemble sizes vary

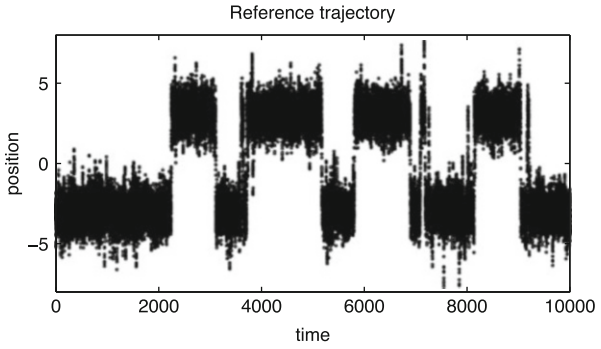

Fig. 10.1 Shown is the reference solution from which observations are generated by adding Gaussian noise with mean zero and variance $R = 36$

between $M = 20, 50$ and 100. Measurements are taken every 10 units of time (i.e., $\Delta t_{\text{obs}} = 10\Delta t$) and a total of 1,000 assimilation steps are performed.

We compute a highly accurate reference solution by solving the Fokker–Planck equation

$$\frac{\partial \pi_X}{\partial t} = \frac{\partial}{\partial x}(\pi_X V') + \frac{1}{2}\frac{\partial^2 \pi_X}{\partial x^2}$$

with stationary PDF

$$\pi_X^*(x) \propto \exp(-2V(x))$$

over a computational grid with mesh-size $\Delta x = 1/16$. The results are used to approximate the prediction step (10.6) directly on the level of PDFs. The grid approximations to $\pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_j)$ are then used to find grid approximation to the analyzed PDFs $\pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_{j+1})$ using (10.7). These density approximations are finally used to approximate the time-evolved means $\bar{x}^{\text{ref}}(t_n)$ which are taken as a reference for the ensemble-based filter algorithms.

Given a set of particles $x_i^f \in \mathbb{R}$ with weights $w_i > 0$ we compute the unweighted ensemble mean

$$\bar{x}^f = \frac{1}{M}\sum_{i=1}^{M} x_i^f$$

and the corresponding covariance matrix

$$P^f = \frac{1}{M-1}\sum_{i=1}^{M}(x_i^f - \bar{x}^f)^2,$$

Table 10.1 RMS errors for ensemble means obtained from an ensemble square root filter (EnKF), the ensemble Gaussian mixture filter (EGMF), the guided sequential Monte Carlo method (GSMC), and the rank histogram filter (RHF) compared to the expected value computed by a Fokker–Planck discretization with error variance $R = 36$ and ensemble sizes of $M = 20, 50, 100$ particles/ensemble members

	EnKF	EGMF	GSMC	RHF
$M = 20$	1.1590	0.7683	1.0200	0.6551
$M = 50$	1.0701	0.5127	0.7172	0.3717
$M = 100$	1.0477	0.4033	0.6534	0.2691

which implies a Gaussian prior

$$\hat{\pi}_X^{\rm prior}(x) = \frac{1}{(2\pi P^f)^{1/2}} \exp\left(-\frac{1}{2P^f}(x-\bar{x}^f)^2\right)$$

for the application of the ensemble square root filter as a proposal map $\hat{T}(x)$.

For the implementation of the GSMC approach we assume furthermore that particles are given an index $\alpha_i \in \{-1, 1\}$ indicating whether they belong to the left or the right, respectively, potential well, i.e., $\alpha_i = +1$ if $x_i^f > 0$ and $\alpha_i = -1$ if $x_i^f < 0$. We then approximate the prior distribution $\pi_X(x, \mathrm{t}_{j+1}|\mathrm{Y}_j)$ by

$$\bar{\pi}_X^{\rm prior}(x) = \frac{\gamma_{-1}}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\bar{x}_{-1})^2\right) + \frac{\gamma_{+1}}{(2\pi)^{1/2}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\bar{x}_{+1})^2\right)$$

with

$$\gamma \mathrm{t}_j = \sum_{i=1}^M \delta_{\alpha,\alpha_i} w_i, \qquad \alpha \in \{-1, +1\},$$

where δ_{jk} denotes the Kronecker delta,

$$\bar{x}_\alpha = \sum_{i=1}^M \delta_{\alpha,\alpha_i} w_i x_i^f$$

and fixed standard deviation $\sigma = \sqrt{2}$. Clearly, a more sophisticated Gaussian mixture model could be fitted to the available ensembles using the expectation-maximization (EM) algorithm [7].

We also implemented the RHF of [1] using piecewise linear ensemble-based approximations to the prior and posterior cumulative distribution functions $F_{\rm prior}$, $F_{\rm post}$ and subsequent construction of a transport map for (10.12) using formula (10.14).

Numerical results are presented in Table 10.1, where the root mean square (RMS) error is defined as

$$\text{RMS error} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(\bar{x}^a(\mathrm{t}_j) - \bar{x}^{\mathrm{ref}}(\mathrm{t}_j))^2},$$

with $\bar{x}^a(\mathrm{t}_j)$ denoting the analyzed ensemble average from the filter at time t_j and $\bar{x}^{\mathrm{ref}}(\mathrm{t}_j)$ its numerical approximation from the Fokker–Planck approach.

The RHF with a transport map based on cumulative distribution functions yields the most accurate filter results. In fact, the RHF converges to the analytic filtering solution as $M \to \infty$. The second best result is obtained for the ensemble Gaussian mixture filter (EGMF) of [23] for which a transport map is constructed using a binary Gaussian mixture approximation for the distributions in x using the EM algorithm. We also find that the GSMC approach yields an improvement over the EnKF while not delivering results as accurate as those from the RHF and the EGMF. Note that the GSMC implementation with $\hat{T}$ based on a ensemble square root filter can be viewed as an inexpensive post-processing step to the associated EnKF algorithm.

10.5 Conclusions

From a mathematical perspective the coupling and optimal transportation approach to Bayesian data assimilation offers attractive opportunities. Practical implementations are limited by the fact that optimal transport maps are difficult to compute numerically. It has, however, been demonstrated for a wide range of problems that rather crude approximations to the coupling/transport problem, such as the EnKF, can lead to surprisingly robust data assimilation algorithms. In this paper, we have followed a recent trend of combining crude ensemble transform approximations with sequential Monte Carlo methods. More specifically, we have proposed an importance sampling approach for post-processing existing ensemble transform filter formulations. Such an approach should be useful whenever the underlying ensemble transform filter is capable of tracking regions of high posterior probability while being of limited statistical accuracy. Under those circumstances, post-processing the data should lead to improved statistics.

References

1. Anderson, J.: A non-Gaussian ensemble filter update for data assimilation. Monthly Weather Rev. **138**, 4186–4198 (2010)
2. Bain, A., Crisan, D.: Fundamentals of stochastic filtering. In: Stochastic Modelling and Applied Probability, vol. 60. Springer, New York (2009)
3. Bengtsson, T., Bickel, P., Li, B.: Curse-of-dimensionality revisited: collapse of the particle filter in very large scale systems. In: Nolan, D., Speed, T. (eds.) Probability and Statistics: Essays in

Honor of David A. Freedman, pp. 316–334. Institute of Mathematical Statistics, Beachwood (2008)
4. Bocquet, M., Pires, C., Wu, L.: Beyond Gaussian statistical modeling in geophysical data assimilation. Monthly Weather Rev. **138**, 2997–3022 (2010)
5. Chorin, A., Morzfeld, M., Tu, X.: Implicit filters for data assimilation. Commun. Appl. Math. Comput. Sci. **5**, 221–240 (2010)
6. Cotter, C., Reich, S.: Ensemble filter techniques for intermittent data assimilation—a survey. In: Engl, H.W. et al. (eds.) Radon Series on Computational and Applied Mathematics. De Gruyter, Boston (2013, in press)
7. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39B**, 1–38 (1977)
8. Doucet, A., de Freitas, N. (eds.) N.G.: Sequential Monte Carlo Methods in Practice. Springer, Berlin (2001)
9. Evensen, G.: Data Assimilation. The Ensemble Kalman Filter. Springer, New York (2006)
10. Gardiner, C.: Handbook on Stochastic Methods, 3rd edn. Springer, New York (2004)
11. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning, 2nd edn. Springer, New York (2009)
12. Jazwinski, A.: Stochastic Processes and Filtering Theory. Academic, New York (1970)
13. Kloeden, P., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1992)
14. Künsch, H.: Recursive Monte Carlo filter: algorithms and theoretical analysis. Ann. Stat. **33**, 1983–2021 (2005)
15. Leeuwen, P.V.: Nonlinear data assimilation in the geosciences: an extremely efficient particle filter. Q. J. R. Meteorolog. Soc. **136**, 1991–1996 (2010)
16. Lei, J., Bickel, P.: A moment matching ensemble filter for nonlinear and non-Gaussian data assimilation. Monthly Weather Rev. **139**, 3964–3973 (2011)
17. del Moral, P.: Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications. Springer, New York (2004)
18. Morzfeld, M., Tu, X., Atkins, E., Chorin, A.: A random map implementation of implicit filters. J. Comput. Phys. **231**, 2049–2066 (2012)
19. Moselhy, T.E., Marzouk, Y.: Bayesian inference with optimal maps. J. Comput. Phys. **231**, 7815–7850 (2012)
20. Moser, J.: On the volume elements on a manifold. Trans. Am. Math. Soc. **120**, 286–294 (1965)
21. Olkin, I., Pukelsheim, F.: The distance between two random vectors with given dispersion matrices. Linear Algebra Appl. **48**, 257–263 (1982)
22. Reich, S.: A dynamical systems framework for intermittent data assimilation. BIT Numer. Math. **51**, 235–249 (2011)
23. Reich, S.: A Gaussian mixture ensemble transform filter. Q. J. R. Meterolog. Soc. **138**, 222–233 (2012)
24. Reich, S.: A non-parametric ensemble transform method for Bayesian inference. Tech. Rep., Potsdam University (2012). SIAM Journal Scientific Computing
25. Villani, C.: Optimal Transportation: Old and New. Springer, Berlin Heidelberg (2009)
26. Wand, M., Jones, M.: Kernel Smoothing. Chapmann and Hall, London (1995)

Chapter 11
Deterministic and Stochastic Dynamics of Chronic Myelogenous Leukaemia Stem Cells Subject to Hill-Function-Like Signaling

Tor Flå, Florian Rupp, and Clemens Woywod

Dedicated to Jürgen Scheurle on the occasion of his 60th birthday

Abstract Based on a discrete Markovian birth-death model including regulated symmetric and asymmetric cell division, we formulate a continuous four-dimensional stochastic (ordinary) differential equation model for the dynamics of Chronic Myelogenous Leukaemia (CML) stem cells in a bone marrow niche involving signaling and competition between active stem cells. Invoking stochastic-deterministic correspondence we then investigate two deterministic subsystems: (a) the competition between active normal and wild-type CML stem cells or also between two developing leukaemic stem cell strains is represented by a two-dimensional equation system, and (b) a three-dimensional model involving both cycling and noncycling normal stem cells as well as cycling wild-type CML stem cells is defined. The four-dimensional equation system finally includes in addition one cycling CML stem cell clone of an anti-CML-drug-resistant mutant. By totally analytic means we discuss the existence and stability of the equilibria of the three systems in the deterministic small noise limit, and establish, by numerical means, connections between these classical results and the original

T. Flå
Department of Mathematics and Statistics, University of Tromsø, 9037 Tromsø, Norway
e-mail: tor.fla@uit.no

F. Rupp
Lehrstuhl für Höhere Mathematik und Analytische Mechanik, Technische Universität München, Fakultät für Mathematik, 85747 Garching, Germany
e-mail: rupp@ma.tum.de

C. Woywod (✉)
Centre for Theoretical and Computational Chemistry (CTCC), Department of Chemistry, University of Tromsø, 9037 Tromsø, Norway
e-mail: clemens.woywod@ch.tum.de

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_11,

stochastic setting. The robust, stable finite population equilibria can be interpreted as homeostatic equilibria of normal and leukaemic stem cell populations, in the case of the four-dimensional model for the scenario of treatment of the wild-type CML clone with a CML suppressing agent, e.g., *imatinib*, which leads to the emergence of a resistant CML strain. The four-dimensional model thus represents a common clinical picture.

11.1 Introduction

One of the major challenges and driving forces of present day medicine is the quest for an understanding of the mechanisms that cause and promote cancer on a cellular level in order to design effective treatment strategies. From a theoretical perspective, mathematical biology and in particular the theory of dynamical systems are playing a key role for the simulation of tumor formation and development.

Chronic myelogenous leukaemia (CML) is the most common cancer of the hematopoietic system, and the relatively simple and well-understood pattern of origin and effects on the molecular level makes CML a very interesting target for both experimentalists and theoreticians.

One assumes that normal cells residing in a stem cell niche mutate by an unregulated crossing-over of certain chromosomes during cell division into wild-type leukaemic cells. Like normal stem cells, CML stem cells differentiate via asymmetric division and produce progenitor cells which are able to leave the stem cell niche and, through the process of differentiation, eventually lead to the observed cancer symptoms. More precisely, CML is generated by mutated hematopoietic stem cells which establish a hierarchy of progenitor and differentiated blood cells with a cancerous phenotype. CML cells are characterized by a reciprocal translocation of chromosomes 9 and 22. The shortened chromosome 22 carries the BCR-ABL1 fusion gene. The corresponding BCR-ABL1 protein is oncogenic and is a constitutively activated tyrosine kinase which is responsible for CML. The agent *imatinib* inhibits the BCR-ABL1 tyrosine kinase by occupying the ATP-binding site and prevents phosphorylation of its substrates [6]. Subsequently, the downstream signaling pathways leading to leukogenesis are switched off. *Imatinib* selectively acts on leukaemia cells by inducing a proliferation inhibiting effect and an increase in the apoptotic rate of actively proliferating cells, cf. [11, 31, 35, 43, 52]. Unfortunately, *imatinib*-resistant cancer clones may appear in a subgroup of CML patients, see [20, 21, 53].

Several models for the description of CML have been proposed, see e.g. [2, 8, 30, 31, 33–35, 38, 41–43, 52] and the literature therein. More generally models for proliferation of metastases are given in, e.g., [7, 15], and for the signaling between cells, which is considered a crucial part when it comes to understanding tumors, in [16].

Michor et al. [11, 38] suggested a CML model including stem cells, progenitors, and differentiated cells in which *imatinib* treatment does not affect the stem cell

compartment. By assuming a sufficiently strong *imatinib* repression of the growth rate of downstream cells, it was found that measurements of BCR-ABL1 transcripts in the blood following *imatinib* treatment could be interpreted by a rapid initial decay of differentiated leukaemic cells succeeded by a slower decay of leukaemic progenitors. Later Roeder et al. [42, 43] have shown that clinical data showing a biphasic decline of BCR-ABL1 transcripts as well as the rapid relapse of BCR-ABL1 levels at treatment cessation can alternatively be explained by a model including proliferating stem cells and quiescent stem cells.

Their CML model explains the biphasic decline of BCR-ABL1 transcripts by competition for stem cell niches and implies that only quiescent CML stem cells are unaffected by the drug. The Roeder model has a state space of a continuous activation parameter and it has recently been shown to be equivalent to a PDE transport model in this state space [41].

We will invoke a simplified picture of stem cell dynamics similar to [31–33, 35, 52] considering only cycling (i.e., stem cells that are actively participating in the cell division cycle) and quiescent (= non-cycling) stem cells (i.e., stem cells that have reversibly entered an inactive mode) explicitly in the model. Progenitor and differentiated cell compartments are not explicitly taken into account.

The brief overview of computational models that have been developed for the description of CML dynamics shows that both deterministic and stochastic formalisms have been employed. Today, the study of stochastic-deterministic correspondences in biological systems is an active field of research, see e.g. [1, 13]. Although we are inherently dealing with a stochastic system, information that may be valuable for the stochastic approach can be gained by a precise knowledge of the underlying deterministic system and in particular of its equilibria.

The stochastic dynamics can, e.g., be formulated as an Itô stochastic differential/ Langevin equation. At an equilibrium point of the corresponding deterministic system the drift term of the stochastic formulation vanishes and only the pure diffusion part survives. In particular, the trajectories representing solutions of an Itô stochastic differential equation will converge toward the location of an asymptotically stable equilibrium of the deterministic system if the diffusive influence can be assumed to be small. Similarly, the trapping of stochastic trajectories at (stable) heteroclinic orbits or the avoidance of regions near unstable deterministic equilibria can be observed. For instance, in the view of a corresponding Fokker–Planck equation this leads to local density maxima at deterministic asymptotically stable equilibria.

In this work, we have not considered the constitutive, deregulated symmetric division of cancer stem cells as described, e.g., in [31–35, 52] since we are not presently interested in modeling the blast phase. Instead, we investigate the dynamics of the chronic phase induced by the wild type and, in one case, additionally by one *imatinib*-resistant CML clone, but still dominated by normal stem cells, progenitors, and differentiated cells close to homeostatic equilibrium.

The compartments of progenitor and differentiated cells, both normal and leukaemic, are not explicitly included in the dynamical models. However, the regulatory feedback effects of the downstream populations are implicitly taken

into account in our stem cell models through appropriate competition parameters. This concept is based on the idea that the time development of progenitors with sufficiently fast decay and differentiated cells are in a quasi-static equilibrium on the slow timescale of stem cell dynamics.

This chapter is structured as follows: in Sect. 11.2 we define a probabilistic four-dimensional model (model C) of the stem cell system consisting of cycling and noncycling normal stem cells as well as cycling wild-type and *imatinib*-resistant CML stem cells. Starting from first principles, i.e., Markov transition probabilities, we derive an approximated Fokker–Planck equation that leads to a stochastic differential model equation system. Based on a biologically motivated scaling argument, we infer that the diffusion part of this equation is rather small and study the dynamics of the deterministic drift part in depth. This is done in Sect. 11.3, where we determine, by totally analytic means, the location of all 13 equilibrium structures in the full four-dimensional deterministic setting. Moreover, specific attention is paid to a two-dimensional (model A, representing competing active normal and wild-type CML stem cells or alternatively two competing leukaemic strains) and a three-dimensional (model B, representing cycling and noncycling normal stem cells plus cycling wild-type CML stem cells) deterministic subsystem for which we study the stability of all equilibria and the complete bifurcation behavior. The two-dimensional subsystem turns out to be completely governed by the dynamics induced by its equilibria since no admissible values of the parameters for periodic orbits exist. We further establish, numerically, connections between the classical results and the original stochastic setting. Finally, Sect. 11.4 summarizes our results.

11.2 Definition of the Governing Probabilistic Four-Dimensional Model (Model C)

11.2.1 Biological Aspects of the Model

We discuss the dynamics of active and quiescent cell types in a stem cell niche including signaling as well as competition effects. In our approach, the growth of the stem cell populations in the bone marrow niche by symmetric division is regulated by negative feedback proportional to the size of the total stem cell population. Differentiation and self-renewal of stem cells are realized by asymmetric division, and the expansion of stem cell pools is balanced by loss of stem cells mainly by symmetric division to progenitors which finally proliferate to differentiated cells, cf. [18, 19, 45]. Stem cell numbers are also controlled by apoptosis and, here only implemented for normal stem cells, by the fraction of cells that enter the cell cycle from a large normal stem cell population in G_0 or quiescent state. It should be mentioned here that leukaemic Ph^+/BCR-ABL1 stem cells in G_0 mode are hypothesized to be capable of generating leukaemic proliferation in a reversible way, but this issue will be addressed in separate studies.

We basically have a birth–death model characterized by a regulated and an unregulated net growth/death rate.

It is straightforward to show for the effectively autoregulatory case that homeostatic, stable, finite stem cell equilibria are potentially obtained if the net death (−net growth) rate is positive with negative feedback regulation of the stem cell division. It is also characteristic for the regime of autoregulation that the corresponding zero population equilibrium is unstable (stable) if the total death rate at this equilibrium is positive (negative) with a bifurcation point for the existence of homeostatic equilibria at the state where the total growth rate is zero. In the autoregulatory case, the existence of homeostatic CML population equilibria and the dynamics toward the equilibria from newly mutated, small cancer populations can be determined in the same way as if one or several independent homeostatic, finite, stable CML stem cell populations were present.

We also notice that without the regulated part of the rate, homeostatic equilibria can only be found at the bifurcation point associated with a zero net death rate since in this case any population represents a marginally stable equilibrium. For any nonzero death rate only the zero population equilibrium exists. Within the scenario of autoregulation, a successful CML treatment is then distinguished by modifying the total growth rate at the zero cancer population equilibrium in a way that the population dynamics regime changes from instability to stability. The desired effect of drug administration is obviously to achieve a switch of the total CML growth rate from positive to negative. In principle, this goal can be achieved by (i) either decreasing the regulated part or (ii) by increasing the net death rate via reduction of the unregulated growth rate or enlargement of the unregulated death rate.

The inclusion of a pool of quiescent stem cells, which are supposed to be in a continuous exchange with the actively dividing stem cells, in the model leads to the same zero population stability regimes, with the only difference that the population dynamics generically displays a biphasic behavior with an initial phase characterized by large decay and growth rates followed by slow asymptotic convergence to stable population limits on the timescale of activation of quiescent stem cells.

In this chapter, we investigate the stability and bifurcations of equilibria (1) of cycling normal stem cells contending with wild-type CML stem cells for resources (model A, Sect. 11.3.1, this model can also be interpreted as competition between two leukaemic stem cell species) and (2) of normal stem cells in a competition regulated by negative feedback with a wild-type leukaemic stem cell strain (model B, Sect. 11.3.2) and up to one *imatinib*-resistant mutant clone (model C, Sect. 11.3.3). We consider the processes of deactivation of cycling normal stem cells and of activation of quiescent normal stem cells. The model consists of two major components:

1. A threshold-controlled protein-signal model featuring a sigmoidal regulatory function that realizes negative feedback by in principle all cell types on the growth of any given stem cell clone. The protein signal regulating the symmetric stem cell division process is here simply assumed to be proportional

to the number of cells of different types. The contributions of the individual cell types to the regulatory switch are determined by weight factors.

The population threshold applicable to an individual switch is assumed to be fixed and equal to an effective stem cell population N_e. N_e is related to the average inverse population scale by the expression $N_e = \langle N^{-1} \rangle^{-1}$. The average is determined with respect to the stem cell niche environment and geometry. This parameter for stem cell growth regulation is based on previous work by Kimura [24, 25] and Waxman [50].

A Hill-type sigmoidal function has been selected for the model presented in this work. This version of a power function switch is commonly used to describe ultra-sensitive switching in certain signaling networks and to approximate Michaelis–Menten functions in enzyme chemistry. In our study, the sigmoidal function is employed for the description of the input/output response in an intra- and inter-cell network that is connected by protein signaling and regulates cell division and cell death processes.

Recent motivation for selecting Hill-type functions for the implementation of stem cell decision models is coming from work on the regulation of gene expression by cell-population density variations ("quorum sensing") [46, 48] and, on the molecular level, by mechanisms that control the interaction of transcription factors and operon binding sites [3, 5, 22, 36, 39, 40, 47].

Similar logistic cell decision functions have been defined not only for the regulation of stem cell reproduction and gene expression, but also in the contexts of other biological network models that can be described by the concept of funneling landscapes of functional activity [9, 36]. The interacting constituents might be amino acids for protein folding, base pairs for gene expression, units of regulated genes with operons for protein networks, modules of genes/proteins for pathways, and modules of pathways for stem cells in their niches [48, 49].

2. Knowledge of the fundamental symmetric and asymmetric stem cell division mechanisms suggests the adoption of an underlying birth–death Markov process and also of a Fokker–Planck equation describing the corresponding Langevin stochastic process from first principles. The Fokker–Planck formalism has been implemented numerically and we compare the stochastic sampling paths to the corresponding deterministic paths.

   The primary mutation leading to the wild-type CML phenotype as well as the pretreatment phase are assumed to be complete and are parameterized via the initial conditions of the normal and CML clones. While suppression of the wild-type CML variant by *imatinib* administration is known to be effective, *imatinib*-resistant CML mutations may develop during the phase of *imatinib* application.

   The bifurcations and stability diagrams of the resulting multi-population equilibria in the different regulation regimes can be interpreted as deformations of the simple, independent autoregulatory and nonregulated homeostatic equilibria discussed above. We note that modified treatment conditions and sub-equilibria of the multi-clone stem cell equilibria, including normal stem cells as well as wild-type and *imatinib*-resistant CML stem cells, need to be classified

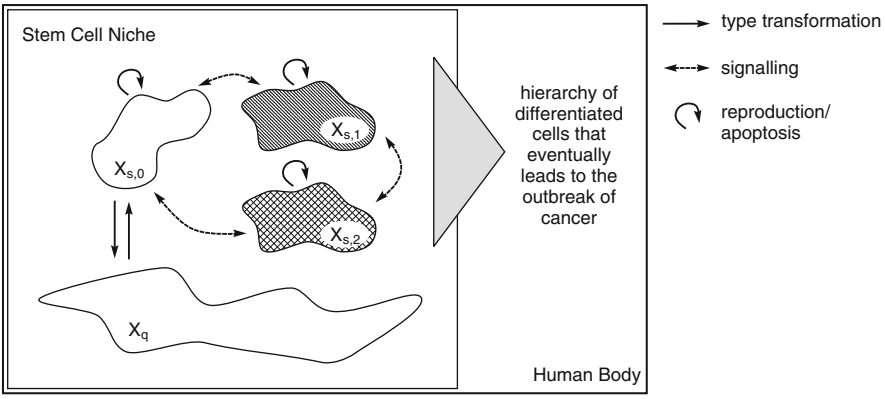

Fig. 11.1 Illustration of the four-species model (model C) for leukaemic stem cell dynamics. Sketched are the different stem cell types in a niche as considered in this chapter and their interactions together with an outlook on how the dynamics in a stem cell niche governs the actual dynamics of normal and leukaemic stem cells in the (human) body

on a personal medication level including stochastic theory taking mutations of the wild-type CML strain into account. Such a classification and the derivation of the biological implications, i.e., of effective stem cell growth, death, and mutation rates, need to be performed in order to reliably predict personalized treatment strategies for scenarios involving drug-resistant CML clones.

11.2.2 Formulation of the Building Blocks of Model C

Specifically, we identify the cycling/active and noncycling/quiescent populations of normal stem cells by $x_{s,0}$ and x_q, respectively. The cycling/active wild-type and *imatinib*-resistant CML clones are denoted by the variables $x_{s,1}$ and $x_{s,2}$, respectively, cf. Fig. 11.1.

By means of first order Markov birth and death processes these types of cells have the following properties:

Let $\mathbb{P}(m, \tau | n, t)$ denote the conditional probability of being in state m at time τ given state n at time t. At each time step $t + \Delta t$ the number of normal active stem cells $n_{s,0}$ can grow in two ways: First, due to activation of quiescent stem cells [10]:

$$\begin{aligned}&\mathbb{P}\left(n_{s,0}+1, n_{s,1}, n_{s,2}, n_q - 1, t + \Delta t \,|\, n_{s,0}, n_{s,1}, n_{s,2}, n_q, t\right) \;=:\; t^{+-}_{0,q}(n)\\ &= \Delta t \cdot \alpha_0 \cdot n_q + o(\Delta t)\,,\end{aligned}$$

where $o(\Delta t)$ means that $o(\Delta t)/\Delta t \to 0$ as $\Delta t \to 0$, $\alpha_0 > 0$ is a proportionality factor and $n_{s,i}$, n_q are population numbers of leukaemic stem cells $x_{s,i}$ $(i = 1, 2)$

and of quiescent normal stem cells x_q, respectively. The four population variables are collected in the vector $n := (n_{s,0}, n_{s,1}, n_{s,2}, n_q)^T$.

Second, normal cycling stem cells can proliferate as a result of cell divisions:

$$\begin{aligned}&\mathbb{P}\left(n_{s,0}+1, n_{s,1}, n_{s,2}, n_q, t+\Delta t \,|\, n_{s,0}, n_{s,1}, n_{s,2}, n_q, t\right) \;=:\; t_0^+(n)\\ &= \Delta t \cdot \left(g_0 + \frac{l_0}{1+(w_{0,0}n_{s,0}+w_{0,1}n_{s,1}+w_{0,2}n_{s,2})^m}\right) n_{s,0} + o(\Delta t)\,,\end{aligned}$$

with proportionality factor $l_0 > 0$ and $g_0 \geq 0$. The cell divisions are regulated by a Hill function of odd index $m \in \mathbb{N}/2\mathbb{N}$ with signaling weights $w_{0,i}$ $(i = 0, 1, 2)$ that take the existence of further active normal and leukaemic stem cells into account[1]. Without loss of generality we can scale the "relative" signaling weights that influence the active stem cell population such that $w_{0,0} = 1$.

Active normal stem cells can become quiescent with a proportionality factor $\beta_0 > 0$, i.e., [10]

$$\begin{aligned}&\mathbb{P}\left(n_{s,0}-1, n_{s,1}, n_{s,2}, n_q+1, t+\Delta t \,|\, n_{s,0}, n_{s,1}, n_{s,2}, n_q, t\right) \;=:\; t_{0,q}^{-+}(n)\\ &= \Delta t \cdot \beta_0 \cdot n_{s,0} + o(\Delta t)\,.\end{aligned}$$

For the cancer-causing stem cells $x_{s,i}$ we assume that these can just reproduce with proportionality factors l_i $(i = 1, 2)$ and not mutate a second time or interact with the quiescent population of normal stem cells, i.e.,

$$\begin{aligned}&\mathbb{P}\left(n_{s,0}, n_{s,1}+1, n_{s,2}, n_q, t+\Delta t \,|\, n_{s,0}, n_{s,1}, n_{s,2}, n_q, t\right) \;=:\; t_1^+(n)\\ &= \Delta t \cdot \left(g_1 + \frac{l_1}{1+(w_{1,0}n_{s,0}+w_{1,1}n_{s,1}+w_{1,2}n_{s,2})^m}\right) n_{s,1} + o(\Delta t)\,,\end{aligned}$$

and

$$\begin{aligned}&\mathbb{P}\left(n_{s,0}, n_{s,1}, n_{s,2}+1, n_q, t+\Delta t \,|\, n_{s,0}, n_{s,1}, n_{s,2}, n_q, t\right) \;=:\; t_2^+(n)\\ &= \Delta t \cdot \left(g_2 + \frac{l_2}{1+(w_{2,0}n_{s,0}+w_{2,1}n_{s,1}+w_{2,2}n_{s,2})^m}\right) n_{s,2} + o(\Delta t)\,,\end{aligned}$$

[1] As we will show in an upcoming article, an even more precise model for the dependence of signaling on stem cell populations would be given by so-called Tsallis functions as they fulfill some important limits better than Hill functions. However, when it comes to an interpretation of the Hill functions, the only difference up to some scales which can be absorbed in the populations is, in the Tsallis-function case, a fixed positive shift in the argument of the respective power function. This means that the effect of the regulation does not kick in before a predefined population size is reached. In this sense, the Hill function can be viewed as a limit of the more exact Tsallis expression.

with $g_1, g_2 \geq 0$ and signaling weights $w_{1,i}, w_{2,i} > 0$ ($i = 0, 1, 2$). Again, we assume $w_{1,1} = w_{2,2} = 1$.

Moreover, we assume that the active stem cells may die with a proportionality factor $\hat{d}_i > 0$ ($i = 0, 1, 2$), i.e.,

$$\mathbb{P}\left(n_{s,i} - 1, t + \Delta t \,|\, n_{s,i}, t\right) \;=\; \Delta t \cdot \hat{d}_i \cdot n_{s,i} + o(\Delta t) \;=:\; t_i^-(n)\,, \quad \text{for } i = 0, 1, 2\,,$$

as well as

$$\mathbb{P}\left(n_{s,0} \,|\, n_{s,0}, t\right) \;=\; 1 - \left(\sum_{i=0}^{2} \left(t_i^+(n) + t_i^-(n) \right) + t_{0,q}^{+-}(n) + t_{0,q}^{-+}(n) \right) .$$

Note, that Δt is chosen such that the respective number of cells can increase or decrease by just one unit.

These first order birth and death Markov processes form the building blocks of an approximate Fokker–Planck / Kolmogorov forward equation that describes the evolution of the continuous probability densities for the involved stem cell types.

11.2.3 The Approximate Fokker-Planck Equation for Model C

The Chapman–Kolmogorov transition equation between two time steps t and $t + \Delta t$ for a discrete population density $p(n, t) : \mathbb{N}^4 \times \mathbb{R} \to \mathbb{R}$ is given by

$$p(n, t + \Delta t) \;=\; \sum_{n'} \mathbb{P}(n, t + \Delta t | n', t) p(n', t)\,.$$

Let the canonical basis vectors $\mathfrak{e}_1, \mathfrak{e}_2, \mathfrak{e}_3, \mathfrak{e}_4$ in $\mathbb{R}^4$ be denoted by $1_0 := \mathfrak{e}_1$, $1_1 := \mathfrak{e}_2$, $1_2 := \mathfrak{e}_3$ and $1_q := \mathfrak{e}_4$. Then, in particular, we have:

$$\begin{aligned}
&p(n, t + \Delta t) \\
&= \left(1 - \left(\textstyle\sum_{i=0}^{2} \left(t_i^+(n) + t_i^-(n) \right) + t_{0,q}^{+-}(n) + t_{0,q}^{-+}(n) \right) \right) \cdot p(n, t) \\
&\quad + \textstyle\sum_{i=0}^{2} t_i^+(n - 1_i) \cdot p(n - 1_i, t) + \sum_{i=0}^{2} t_i^-(n + 1_i) \cdot p(n + 1_i, t) \\
&\quad + t_{0,q}^{+-}(n - 1_0 - 1_q) \cdot p(n - 1_0 - 1_q, t) + t_{0,q}^{-+}(n + 1_0 + 1_q) \cdot p(n + 1_0 + 1_q, t)\,.
\end{aligned}$$

To come from discrete population sizes n to continuous population sizes x, we introduce the scaling $x = n/N_e$ for some natural effective population scale N_e such that $\mathbb{E}(N(t))/N_e = \mathscr{O}(1)$ and $\mathbb{V}ar(N(t))/N_e \leq \mathscr{O}(1)$ for the total population size $N(t) = \sum_{i=0}^{2} n_{s,i} + n_q$.

Again, let 1_i be the ith canonical basis vector in $\mathbb{R}^d$, and Δx a small scalar positive quantity. For two analytic functions $a, b : \mathbb{R}^d \to \mathbb{R}$ with Hesse-matrices H_a and H_b one can show, by expansion in a Taylor-series, that

$$
\begin{aligned}
&(1 - a(x))b(x) + a(x + \Delta x 1_i)b(x + \Delta x 1_i) \\
&= b(x) - a(x)b(x) + \left(a(x) + \nabla a(x)^T 1_i \Delta x + \tfrac{1}{2}\Delta x 1_i^T H_a(x) 1_i \Delta x + o(\Delta x^3)\right) \\
&\quad \cdot \left(b(x) + \nabla b(x)^T 1_i \Delta x + \tfrac{1}{2}\Delta x 1_i^T H_b(x) 1_i \Delta x + o(\Delta x^3)\right) \\
&= b(x) + \left(b(x)\nabla a(x)^T + a(x)\nabla b(x)^T\right) 1_i \Delta x \\
&\quad + \tfrac{1}{2}\left(b(x) 1_i^T H_a(x) 1_i + 2\nabla a(x)^T 1_i \cdot \nabla b(x)^T 1_i + a(x) 1_i^T H_b(x) 1_i\right) \Delta x^2 \\
&\quad + o(\Delta x^3) \\
&= b(x) + \nabla(a \cdot b)(x)^T 1_i \Delta x + \tfrac{1}{2}\Delta x 1_i^T H_{a\cdot b}(x) 1_i \Delta x + o(\Delta x^3) \\
&= b(x) + \partial_{x_i}(a \cdot b)(x)\Delta x + \tfrac{1}{2}\partial_{x_i}\partial_{x_i}(a \cdot b)(x)\Delta x^2 + o(\Delta x^3)\,,
\end{aligned}
$$

and

$$
\begin{aligned}
&-a(x)b(x) + a(x - \Delta x 1_i)b(x - \Delta x 1_i) \\
&= -\partial_{x_i}(a \cdot b)(x)\Delta x + \tfrac{1}{2}\partial_{x_i}\partial_{x_i}(a \cdot b)(x)\Delta x^2 + o(\Delta x^3)\,.
\end{aligned}
$$

Analogously, we get for a two-point change, associated with $t_{0,q}^{+-}$ or $t_{0,q}^{-+}$, by first expanding with respect to $\Delta x 1_j$ and then with respect to $\Delta x 1_i$ $(i \neq j)$

$$
\begin{aligned}
&-a(x)b(x) + a(x + \Delta x 1_i - \Delta x 1_j)b(x + \Delta x 1_i - \Delta x 1_j) \\
&= \left(\partial_{x_i} - \partial_{x_j}\right)(a \cdot b)(x)\Delta x + \tfrac{1}{2}\left(\partial_{x_i} - \partial_{x_j}\right)^2 (a \cdot b)(x)\Delta x^2 + o(\Delta x^3)\,,
\end{aligned}
$$

where $\left(\partial_{x_i} - \partial_{x_j}\right)^2 = \partial_{x_i}\partial_{x_i} - 2\partial_{x_i}\partial_{x_j} + \partial_{x_j}\partial_{x_j}$, together with

$$
\begin{aligned}
&-a(x)b(x) + a(x - \Delta x 1_i + \Delta x 1_j)b(x - \Delta x 1_i + \Delta x 1_j) \\
&= \left(-\partial_{x_i} + \partial_{x_j}\right)(a \cdot b)(x)\Delta x + \tfrac{1}{2}\left(\partial_{x_i} - \partial_{x_j}\right)^2 (a \cdot b)(x)\Delta x^2 + o(\Delta x^3)\,.
\end{aligned}
$$

With $\Delta x = N_e^{-1}$, we thus obtain the state-continuous, time-discrete model

$$
\begin{aligned}
\frac{p(x, t + \Delta t) - p(x, t)}{\Delta t} = &- \sum_{i=1}^{4} \partial_{x_i}\left(V_i(x) \cdot p(x, t)\right)^T \Delta x \\
&+ \tfrac{1}{2N_e} \sum_{i,j=1}^{4} \partial_{x_i}\partial_{x_j}\left(G_{ij}(x) p(x, t)\right) \Delta x^2 + o(\Delta x^3)\,,
\end{aligned}
$$

where for $x := (x_{s,0}, x_{s,1}, x_{s,2}x_q)^T$ the vector-valued function $V : \mathbb{R}^4 \to \mathbb{R}^4$ with rescaled weights is given as

$$V(x) \;=\; \begin{pmatrix} \left(\frac{l_0}{1+(x_{s,0}+w_{0,1}x_{s,1}+w_{0,2}x_{s,2})^m} - d_0 - \beta_0\right) x_{s,0} + \alpha_0 x_q \\ \left(\frac{l_1}{1+(x_{s,1}+w_{1,0}x_{s,0}+w_{1,2}x_{s,2})^m} - d_1\right) x_{s,1} \\ \left(\frac{l_2}{1+(x_{s,2}+w_{2,0}x_{s,0}+w_{2,1}x_{s,1})^m} - d_2\right) x_{s,2} \\ -\alpha_0 x_q + \beta_0 x_{s,0} \end{pmatrix}, \quad (11.1)$$

with $d_i := \hat{d}_i - g_i$, $i = 0, 1, 2$, and the symmetric matrix $G = (\gamma_{i,j})_{i,j=1,\ldots,4} \in \mathbb{R}^{4\times 4}$can be written according to

$$\begin{cases} \gamma_{1,1} = \left(\frac{l_0}{1+(x_{s,0}+w_{0,1}x_{s,1}+w_{0,2}x_{s,2})^m} + \hat{d}_0 + g_0 + \beta_0\right) x_{s,0} + \alpha_0 x_q \\ \gamma_{2,2} = \left(\frac{l_1}{1+(x_{s,1}+w_{1,0}x_{s,0}+w_{1,2}x_{s,2})^m} + \hat{d}_1 + g_1\right) x_{s,1} \\ \gamma_{3,3} = \left(\frac{l_2}{1+(x_{s,2}+w_{2,0}x_{s,0}+w_{2,1}x_{s,1})^m} + \hat{d}_2 + g_2\right) x_{s,2} \\ \gamma_{4,4} = \alpha_0 x_q + \beta_0 x_{s,0} \\ \gamma_{1,4} = \gamma_{4,1} \;=\; -\left(\alpha_0 x_q + \beta_0 x_{s,0}\right) \;=\; -\gamma_{4,4} \end{cases}, \quad (11.2)$$

with all remaining entries being zero. Note that G is diagonally dominant with positive entries on its diagonal (for $x_{s,0}, x_{s,1}, x_{s,2}x_q > 0$). Thus, due to the Gershgorin circle theorem, G has nonnegative eigenvalues only. This enables us to determine its "square root" later on.

Finally, if we scale the time as $\tau = t/N_e = t\Delta x$ and scale the continuous density distribution as $\rho(x, \tau)\Delta x = p(x, t)$, we get an approximate Fokker–Planck equation representing model C:

$$\partial_\tau \rho(x,\tau) \;=\; -\sum_{i=1}^{4} \partial_{x_i} \left(V_i(x)\rho(x,\tau)\right) \;+\; \tfrac{1}{2N_e} \sum_{i,j=1}^{4} \partial_{x_i}\partial_{x_j} \left(G_{ij}(x)\rho(x,\tau)\right), \quad (11.3)$$

with drift-vector V and diffusion matrix G as in Eqs. (11.1) and (11.2), respectively.

11.2.4 The Stochastic Version of Model C in Terms of Itô/Langevin Equations

There exists an easy translation from the Fokker–Planck Eq. (11.3) to a stochastic differential equation, cf. [44], namely

$$\mathrm{d}X_t \;=\; V(x)\mathrm{d}t \;+\; L(x)\,\mathrm{d}W_t\,, \quad \text{with} \quad G(x) \;=\; L(x)L^T(x)\,,$$

where W_t is a vector-valued Wiener process and V and $G = (\gamma_{i,j})_{i,j=1,\ldots,4}$ as specified in (11.1) and (11.2), respectively. Here, the (Langevin) diffusion matrix L is formulated as

$$L = \sqrt{\frac{1}{N_e}} \begin{pmatrix} \sqrt{\gamma_{1,1} - \gamma_{4,4}} & 0 & 0 & -\sqrt{\gamma_{4,4}} \\ 0 & \sqrt{\gamma_{2,2}} & 0 & 0 \\ 0 & 0 & \sqrt{\gamma_{3,3}} & 0 \\ 0 & 0 & 0 & \sqrt{\gamma_{4,4}} \end{pmatrix}.$$

Note, L is unique only up to orthogonal transformations.

This technique leads to the following four-dimensional system of coupled Itô stochastic differential equations (Langevin type) for the dynamics of cancer stem cells in a niche, corresponding to model C:

$$\begin{aligned} \mathrm{d}x_{s,0} = &\left(\left(\frac{l_0}{1 + (x_{s,0} + w_{0,1}x_{s,1} + w_{0,2}x_{s,2})^m} - d_0 - \beta_0 \right) x_{s,0} + \alpha_0 x_q \right) \mathrm{d}t \\ &+ \sqrt{\frac{1}{N_e} \left(\left(\frac{l_0}{1 + (x_{s,0} + w_{0,1}x_{s,1} + w_{0,2}x_{s,2})^m} + \hat{d}_0 + g_0 \right) x_{s,0} \right)} \, \mathrm{d}W_t^{(0)} \\ &- \sqrt{\frac{1}{N_e} \left(\alpha_0 x_q + \beta_0 x_{s,0} \right)} \mathrm{d}W_t^{(q)}, \end{aligned}$$

$$\begin{aligned} \mathrm{d}x_{s,1} = &\left(\frac{l_1}{1 + (x_{s,1} + w_{1,0}x_{s,0} + w_{1,2}x_{s,2})^m} - d_1 \right) x_{s,1} \mathrm{d}t \\ &+ \sqrt{\frac{1}{N_e} \left(\frac{l_1}{1 + (x_{s,1} + w_{1,0}x_{s,0} + w_{1,2}x_{s,2})^m} + \hat{d}_1 + g_1 \right) x_{s,1}} \, \mathrm{d}W_t^{(1)}, \end{aligned}$$

$$\begin{aligned} \mathrm{d}x_{s,2} = &\left(\frac{l_2}{1 + (x_{s,2} + w_{2,0}x_{s,0} + w_{2,1}x_{s,1})^m} - d_2 \right) x_{s,2} \mathrm{d}t \\ &+ \sqrt{\frac{1}{N_e} \left(\frac{l_2}{1 + (x_{s,2} + w_{2,0}x_{s,0} + w_{2,1}x_{s,1})^m} + \hat{d}_2 + g_2 \right) x_{s,2}} \, \mathrm{d}W_t^{(2)}, \end{aligned}$$

$$\mathrm{d}x_q = \left(-\alpha_0 x_q + \beta_0 x_{s,0} \right) \mathrm{d}t + \sqrt{\frac{1}{N_e} \left(\alpha_0 x_q + \beta_0 x_{s,0} \right)} \mathrm{d}W_t^{(q)},$$

where $W_t^{(0)}, \ldots, W_t^{(q)}$ are independent scalar standard Wiener processes. In this context $d_i = \hat{d}_i - g_i$, $i = 0, 1, 2$ can be interpreted as the net unregulated death–birth rate. As already mentioned, the effective stem cell niche reference population

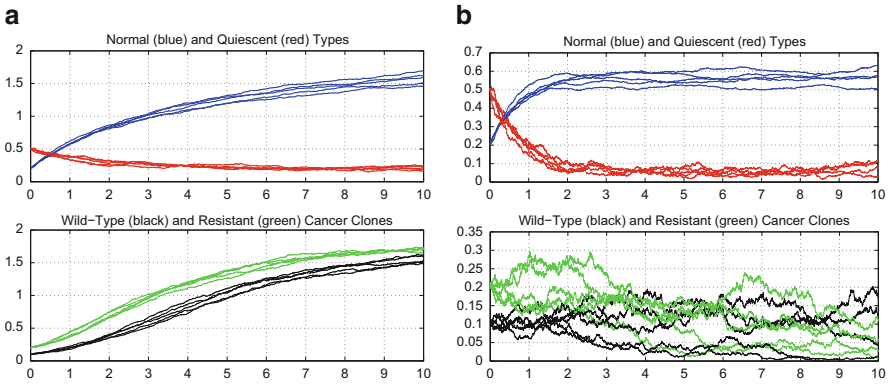

Fig. 11.2 Model C: five simulation runs for each of the stochastically driven stem cell populations $x_{s,0}$ (*blue*), x_q (*red*) in the top subfigures and $x_{s,1}$ (*black*), $x_{s,2}$ (*green*) in the lower subfigures over time for different parameter values that highlight the stochastic influence: (**a**) $g_0 = g_1 = g_2 = 0.1$, $\hat{d}_0 = \hat{d}_1 = \hat{d}_2 = 0.3$ and (**b**) $g_0 = g_1 = g_2 = 4$, $\hat{d}_0 = \hat{d}_1 = \hat{d}_2 = 5$. The remaining parameters are the same for columns (a) and (b); they are given as $\alpha_0 = 0.1$, $\beta_0 = 1$, $l_0 = l_1 = l_2 = 1$, together with $w_{0,1} = w_{1,0} = 0.5$, $w_{0,2} = w_{1,2} = w_{2,0} = w_{2,1} = 0.1$, as well as $d_i = \hat{d}_i - g_i = 0.2$ for $i = 1, 2, 3$. The initial conditions are $x_{s,0}(0) = 0.5$, $x_q(0) = 0.2$ in the top subfigures and $x_{s,1} = 0.1$, $x_{s,2} = 0.2$ in the lower subfigures

is defined as N_e and is here taken to be fixed for simplicity. It can be seen that N_e only affects the variance in the rescaled model, but defines the population measurement unit. In practice, one can scale $w_{i,i} = 1/N_e$ and take $N_e w_{i,j} \to w_{i,j}$, $n_i/N_e \to x_i$ for $i = 0, 1, 2$. The model C might be improved by adopting nonuniform population scales $w_{i,i} = 1/N_{e,i}$. These population scalings correspond to the regulated population capacity of the stem cell niche.

Inspection of this system, especially of its underlying Markovian birth and death dynamics, leads to the conclusion that once a population vanishes it stays zero for all consecutive times unless reemergence of a specific clone by activation of associated quiescent stem cells or by reoccurrence of the mutagenesis is taken into account. Hence, the phase space orthant of the nonnegative populations $\{(x_{s,0}, x_{s,1}, x_{s,2}, x_q) \in \mathbb{R}^4 : x_{s,0}, x_{s,1}, x_{s,2}, x_q \geq 0\}$ may stay invariant. Moreover, the drift coefficients of this system are Lipschitz-continuous and fulfill the usual linear growth conditions for all elements of the state space. The diffusion coefficients are Lipschitz-continuous and fulfill the usual linear growth conditions outside any environment of the origin. Thus, outside the origin this system has a path-wise unique strong solution for any final and nonvanishing initial condition, cf. [14], p. 98. This allows the application of path-wise numerical schemes to simulate the dynamics as in Fig. 11.2 where the influence of the stochastic perturbations is highlighted. Part (a) of Fig. 11.2 shows a very weak influence of the stochastic diffusion whereas the influence of stochasticity is well recognizable in part (b). Interestingly enough, the parameters chosen in (b) seem to lead to an extinction of the cancer clones compared to those taken in (a). In order to understand, at

least from a small noise perspective, what really happens a thorough discussion of the parameter space and its corresponding deterministic phase spaces needs to be carried out. As we will see in the following section multiple bifurcation scenarios are present in the parameter space.

For the preparation of Fig. 11.2 these stochastic simulations were carried out with the Milstein scheme, a numerical scheme of strong order of convergence one, cf. [28, p. 345], and supplemented with the positions of the equilibria of the underlying deterministic system. We will formulate the Milstein scheme for the stochastic dynamics of competing normal and leukaemic stem cells in a niche in the appendix of this chapter. Of, course, like the existence and path-wise uniqueness of strong solutions, the order of convergence depends crucially on the assumption that the coefficient functions are (globally) Lipschitz-continuous. Because of the square roots at the diffusion coefficients these assumptions hold outside a neighborhood of the origin, only. This may cause the Milstein scheme to deliver negative numerical results even though the analytic solution always stays positive. This is a well-known difficulty in financial mathematics, cf. [27].

The induced deterministic dynamics will be analyzed in the remainder of this chapter by a complete discussion of the deterministic equilibria, their stability, and the global dynamics they contribute to. Such a study of the deterministic dynamics is quite valuable for the stochastic case, as we will show.

11.3 Equilibria and Their Stability in the Deterministic Small Noise Limit

The discussion of the underlying deterministic dynamics forms an important foundation for the proper understanding of any stochastic system. Though the deterministic system is the noise-free representation of the stochastic system, one can expect the stochastic system's behavior for small noise to be close to that of the free noise system, cf. [4, 12]. Moreover, deterministic equilibria remain as sources and sinks in a stochastic density interpretation[2].

We therefore analyze the global dynamics of the deterministic normal/leukaemic stem cell system (model C) [10]:

$$\dot{x}_{s,0} = \left(\frac{l_0}{1 + (x_{s,0} + w_{0,1} x_{s,1} + w_{0,2} x_{s,2})^m} - d_0 - \beta_0 \right) x_{s,0} + \alpha_0 x_q \,, \quad (11.4)$$

$$\dot{x}_{s,1} = \left(\frac{l_1}{1 + (x_{s,1} + w_{1,0} x_{s,0} + w_{1,2} x_{s,2})^m} - d_1 \right) x_{s,1} \,, \quad (11.5)$$

[2] Of course, it is not uncommon that "large" noise perturbations may drive the trajectories away even from an asymptotically stable equilibrium of the deterministic system, cf. [4, 23].

$$\dot{x}_{s,2} = \left(\frac{l_2}{1 + (x_{s,2} + w_{2,0} x_{s,0} + w_{2,1} x_{s,1})^m} - d_2 \right) x_{s,2} \,, \tag{11.6}$$

$$\dot{x}_q = -\alpha_0 x_q + \beta_0 x_{s,0} \,, \tag{11.7}$$

with $l_i, d_i > 0$ $(i = 0, 1, 2)$, $w_{0,1}, w_{0,2}, w_{1,0}, w_{1,2}, w_{2,0}, w_{2,1} > 0$, $\alpha_0, \beta_0 > 0$ and $m \in \mathbb{N}/2\mathbb{N}$.

Due to "symmetries" in these equations, the discussion is presented sequentially: First, we propose model A to study the dynamics of two competing cell species, which could be normal and wild-type CML stem cells or alternatively two different leukaemic strains, i.e., a subsystem involving the variables $x_{s,1}$ and $x_{s,2}$ is considered (i.e., $x_{s,0} = 0$ and $x_q = 0$). This corresponds to a two-dimensional manifold in the solution space and allows insights into the behavior of two developing cell lines. Second, the three-dimensional state space manifold $(x_{s,0}, x_{s,1}, x_q)$ is discussed (model B). This means that two important sub-manifolds are completely analyzed. Finally, the full four-dimensional system (model C) including both cycling wild-type and *imatinib*-resistant CML strains is considered. This sequential procedure has the advantage that analysis of the four-dimensional state space can be restricted to the dynamics outside of the lower-dimensional sub-manifolds.

11.3.1 Model A: The Dynamics of Two Competing Clones

As a starting point, we completely analyze the deterministic dynamics of the two species sub-model. By introducing the variables $x = x_{s,1}$ and $y = x_{s,2}$, the first order system of ordinary differential equations corresponding to model A is given by

$$\dot{x} = \left(\frac{l_1}{1 + (x + w_1 y)^m} - d_1 \right) x \,, \tag{11.8}$$

$$\dot{y} = \left(\frac{l_2}{1 + (y + w_2 x)^m} - d_2 \right) y \,, \tag{11.9}$$

with $l_i, w_i, d_i > 0$ $(i = 1, 2)$ and $m \in \mathbb{N}/2\mathbb{N}$.

The variables x and y may represent either normal and wild-type CML stem cells, respectively, or alternatively two competing leukaemic clones.

The right-hand side's coefficient functions of this system have uniformly bounded derivatives and are thus globally Lipschitz-continuous. Moreover, they satisfy the usual linear growth condition. Hence, our two-dimensional system has an unique solution on its maximal domain of definition, $\mathbb{R}^2$. Moreover, the first quadrant $\{(x, y) \in \mathbb{R}^2 : x, y \geq 0\}$ is invariant and negative populations are not accessible from positive initial conditions.

A graphical representation of the dynamics determined by Eqs. (11.8)–(11.9) is given in Fig. 11.3 for the parameters $l_1 = l_2 = 1$, $w_1 = 1$, $w_2 = 1.1$, $d_1 = d_2 = 0.9$

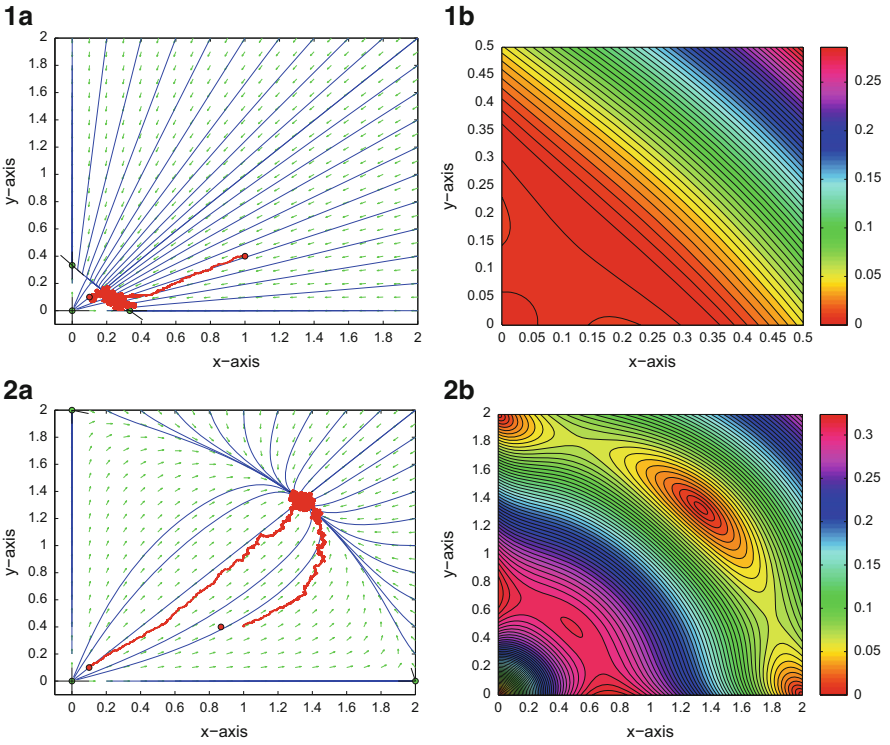

Fig. 11.3 Model A: simulation of the deterministic dynamics of normal and wild-type CML stem cells (or of wild-type and resistant CML clones) described by Eqs. (11.8)–(11.9) for the parameters $l_1 = l_2 = 1$, $w_1 = 1$, $w_2 = 1.1$, $d_1 = d_2 = 0.9$ and $m = 2$ in the first row (*1*) and for the parameters $l_1 = l_2 = 1$, $w_1 = w_2 = 0.5$, $d_1 = d_2 = 0.2$ and $m = 2$ in the second row (*2*) (cf. [10] for the parameterization of a similar equation system). Subfigures in row (*a*) show the deterministic phase space including two stochastic trajectories (with $N_e = 1{,}000$) and row (*b*) provides the norm of the corresponding deterministic vector-field

and $m = 2$ (first row) as well as $l_1 = l_2 = 1$, $w_1 = w_2 = 0.5$, $d_1 = d_2 = 0.2$ and $m = 2$ (second row) (cf. [10] for the parameterization of a similar equation system) Figs. 11.3 (1a) and (2a) display the phase space together with the vector-field and a bundle of specific deterministic trajectories for the corresponding parameters. In particular, they show the system's equilibria together with the directions of their stable and unstable eigen-spaces, and, starting at $(x, y) = (0.1, 0.1)$ and $(x, y) = (1, 0.4)$, two stochastic trajectories which are characterized by a diffusion part that is analogous to the full model presented in Sect. 11.2.4 (with $N_e = 1000$). Moreover, Figs. 11.3 (1b) and (2b) illustrate the norm of the corresponding vector-field thus indicating the speed of a population combination. We observe that this speed declines and nearly vanishes at the heteroclinic orbit connecting the equilibria at the coordinate-axes (1b) and locally around the mixed species equilibrium (2b),

respectively. This explains why the stochastic trajectories stay in the corresponding region as their drift component no longer contributes significantly.

To understand the dynamics for all parameter regimes we discuss the existence and stability of the equilibria of the system Eqs. (11.8)–(11.9).

Theorem 11.1 (Existence of Equilibria). *Besides the point at infinity, for admissible parameters $l_i, w_i, d_i > 0$ $(i = 1, 2)$ and $m \in \mathbb{N}/2\mathbb{N}$ the following points in $\mathbb{R}^2$ are equilibria of the system Eqs. (11.8)–(11.9):*

- *the origin $(x, y) = (0, 0)$ for all choices of the parameter.*
- *the line equilibrium $(x, 0) = (x_s, 0)$, with $x_s := \sqrt[m]{d_1^{-1}(l_1 - d_1)}$, provided $l_1 - d_1 > 0$ holds.*
- *the line equilibrium $(0, y) = (0, y_s)$, with $y_s := \sqrt[m]{d_2^{-1}(l_2 - d_2)}$, provided $l_2 - d_2 > 0$ holds.*
- *the face equilibrium $(x, y) = (x_f, y_f)$, where*

$$x_f := \frac{x_s - w_1 y_s}{1 - w_1 w_2}, \quad \text{and} \quad y_f := \frac{y_s - w_2 x_s}{1 - w_1 w_2},$$

  given that $x_f, y_f > 0$, $w_1 w_2 \neq 1$ and x_s, y_s exist as defined above.
- *the line of equilibria $(x, y) = (x, w_2(x_s - x))$, if $w_1 w_2 = 1$ and x_s, y_s exists as defined above, for all x such that $x_s > x$. Note, that $y_s = w_2 x_s$ holds in this case.*

Moreover, these are the only candidates for equilibria in the first quadrant.

Proof. By definition, the equilibria of Eqs. (11.8)–(11.9) are those points for which the right-hand sides of the system equations vanish. This is correct for $(x, y) = (0, 0)$. Next, let $x > 0$ and $y = 0$, then

$$0 \stackrel{!}{=} \frac{l_1}{1 + x^m} - d_1 \quad \Leftrightarrow \quad x = \overset{+}{[-]} \sqrt[m]{d_1^{-1}(l_1 - d_1)} = x_s \,,$$

provided $l_1 - d_1 > 0$ holds. The analogous argumentation shows that $(0, y_s)$ is the unique equilibrium for $y > 0$ and $x = 0$, provided $l_2 - d_2 > 0$ holds. Finally, let $x, y > 0$. Then, provided x_s and y_s exist,

$$\begin{cases} x + w_1 y = x_s \\ y + w_2 x = y_s \end{cases} \quad \Leftrightarrow \quad \begin{cases} x + w_1 y = x_s \\ y(1 - w_1 w_2) = y_s - w_2 x_s \end{cases} .$$

The assertion on the face equilibrium follows if $1 - w_1 w_2 \neq 0$ and $(y_s - w_2 x_s)$ $(1 - w_1 w_2)^{-1} > 0$ together with $(x_s - w_1 y_s)(1 - w_1 w_2)^{-1}$ hold. If $w_1 w_2 = 1$ it results that $y_s = w_2 x_s$ and consequently $y = w_2(x_s - x)$ together with $0 < x < x_s$. There are no further candidates in the first quadrant that lead to equilibria. □

For further reference, we define the following quantities that will be utilized later as bifurcation parameters:

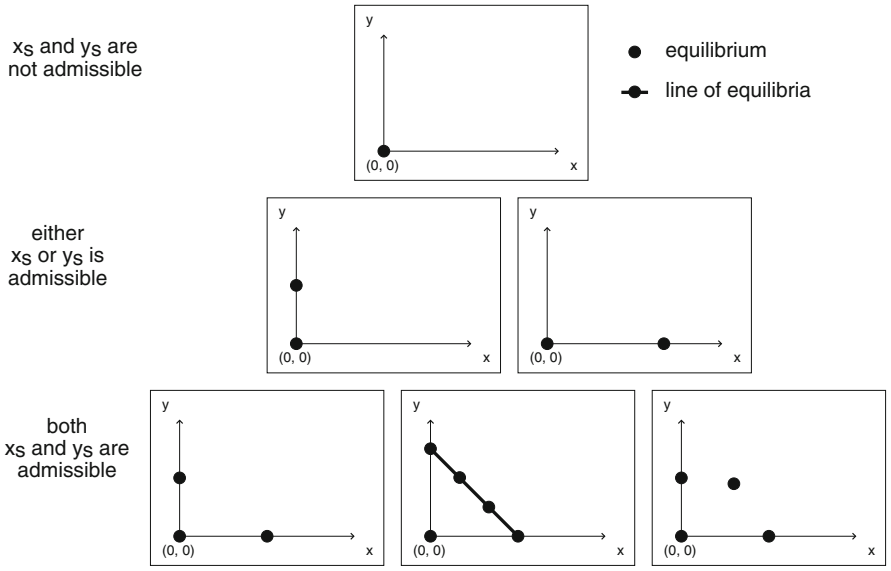

Fig. 11.4 Model A: graph of the possible equilibrium situations of the system Eqs. (11.8)–(11.9) as stated in Theorem 11.1

$$\gamma_x := (l_1 - d_1)d_1^{-1}\,, \qquad \gamma_x = x_s^m\,, \text{ if } \gamma_x > 0\,,$$
$$\gamma_y := (l_2 - d_2)d_2^{-1}\,, \qquad \gamma_y = y_s^m\,, \text{ if } \gamma_y > 0\,.$$

Counting the number of equilibria with respect to those quantities (cf. Fig. 11.4) reveals that the origin is always an equilibrium. If x_s and / or y_s are admissible by the parameters, one, two, or infinitely many (for $w_1 w_2 = 1$) equilibria exist. Finally, if x_s and y_s are both admissible and the weights w_1 and w_2 are such that x_f and y_f exist, then the system Eqs. (11.8)–(11.9) has four isolated equilibria.

We continue with a discussion of the linearized stability properties of the equilibria. Therefore, evaluation of the 2×2-Jacobi-matrix $J(x^*, y^*)$ as given by the right-hand side of the system Eqs. (11.8)–(11.9) at an equilibrium (x^*, y^*) in the first quadrant is required. In general, the entries of $J(x^*, y^*)$ are given by the expressions:

$$\begin{aligned} J_{1,1}(x^*, y^*) &= \frac{\partial}{\partial x}\left(\left(\frac{l_1}{1 + (x + w_1 y)^m} - d_1\right) x\right)\Big|_{(x,y)=(x^*,y^*)} \\ &= \underbrace{\frac{l_1}{1 + (x^* + w_1 y^*)^m} - d_1}_{=0\,,\quad \text{if } x^* \neq 0} - \frac{m l_1 (x^* + w_1 y^*)^{m-1}}{\left(1 + (x^* + w_1 y^*)^m\right)^2} x^*\,, \end{aligned}$$

$$J_{1,2}(x^*, y^*) = -\frac{m l_1 w_1 (x^* + w_1 y^*)^{m-1}}{\left(1 + (x^* + w_1 y^*)^m\right)^2} x^*\,,$$

and

$$J_{2,2}(x^*,y^*) = \frac{\partial}{\partial y}\left(\left(\frac{l_2}{1+(y+w_2x)^m} - d_2\right)y\right)\Big|_{(x,y)=(x^*,y^*)}$$
$$= \underbrace{\frac{l_2}{1+(y^*+w_2x^*)^m} - d_2}_{=0\,,\quad \text{if } y^*\neq 0} - \frac{ml_2(y^*+w_2x^*)^{m-1}}{\left(1+(y^*+w_2x^*)^m\right)^2}y^*\,,$$

$$J_{2,1}(x^*,y^*) = -\frac{ml_2w_2(y^*+w_2x^*)^{m-1}}{\left(1+(y^*+w_2x^*)^m\right)^2}y^*\,.$$

Thus, the stability properties of the origin become obvious:

Proposition 11.1 (Linearized Stability of the Origin). *The Jacobi-matrix of system Eqs. (11.8)–(11.9) evaluated at the origin corresponds to* $J(0,0) = diag\,(l_1-d_1, l_2-d_2)$*, and consequently* $(x,y) = (0,0)$ *is hyperbolic if* $l_1-d_1, l_2-d_2 \neq 0$*. In particular,* $J(0,0)$ *is asymptotically stable in the linearized system if both* l_1-d_1 *and* l_2-d_2 *are negative, stable if either* $l_1-d_1=0$ *(or* $l_2-d_2=0$*) together with* $l_2-d_2\leq 0$ *(or* $l_1-d_1\leq 0$*), and unstable if* $l_1-d_1>0$ *or* $l_2-d_2>0$*.*

Moreover, $\mathfrak{e}_1 := (1,0)^T$ *is the eigenvector corresponding to the eigenvalue* l_1-d_1*, and* $\mathfrak{e}_2 := (0,1)^T$ *corresponds to the eigenvalue* l_2-d_2*.*

The principle of linearized stability allows to transfer the stability properties as determined for the linearized system at the origin to the nonlinear system Eqs. (11.8)–(11.9) if $(x,y)=(0,0)$ is hyperbolic. In the case that $(x,y)=(0,0)$ is non-hyperbolic, we utilize the function $V(x,y)=\frac{1}{2}x^2+\frac{1}{2}y^2$ as a strict Lyapunov function to show that the origin is asymptotically stable if $l_1-d_1\leq 0$ and $l_2-d_2\leq 0$. The conditions $V(x,y)>0$ for all $(x,y)\in\mathbb{R}^2\setminus\{0\}$ and $V(x,y)=0$ if and only if $(x,y)=(0,0)$ follow immediately, as well as

$$\begin{aligned} V'(x,y) &= \langle \nabla V(x,y), (\dot{x},\dot{y})^T\rangle \\ &= \left(\tfrac{l_1}{1+(x+w_1y)^m} - d_1\right)x^2 + \left(\tfrac{l_2}{1+(y+w_2x)^m} - d_2\right)y^2 \\ &\leq 0\,, \qquad \text{as } l_1-d_1\leq 0 \text{ and } l_2-d_2\leq 0\,, \end{aligned}$$

for the orbital derivative V' together with $V'(x,y)=0$ if and only if $(x,y)=(0,0)$. Thus, as claimed, V is a strict Lyapunov function and the origin is asymptotically stable in the non-hyperbolic case $l_1-d_1=0$ and $l_2-d_2=0$, too. Note that if the line equilibria $(x_s,0)$ and / or $(0,y_s)$ exist, then $l_1-d_1>0$ and / or $l_2-d_2>0$ holds, which leads to an unstable eigen-direction of the origin. Consequently, this leads to instability in the non-hyperbolic cases ($l_1-d_1=0$, $l_2-d_2>0$) and ($l_1-d_1>0$, $l_2-d_2=0$).

Moreover, for any pair (l_1,d_1), (l_2,d_2) there exist large enough values of x and y such that the right-hand side of Eqs. (11.8)–(11.9) becomes negative and stays

negative for even larger values of x and y. Thus, the point at infinity acts as a source for all parameters, and if the origin becomes repellent, too, another attractor has to form.

The coordinate lines $\{(x,0) \in \mathbb{R}^2 : x \geq 0\}$ and $\{(0,y) \in \mathbb{R}^2 : y \geq 0\}$ are invariant under the dynamics induced by Eqs. (11.8)–(11.9). Proposition 11.1 thus already proves (for this one-dimensional setting) that the line equilibria $(x_s, 0)$ and $(0, y_s)$, if they exist, have at least one (asymptotically) stable eigen-direction, and that $(x_s, 0)$ and $(0, y_s)$, if they exist, are trivially connected to the origin by an orbit lying totally in the respective coordinate line.

Proposition 11.2 (Linearized Stability of the Line Equilibria). *Let x_s be admissible by the parameter, then the Jacobi-matrix of system Eqs. (11.8)–(11.9) evaluated at the line equilibrium $(x_s, 0)$ is given by*

$$J(x_s,0) \;=\; \begin{pmatrix} -md_1(l_1-d_1)l_1^{-1} & -mw_1d_1(l_1-d_1)l_1^{-1} \\ 0 & j_{2,2} \end{pmatrix}$$

with

$$j_{2,2} \;=\; \frac{d_1d_2}{(1-w_2^m)d_1 + w_2l_1}\left(\gamma_y - w_2^m x_s^m\right) .$$

The value $\lambda_1 := -md_1(l_1-d_1)l_1^{-1} < 0$ is an eigenvalue of $J(x_s,0)$, and $\mathfrak{e}_1$ is its corresponding eigenvector[3]*. The value $\lambda_2 := j_{2,2}$ is the second eigenvalue of $J(x_s,0)$, and*

$$\left(-\frac{j_{2,2} + mw_1d_1(l_1-d_1)l_1^{-1}}{md_1(l_1-d_1)l_1^{-1}},\, 1\right)^T$$

is the corresponding eigenvector of the Jacobi matrix.

Thus, $(x_s, 0)$ is asymptotically stable if $\gamma_y < w_2^m x_s^m$, it is a saddle if $\gamma_y > w_2^m x_s^m$, and its linearization has a neutral component if $w_2^m x_s^m = y_s^m = \gamma_y$. For $w_1 w_2 = 1$, this neutral component corresponds to the line of equilibria as described in Theorem 11.1.

The analogous result can be established for the line equilibrium $(0, y_s)$. In particular, the eigenvalues of the corresponding Jacobi-matrix are $-md_2(l_2 - d_2)l_2^{-1} < 0$ and $d_1d_2\left((1-w_1^m)d_2 + w_1l_2\right)^{-1}(\gamma_x - w_1^m y_s^m)$.

Proof. It is sufficient to take a more detailed look at $j_{2,2}$, the remaining assertions follow immediately from the Jacobi-matrix:

[3] The line equilibrium $(x_s, 0)$ is thus always asymptotically stable with respect to the eigen-direction corresponding to λ_1.

$$j_{2,2} = \frac{l_2 d_1}{(1-w_2^m)d_1 + w_2^m l_1} - d_2 = \frac{(1-w_2^m)d_1(l_2-d_2) + w_2^m(l_2 d_1 - l_1 d_2)}{(1-w_2^m)d_1 + w_2^m l_1}$$
$$= \frac{d_1 d_2}{(1-w_2^m)d_1 + w_2 l_1}\left(\gamma_y - w_2^m x_s^m\right) .$$

Following the proof of Theorem 11.1 it holds that $w_2 x_s = y_s$ for $w_1 w_2 = 1$ and existing x_s, y_s. □

If $\lambda_2 < 0$ ($\lambda_2 > 0$), the nonlinear asymptotic stability (instability) of $(x_s, 0)$ follows immediately from the principle of linearized stability, and the nonlinear stability in the special case $w_1 w_2 = 1$ and existing x_s, y_s is obvious due to the line of equilibria.

The nonlinear stability of the non-hyperbolic case $\gamma_y = w_2^m x_s^m$ for $w_1 w_2 \neq 1$ remains to be investigated. Therefore, we apply the method of Markov partitions that constitute transitions between certain regions of the phase space near the equilibrium. Let $\varepsilon_x \in \mathbb{R}$ and $\varepsilon_y > 0$ be small enough such that $(x_s + \varepsilon_x, \varepsilon_y)$ is near $(x_s, 0)$ and in the first quadrant. Then, the dynamics of the initial point $(x_s + \varepsilon_x, \varepsilon_y)$ as described by Eqs. (11.8)–(11.9) is governed by

$$\dot{\varepsilon}_x = \left(\frac{l_1}{1 + (x_s + \varepsilon_x + w_1 \varepsilon_y)^m} - d_1\right)(x_s - \varepsilon_x) ,$$
$$\dot{\varepsilon}_y = \left(\frac{l_2}{1 + (\varepsilon_y + w_2(x_s + \varepsilon_x))^m} - d_2\right)\varepsilon_y .$$

For all $(\varepsilon_x, \varepsilon_y)$ in

$$S_{--} := \left\{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1 \varepsilon_y > 0 \text{ and } \varepsilon_y + w_2 \varepsilon_x > 0\right\}$$

it holds that $\dot{\varepsilon}_x, \dot{\varepsilon}_y < 0$, as

$$\frac{l_1}{1+(x_s+\varepsilon)^m} - d_1 < 0 , \quad \text{and} \quad \frac{l_2}{1+(\varepsilon + w_2 x_s)^m} - d_2 < \frac{l_2}{1+(w_2 x_s)^m} - d_2 = 0 ,$$

for all $\varepsilon > 0$. That is, a positive value of $\varepsilon_x + w_1 \varepsilon_y$ leads to a negative value of $\dot{\varepsilon}_x$ and thus a decrease of ε_x. Similarly, for $w_1^{-1} > w_2$, the Markov partitions

$$S_{0-} := \{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1 \varepsilon_y = 0 \text{ and } \varepsilon_y + w_2 \varepsilon_x > 0\} ,$$
$$S_{+-} := \{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1 \varepsilon_y < 0 \text{ and } \varepsilon_y + w_2 \varepsilon_x > 0\} ,$$
$$S_{+0} := \{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1 \varepsilon_y < 0 \text{ and } \varepsilon_y + w_2 \varepsilon_x = 0\} ,$$
$$S_{++} := \{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1 \varepsilon_y < 0 \text{ and } \varepsilon_y + w_2 \varepsilon_x < 0\} ,$$

are defined and yield specific signs of $\dot{\varepsilon}_x$ and $\dot{\varepsilon}_y$. In particular, S_{+-} is positively invariant and its elements tend toward the equilibrium $(x_s, 0)$. Moreover, elements of all other partitions eventually reach S_{+-} or the equilibrium $(x_s, 0)$ itself, see Fig. 11.5. This describes the nonlinear dynamics around $(x_s, 0)$ completely and

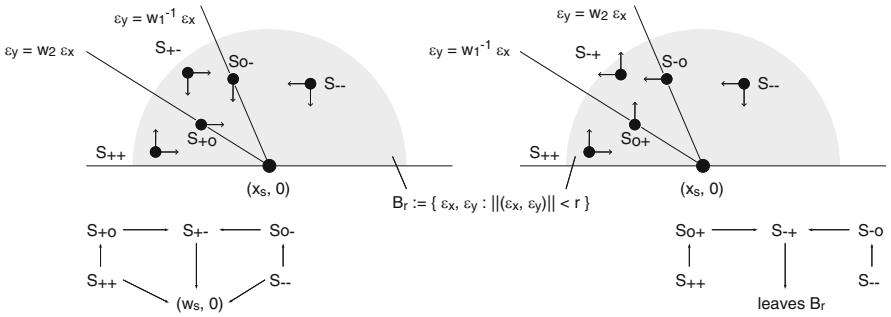

Fig. 11.5 Model A: sketch of the Markov partitions near the equilibrium $(x_s, 0)$ and the transitions between these phase space partitions

constitutes the asymptotic stability of $(x_s, 0)$ for $w_1^{-1} > w_2$. An analogous discussion, as graphically represented in Fig. 11.5, shows the saddle point character of $(x_s, 0)$ for $w_1^{-1} < w_2$. There, additionally the Markov partitions

$$S_{-+} := \{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1\varepsilon_y > 0 \text{ and } \varepsilon_y + w_2\varepsilon_x < 0\} ,$$
$$S_{-0} := \{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1\varepsilon_y > 0 \text{ and } \varepsilon_y + w_2\varepsilon_x = 0\} ,$$
$$S_{0+} := \{(\varepsilon_x, \varepsilon_y) \in \mathbb{R} \times \mathbb{R}^+ : \varepsilon_x + w_1\varepsilon_y = 0 \text{ and } \varepsilon_y + w_2\varepsilon_x < 0\} ,$$

are defined. For $w_1^{-1} < w_2$ elements of any other partition may enter S_{-+} and then be transported away from the equilibrium.

As next step we focus on the linearized stability of the elements of the line of equilibria:

Proposition 11.3 (Linearized Stability of the Line of Equilibria). *Let $w_1 w_2 = 1$ and x_s be admissible by the parameters. Then the Jacobi-matrix of system Eqs. (11.8)–(11.9) evaluated at an interior element of the line of equilibria $(x, w_2(x_s - x))$ $(x \in (0, x_s))$ reads as*

$$J(x, w_2(x_s - x)) = -\begin{pmatrix} \alpha & w_1\alpha \\ w_2\beta & \beta \end{pmatrix} ,$$

where

$$\alpha := \frac{m l_1 x_s^{m-1}}{(1 + x_s^m)^2} x > 0 , \qquad \beta := \frac{m l_2 (w_2 x_s)^{m-1}}{(1 + (w_2 x_s)^m)^2} w_2 (x_s - x) > 0 .$$

Each interior equilibrium on the line of equilibria is non-hyperbolic with a vanishing eigenvalue as well as a real negative eigenvalue.

Proof. The existence of a vanishing eigenvalue can be read off directly from the rank one matrix $J(x, w_2(x_s - x))$, as $w_1(\alpha, w_2\beta)^T = (w_1\alpha, \beta)$, as $w_1 w_2 = 1$. The characteristic polynomial of $J(x, w_2(x_s - x))$ is given by $\chi_J(\lambda) = \lambda\,(\lambda + (\alpha + \beta))$, which entirely verifies the assertion. □

For $w_1 w_2 = 1$ and $x_s = w_2 y_s$ any constant x_b and y_b is a solution of Eqs. (11.8)–(11.9) if the pair (x_b, y_b) lies on the line of equilibria. This result indicates that an equilibrium on the line of equilibria might be a biological state which is preferred in the signal-triggered competition and could be a target of regulation if robustness toward homeostatic equilibrium is required. The reason is that if $w_1 = w_2 = 1$ such equilibria correspond to $x + w_1 y = N =$ fixed, i.e., the total number N of stem cells is fixed. This must then be true also for the total number of differentiated cells since the populations of stem and differentiated cells are to a good approximation proportional. This might be a preferred state also for a model that includes both normal and leukaemic stem cells including regulation.

In [37] an intensive study of a similar system exhibiting such a line of deterministic equilibria is carried out with the special focus on evolutionary selection and survival times of the two involved species. The stochastic impact eventually drives a state of coexistence (e.g., an initial point at the middle of the line of equilibria) to one of the equilibria at the axis. This leads to a dynamics evolving along the line of equilibria and thus finally to the extinction of one of the species, cf. Fig. 11.3 (1a).

Proposition 11.4 (Linearized Stability of the Face Equilibria). *Let x_f and y_f be admissible. Then the Jacobi-matrix of system Eqs. (11.8)–(11.9) evaluated at a face equilibrium (x_f, y_f) can be written as*

$$J(x_f, y_f) = \begin{pmatrix} -\frac{m l_1 x_s^{m-1}}{(1+x_s^m)^2} x_f & -w_1 \frac{m l_1 x_s^{m-1}}{(1+x_s^m)^2} x_f \\ -w_2 \frac{m l_2 y_s^{m-1}}{(1+y_s^m)^2} y_f & -\frac{m l_2 y_s^{m-1}}{(1+y_s^m)^2} y_f \end{pmatrix}.$$

If $w_1 w_2 > 1$ the equilibrium (x_f, y_f) is a saddle, and if $w_1 w_2 < 1$ it is asymptotically stable.

Proof. Define $a, b > 0$ such that

$$J(x_f, y_f) = -\begin{pmatrix} a & w_1 a \\ w_2 b & b \end{pmatrix} \quad \Rightarrow \quad \chi_J(\lambda) = \lambda^2 + (a+b)\lambda + (1 - w_1 w_2) ab\,.$$

The sequence of signs of the coefficients of the quadratic characteristic polynomial $\chi_J(\lambda)$ of $J(x_f, y_f)$ has exactly one sign change if $w_1 w_2 > 1$. Accordingly, Descartes sign rule implies the existence of one positive real root and thus of one negative real root. This establishes the saddle character of (x_f, y_f).

Moreover, the usual solution formula for quadratics leads to

$$\lambda_{1/2} = \tfrac{1}{2}\left(-(a+b) \pm \sqrt{(a-b)^2 + 4 w_1 w_2 ab}\right),$$

hence, $\chi_J(\lambda)$ always has two real roots. It follows that $(a-b)^2+4w_1w_2ab < (a+b)^2$ if $w_1w_2 < 1$ and thus that in this case both roots are negative. This establishes the asymptotic stability of (x_f, y_f). □

Due to the complete classification of the equilibria in Theorem 11.1 and the invariance of the x- and y-coordinate spaces periodic orbits could exist only for parameters that allow face equilibria with the face equilibrium located at the center of the periodic orbit. However, the following proposition shows that even in this case no periodic orbits exist in the system Eqs. (11.8)–(11.9).

Proposition 11.5 (Non-Existence of Periodic Orbits). *Let the parameters of system Eqs. (11.8)–(11.9) be such that a face equilibrium at (x_f, y_f) exists. Then, there is no periodic orbit lying in the open first quadrant.*

Proof. Let $D := \{(x, y) \in \mathbb{R}^2 : x, y > 0\} \subset \mathbb{R}^2$ be the open first quadrant, and $B(x, y) : D \to \mathbb{R}$ be defined as $B(x, y) := (xy)^{-1}$. Moreover, let $F(x, y)$ denote the right-hand side of Eq. (11.8) and $G(x, y)$ that of Eq. (11.9). Then

$$\begin{aligned}&\partial_x \left(B(x,y)F(x,y)\right) + \partial_y \left(B(x,y)G(x,y)\right)\\ &= -\tfrac{ml_1(x+w_1y)^{m-1}}{y(1+(x+w_1y)^m)^2} - \tfrac{ml_2(y+w_2x)^{m-1}}{x(1+(y+w_2x)^m)^2} \;<\; 0 \quad \text{for all } (x,y) \in D\,.\end{aligned}$$

Due to Dulac's criterion, there are no periodic orbits lying entirely in D. □

This concludes the discussion of the global dynamics of the system Eqs. (11.8)–(11.9) since, according to the Theorem of Poincare–Bendixon, the ω-limit of every regular point is an equilibrium point (neither periodic orbits nor attractive connecting orbits exist due to the stability properties of the equilibria).

Finally, Fig. 11.6 summarizes our results by showing the typical dynamics for successively increased values of γ_x and γ_y, and hence illustrating various bifurcation patterns. In particular for $w_1w_2 \neq 1$ (and $\gamma_x, \gamma_y > 0$) the transition paths

$$\begin{array}{lll}
\gamma_x > w_2^{-m}\gamma_y \text{ and } \gamma_x > w_1^m\gamma_y & & \gamma_x > w_2^{-m}\gamma_y \text{ and } \gamma_x > w_1^m\gamma_y \\
\gamma_x = w_2^{-m}\gamma_y \text{ and } \gamma_x > w_1^m\gamma_y & & \gamma_x > w_2^{-m}\gamma_y \text{ and } \gamma_x = w_1^m\gamma_y \\
\gamma_x < w_2^{-m}\gamma_y \text{ and } \gamma_x > w_1^m\gamma_y & \text{or} & \gamma_x > w_2^{-m}\gamma_y \text{ and } \gamma_x < w_1^m\gamma_y \\
\gamma_x < w_2^{-m}\gamma_y \text{ and } \gamma_x = w_1^m\gamma_y & & \gamma_x = w_2^{-m}\gamma_y \text{ and } \gamma_x < w_1^m\gamma_y \\
\gamma_x < w_2^{-m}\gamma_y \text{ and } \gamma_x < w_1^m\gamma_y & & \gamma_x < w_2^{-m}\gamma_y \text{ and } \gamma_x < w_1^m\gamma_y
\end{array}$$

from $\gamma_x > w_2^{-m}\gamma_y$ and $\gamma_x > w_1^m\gamma_y$ to $\gamma_x < w_2^{-m}\gamma_y$ and $\gamma_x < w_1^m\gamma_y$ via the face equilibria are visualized. Here, the combination $0 < \gamma_x < w_2^{-m}\gamma_y$, $\gamma_x > w_1^m\gamma_y$ ($\gamma_x > w_2^{-m}\gamma_y > 0$, $\gamma_x < w_1^m\gamma_y$) immediately implies $1 > w_1w_2$ ($1 < w_1w_2$), i.e., the face equilibria are trivially admissible by the parameters.

In biological terms, Fig. 11.6 reveals that only y stem cells can survive if $\gamma_x > 0 > \gamma_y$, independent of the value of w_1 and w_2. Under the condition $\gamma_x, \gamma_y > 0$,

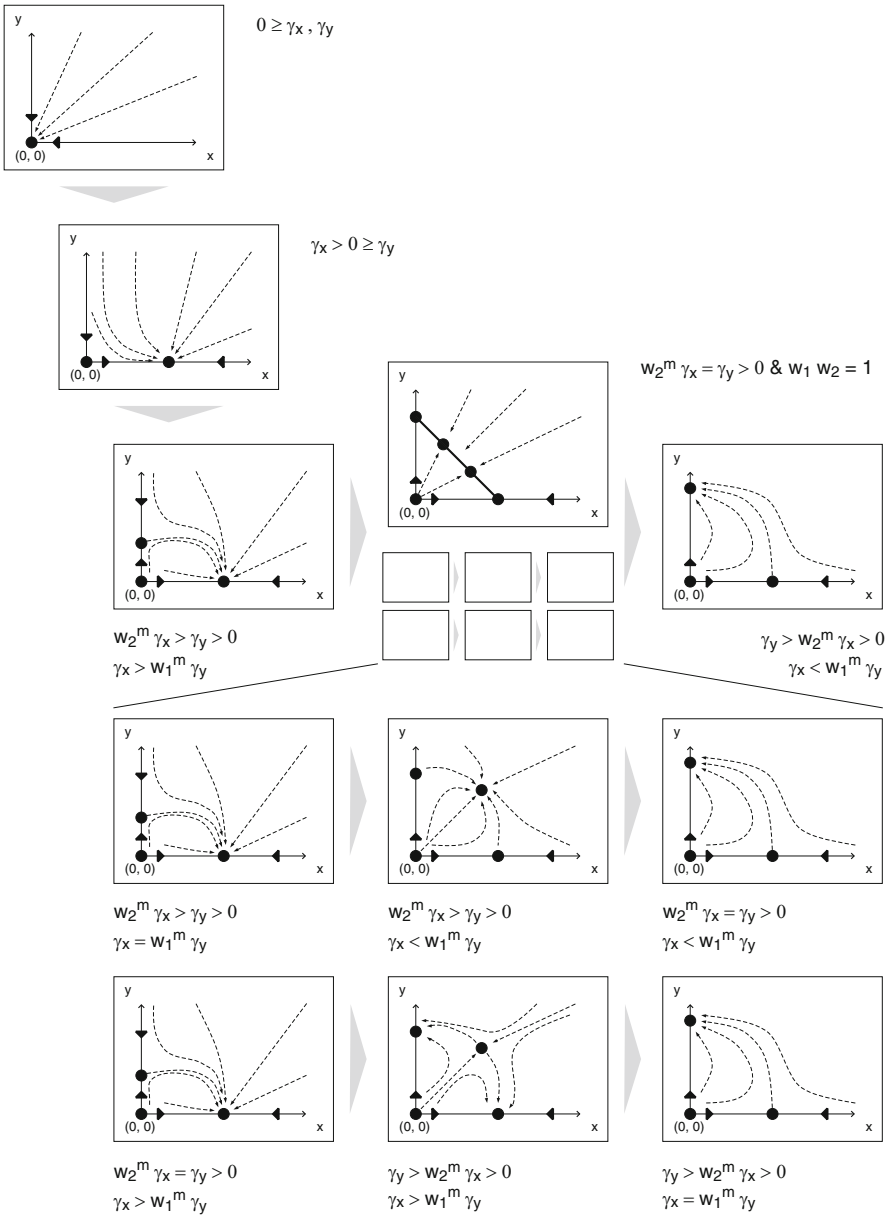

Fig. 11.6 Model A: graphical representation of bifurcation scenario showing the global dynamics for increasing values of γ_x and γ_y

both scenarios that allow for the existence of exclusively one species, x or y, and also of mixed population states are possible. The eventual picture is determined by the ratio γ_x/γ_y as well as by the magnitude of the signaling weights w_1 and w_2 and of the Hill-function index m.

If we assume that x and y represent normal and wild-type leukaemic stem cells, respectively, then the desired effect of treatment would be to shift the system toward equilibria characterized by an elimination of the y clone and a nonzero x population.

Inspection of Fig. 11.6 shows that a successful medication can, e.g., be realized with the restrictions $w_2^m\gamma_x \geq \gamma_y > 0$ and $\gamma_x \geq w_1^m\gamma_y$. Within these limits, preparation of the system in any mixed state will lead to survival of the x strain only. The reverse situation is obtained with $0 < w_2^m\gamma_x \leq \gamma_y$ and $\gamma_x \leq w_1^m\gamma_y$.

Pure x populations can also be reached from initial conditions satisfying $y < x$ if $0 < w_2^m\gamma_x < \gamma_y$ and $\gamma_x > w_1^m\gamma_y$.

This result reflects the notion that according to Eqs. (11.8)–(11.9) an effective control of leukaemic stem cells y by an agent could imply an increase of the death rate d_2, a reduction of the growth rate of y, e.g., by decreasing l_2, or achieving both effects simultaneously. These measures will lead to the desired decline of γ_y.

Figure 11.7 illustrates in a rather compact way the bifurcation pattern shown in Fig. 11.6. The two phase space dimensions x and y are supplemented by a third "pseudo"-axis representing the consecutive increase of the bifurcation parameters γ_x and γ_y. Bifurcation points, i.e., locations where the stability of one or more equilibria changes, are represented as gray circles with respect to this third coordinate, and stability of the equilibria is displayed by certain line styles (solid for stable character, dashed for saddle character, and dotted for a completely unstable character). The left figure illustrates the behavior for the case $w_1w_2 < 1$ (the right one can be interpreted analogously): by fixing a suitable value of $\gamma_y < 0$ and increasing $\gamma_x > 0$, the equilibrium at the origin stays as the only attractive equilibrium until the first bifurcation point is reached and an asymptotically stable equilibrium on the x-axis occurs (at that point the origin just becomes a saddle point). Now, we increase γ_y while fixing γ_x such that a saddle-point equilibrium occurs at the y-axis reducing the origin to an equilibrium with two unstable directions. Further increase in γ_y leads to the emergence of a stable line of equilibria. This is indicated by the solid line connecting the two bifurcation points on the x- and y-axis. These last bifurcation points finally display the change of stability for further increased values of γ_y where the line of equilibria breaks.

11.3.2 Model B: The Formation of Cancer—Competition Between Normal and Wild-Type Leukaemic Stem Cells

Next, we consider a model system including cycling and noncycling normal stem cells as well as cycling wild-type CML stem cells. We introduce the variables $x = x_{s,0}$, $y = x_{s,1}$ and $z = x_q$ to write the first order system of ordinary differential

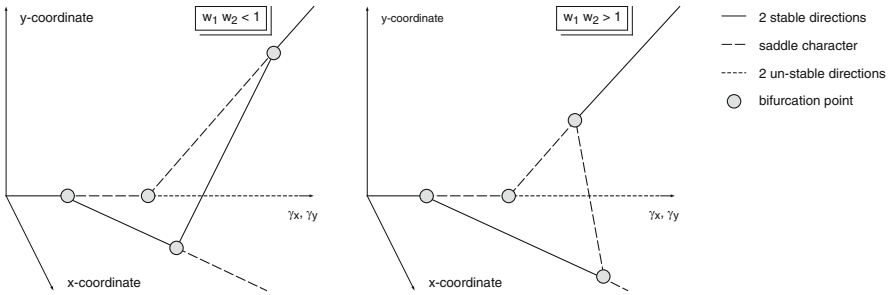

Fig. 11.7 Model A: bifurcation diagram illustrating the dependence of the equilibria on the bifurcation parameters γ_x, γ_y for $w_1 w_2 < 1$ (*left*) and $w_1 w_2 > 1$ (*right*). See the text for more details

equations corresponding to model B as

$$\dot{x} = \left(\frac{l_0}{1 + (x + w_0 y)^m} - d_0 - \beta_0 \right) x + \alpha_0 z \,, \tag{11.10}$$

$$\dot{y} = \left(\frac{l_1}{1 + (y + w_1 x)^m} - d_1 \right) y \,, \tag{11.11}$$

$$\dot{z} = -\alpha_0 z + \beta_0 x \,, \tag{11.12}$$

with $l_i, w_i, d_i > 0$ $(i = 0, 1)$, $\alpha_0, \beta_0 > 0$ and $m \in \mathbb{N}/2\mathbb{N}$. For a small initial number $y(0)$ of wild-type CML stem cells the state of the system can be interpreted as the origin of leukaemia: due to genetic aberration (cf. Sect. 17.1), leukaemic stem cells are formed and it is now interesting to simulate their development and in particular the response of the system to the administration of *imatinib*. We ask: are CML stem cells able to persist, displace normal stem cells, and take over the stem cell niche or can the growth of CML stem cells be suppressed by medication, with the goal of an asymptotically stable "healthy" equilibrium?

The right-hand side's coefficient functions of this system have uniformly bounded derivatives and are thus globally Lipschitz-continuous. Moreover, they satisfy the usual linear growth condition. Hence, our three-dimensional system has an unique solution on its maximal domain of definition, $\mathbb{R}^3$. Moreover, the first octant $\{(x, y, z) \in \mathbb{R}^3 : x, y, z \geq 0\}$ is invariant and negative populations are not accessible from positive initial conditions.

Figure 11.8 provides a first impression of the dynamics described by Eqs. (11.10)–(11.12) for the parameters $l_1 = l_2 = 1$, $w_1 = 1$, $w_2 = 1.1$, $d_1 = d_2 = 0.9$ and (first row) as well as $l_1 = l_2 = 1$, $w_1 = w_2 = 0.5$, $d_1 = d_2 = 0.2$ (second row) together with $m = 2$, $\alpha_0 = 0.1$ and $\beta_0 = 1$. Figure 11.8 (1a) and (2a) shows the purely deterministic phase space represented by a bundle of trajectories obtained with the specified parameters, and provide equilibria together with their stable and

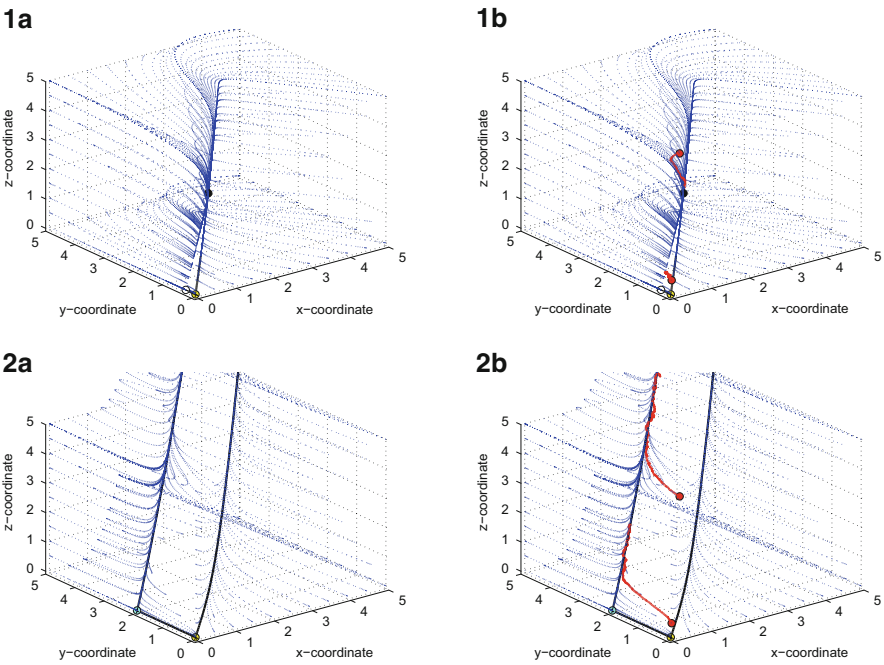

Fig. 11.8 Model B: simulation of the deterministic dynamics described by Eqs. (11.10)–(11.12) for the parameters $l_1 = l_2 = 1$, $w_1 = 1$, $w_2 = 1.1$, and $d_1 = d_2 = 0.9$ in panels (**1a**)–(**1b**) and for the parameters $l_1 = l_2 = 1$, $w_1 = w_2 = 0.5$, and $d_1 = d_2 = 0.2$ in panels (**2a**)–(**2b**) as well as $m = 2$, $\alpha_0 = 0.1$ and $\beta_0 = 1$. Two stochastic trajectories (with $N_e = 1{,}000$) starting at $(x, y, z) = (0.1, 0.1, 0.4)$ and $(x, y, z) = (1, 1, 4)$ are added to the corresponding deterministic phase representation in subfigures (**1b**) and (**2b**), respectively

unstable manifolds. In Fig. 11.8 (1b) and (2b), the deterministic phase space picture is complemented by two stochastic trajectories starting at $(x, y, z) = (0.1, 0.1, 0.4)$ and $(x, y, z) = (1, 1, 4)$. The diffusion part of the stochastic solutions is analogous to the full model presented in Sect. 11.2.4 (with $N_e = 1000$). We find again that, in particular due to the small diffusion part, these stochastic trajectories condense around the deterministic stable manifolds and converge to the asymptotically stable deterministic equilibrium, cf. [4, 12].

Due to the structural analogy to system Eqs. (11.8)–(11.9), the equilibria of our new system Eqs. (11.10)–(11.12) follow immediately:

Theorem 11.2 (Existence of Equilibria). *Besides the point at infinity, for admissible parameters $l_i, w_i, d_i > 0$ $(i = 0, 1)$, $\alpha_0, \beta_0 > 0$ and $m \in \mathbb{N}/2\mathbb{N}$ the following points in $\mathbb{R}^3$ are equilibria of the system Eqs. (11.10)–(11.12):*

- *the origin $(x, y, z) = (0, 0, 0)$ for all choices of the parameter.*
- *the point $(x, 0, z) = (x_s, 0, \alpha_0^{-1}\beta_0 x_s)$, with $x_s := \sqrt[m]{d_0^{-1}(l_0 - d_0)}$, provided $l_0 - d_0 > 0$ holds.*

- *the point* $(0, y, 0) = (0, y_s, 0)$*, with* $y_s := \sqrt[m]{d_1^{-1}(l_1 - d_1)}$*, provided* $l_1 - d_1 > 0$ *holds.*
- *the point* $(x, y, z) = (x_f, y_f, \alpha_0^{-1}\beta_0 x_f)$*, where*

$$x_f := \frac{x_s - w_0 y_s}{1 - w_0 w_1}, \quad \text{and} \quad y_f := \frac{y_s - w_1 x_s}{1 - w_0 w_1},$$

  provided that $x_f, y_f > 0$*,* $w_0 w_1 \neq 1$ *and* x_s, y_s *exist as defined above.*
- *the line of equilibria* $(x, y, z) = (x, w_2(x_s - x), \alpha_0^{-1}\beta_0 x)$*, if* $w_0 w_1 = 1$ *and* x_s *exists as defined above, for all* x *such that* $x_s > x$*.*

Moreover, these are the only candidates for equilibria in the first octant.

Proof. Compared to Theorem 11.1 the only new component is the dynamics on the z-coordinate. Here, $z^* = \alpha_0^{-1}\beta_0 x^*$ has to hold for any equilibrium of the system Eqs. (11.10)–(11.12). The remaining assertions follow analogously to Theorem 11.1. □

Let $K(x^*, y^*, z^*) \in \mathbb{R}^3$ denote the Jacobi-matrix of the right-hand side of system Eqs. (11.10)–(11.12) evaluated at one of the equilibria (x^*, y^*, z^*) stated in Theorem 11.2.

The right-hand side of system Eqs. (11.8)–(11.9) defines the Jacobian $J(x, y) = (J_{i,j}(x, y))_{i,j=1,2}$ (cf. Sect. 11.3.1) which leads to

$$K(x^*, y^*, z^*) = \begin{pmatrix} J_{1,1}(x^*, y^*) - \beta_0 & J_{1,2}(x^*, y^*) & \alpha_0 \\ J_{2,1}(x^*, y^*) & J_{2,2}(x^*, y^*) & 0 \\ \beta_0 & 0 & -\alpha_0 \end{pmatrix}.$$

By suppressing the arguments (x^*, y^*, z^*) and applying Sarrus's rule, the characteristic polynomial $\chi_K(\lambda)$ can be written as

$$\chi_K(\lambda) = (J_{2,2} - \lambda)\left(-(J_{1,1} - \beta_0 - \lambda)(\alpha_0 + \lambda) - \alpha_0\beta_0\right) + J_{1,2}J_{2,1}(\alpha_0 + \lambda).$$

Hence, if $J_{1,2} = 0$ or $J_{2,1} = 0$ as in the case of an equilibrium at the origin or at the points $(x_s, 0, \alpha_0^{-1}\beta_0 x_s)$ and $(0, y_s, 0)$, the eigenvalues λ_i $(i = 1, 2, 3)$ of K follow immediately as

$$\lambda_2 = J_{2,2}, \text{ and } \lambda_{1/3} = \frac{1}{2}\left(J_{1,1} - \alpha_0 - \beta_0 \pm \sqrt{(J_{1,1} - \alpha_0 - \beta_0)^2 + 4\alpha_0 J_{1,1}}\right).$$

Because the systems' behavior with respect to the eigenvalue $\lambda_2 = J_{2,2}$ has already been analyzed in Sect. 11.3.1, we proceed with the discussion of λ_1 and λ_3 at the origin and at the points $(x_s, 0, \alpha_0^{-1}\beta_0 x_s)$ and $(0, y_s, 0)$.

Proposition 11.6 (Linearized Stability of the Origin in the x-z-Plane). *The eigenvalues λ_1 and λ_3 of the linearization of the system Eqs. (11.10)–(11.12) at the origin*

- *are real and of opposite sign if $l_0 - d_0 > 0$ (saddle structure),*
- *are zero and a negative value if $l_0 - d_0 = 0$ (stable structure),*
- *have negative real part if $l_0 - d_0 < 0$ (asymptotically stable structure). In particular, λ_1 and λ_3 are complex conjugate if $4\alpha_0|l_0 - d_0| > (|l_0 - d_0| + \alpha_0 + \beta_0)^2$.*

Proof. The assertions follow immediately as $J_{11} = l_0 - d_0$. □

Proposition 11.7 (Linearized Stability of $(0, y_s, 0)$ and $(x_s, 0, \alpha_0^{-1}\beta_0 x_s)$ in the x-z-Plane). *For $i = 0, 1$, the eigenvalues λ_1 and λ_3 of the linearization of the system Eqs. (11.10)–(11.12) at the point $(0, y_s, 0)$ and at the point $(x_s, 0, \alpha_0^{-1}\beta_0 x_s)$, respectively,*

- *are real and of opposite sign if $l_i - d_i < 0$ (saddle structure),*
- *are zero and a negative value if $l_i - d_i = 0$ (stable structure),*
- *have negative real part if $l_i - d_i > 0$ (asymptotically stable structure).*

Here, the index $i = 0$ corresponds to the discussion of the point $(x_s, 0, \alpha_0^{-1}\beta_0 x_s)$ and the index $i = 1$ to that of the point $(0, y_s, 0)$.

Proof. The assertions follow immediately as $J_{11} = -ml_i(l_i - d_i)l_i^{-1}$ for $i = 0, 1$. □

Remark 11.1. Mathematically, treatment of the wild-type CML clone would thus correspond to a variation of the parameters w_i, l_i, d_i $(i = 0, 1)$ in a way that $(x_s, 0, \alpha_0^{-1}\beta_0 x_s)$ becomes an asymptotically stable equilibrium and the remaining equilibria are unstable. Hence, one would try to affect, by suitable medication, the death and birth rates such that $l_0 - d_0 > 0$ and $l_1 - d_1 > 0$ as well as $w_0 w_1 > 1$ to create an unstable mixed-species equilibrium at $(x_f, y_f, \alpha_0^{-1}\beta_0 x_f))$ (if it exists), cf. Proposition 11.9.

The only difference between model A [Eqs. (11.8)–(11.9)] and model B [Eqs. (11.10)–(11.12)] is the inclusion of a pool of quiescent stem cells z in model B that is in continuous exchange with the cycling normal stem cells x. According to Theorem 2, extension of model A by quiescent stem cells to yield model B does not lead to the formation of additional equilibria. The structure of the terms describing deactivation of normal cells and activation of quiescent cells, respectively, implies that a steady state is established. The size of the residual long-time population of z is determined by the parameters of model B.

The discussion of the biological relevance of the solutions of Eqs. (11.8)–(11.9) obtained for different parameter regimes at the end of Sect. 11.3.1 also applies to Eqs. (11.10)–(11.12), since the dependence of the relative topology of the equilibria positions in (x,y) space on the bifurcation parameters γ_x and γ_y as documented for model A in Fig. 11.6 is not changed by adopting model B.

We now address the stability of the line of equilibria and that of the point $(x_f, y_f, \alpha_0^{-1}\beta_0 x_f)$. In this case, the characteristic polynomial $\chi_K(\lambda)$ can be expressed via $\mathrm{tr}(J)$ and $\det(J)$ leading to

$$p(\lambda) \;=\; \lambda^3 - (\mathrm{tr}(J) - \alpha_0 - \beta_0)\,\lambda^2 - (\alpha_0 \mathrm{tr}(J) - \det(J) + \beta_0 J_{2,2})\,\lambda + \alpha_0 \det(J)\,,$$

with $p(\lambda) = -\chi_K(\lambda)$.

Proposition 11.8 (Linearized Stability of the Line of Equilibria). *Let $w_0 w_1 = 1$ and x_s be admissible by the parameters. Then the Jacobi-matrix K of system Eqs. (11.10)–(11.12) evaluated at an interior element of the line of equilibria $(x, w_2(x_s - x), \alpha_0^{-1}\beta_0 x)$ $(x \in (0, x_s))$ has a vanishing eigenvalue as well as two eigenvalues (counted by multiplicity) with negative real part.*

Proof. Since $\det(J)$ is the product and $\mathrm{tr}(J)$ the sum of J's eigenvalues, we arrive, due to Proposition 11.3, for $J(x_s, w_2(x_s - x))$ at the following properties of J: $\det(J) = 0$, $\mathrm{tr}(J) = J_{1,1} + J_{2,2} < 0$ and $J_{1,1}, J_{2,2} < 0$. Hence, $p(\lambda)$ becomes

$$p(\lambda) \;=\; \lambda\left(\lambda^2 - (\mathrm{tr}(J) - \alpha_0 - \beta_0)\,\lambda - \alpha_0 \mathrm{tr}(J) - \beta_0 J_{2,2}\right)\,,$$

and the usual solution formula for quadratics leads to

$$\lambda_{2,3} \;=\; \tfrac{1}{2}\left(\mathrm{tr}(J) - \alpha_0 - \beta_0 \pm \sqrt{(\mathrm{tr}(J) - \alpha_0 - \beta_0)^2 + 4\,(\alpha_0 \mathrm{tr}(J) + \beta_0 J_{2,2})}\right)\,.$$

This confirms the assertion. □

Proposition 11.9 (Linearized Stability of the Point $(x_f, y_f, \alpha_0^{-1}\beta_0 x_f)$). *Let x_f and y_f be admissible, i.e., $w_0 w_1 \neq 1$. The equilibrium $(x_f, y_f, \alpha_0^{-1}\beta_0 x_f)$ is hyperbolic. It is asymptotically stable for $w_0 w_1 < 1$ and unstable for $w_0 w_1 > 1$.*

Proof. Proposition 11.4 implies that $J_{1,1}, J_{2,2} < 0$, $\mathrm{tr}(J) < 0$ as well as $\det(J) < 0$ if $w_0 w_1 > 1$ and $\det(J) > 0$ if $w_0 w_1 < 1$. Let the coefficients of $p(\lambda)$ be identified by $\kappa_0, \kappa_1, \kappa_2$ such that $p(\lambda) = \lambda^3 + \kappa_2\lambda^2 + \kappa_1\lambda + \kappa_0$. As $\kappa_0 \neq 0$ none of the roots of $p(\lambda)$ vanishes.

If $\kappa_0, \kappa_2 > 0$ and $\kappa_1\kappa_2 > \kappa_0$, then all roots of $p(\lambda)$ are negative or have negative real part due to the Routh–Hurwitz criterion. Let $w_0 w_1 < 1$, then $\kappa_0, \kappa_1, \kappa_2 > 0$ and

$$\begin{aligned}\kappa_1\kappa_2 - \kappa_0 &= (\alpha_0 \mathrm{tr}(J) - \det(J) + \beta_0 J_{2,2})\,(\mathrm{tr}(J) - \alpha_0 - \beta_0) - \alpha_0 \det(J) \\ &= \alpha_0(\mathrm{tr}(J))^2 - \alpha_0(\alpha_0 + \beta_0)\mathrm{tr}(J) - \mathrm{tr}(J)\det(J) + \beta_0 \det(J) \\ &\quad + \beta_0 J_{2,2}\mathrm{tr}(J) - \beta_0(\alpha_0 + \beta_0)J_{2,2} \;>\; 0\,.\end{aligned}$$

Thus, the assertion follows for $w_0 w_1 < 1$.

For $w_0 w_1 > 1$ it follows that $\kappa_2 > 0$ and $\kappa_0 < 0$ holds. Hence, due to Descartes sign rule, there is exactly one positive real root of $p(\lambda)$. □

The bifurcation pattern is analogous to that described in Sect. 11.3.1 in the context of the dynamics of two competing stem cell strains.

11.3.3 *Model C: The Full Four-Dimensional Problem, Including Cycling and Noncycling Normal Stem Cells Plus Two Cycling Leukaemic Stem Cell Clones*

We are now ready to investigate the equilibria of the initial four-dimensional model equations (11.4)–(11.7). We recall the governing equations for model C with some notational modifications:

$$\dot{x}_0 = \left(\frac{l_0}{1 + (x_0 + w_{0,1}x_1 + w_{0,2}x_2)^m} - d_0 - \beta_0 \right) x_0 + \alpha_0 x_3 \,, \tag{11.13}$$

$$\dot{x}_1 = \left(\frac{l_1}{1 + (x_1 + w_{1,0}x_0 + w_{1,2}x_s)^m} - d_1 \right) x_1 \,, \tag{11.14}$$

$$\dot{x}_2 = \left(\frac{l_2}{1 + (x_2 + w_{2,0}x_0 + w_{2,1}x_1)^m} - d_2 \right) x_2 \,, \tag{11.15}$$

$$\dot{x}_3 = -\alpha_0 x_3 + \beta_0 x_0 \,, \tag{11.16}$$

with $l_i, d_i > 0$ $(i = 0, 1, 2)$, $w_{0,1}, w_{0,2}, w_{1,0}, w_{1,2}, w_{2,0}, w_{2,1} > 0$, $\alpha_0, \beta_0 > 0$ and $m \in \mathbb{N}/2\mathbb{N}$.

Due to the previous considerations in Sects. 11.3.1 and 11.3.2, combinatorics can be employed to analyze the existence and location of the equilibria.

Theorem 11.3 (Existence of Equilibria). *Besides the point at infinity, for admissible parameters $l_i, w_i, d_i > 0$ $(i = 0, 1, 2)$, $\alpha_0, \beta_0 > 0$ and $m \in \mathbb{N}/2\mathbb{N}$ the following points in $\mathbb{R}^4$ are equilibria of the system Eqs. (11.13)–(11.16):*

- *the origin $(x_0, x_1, x_2, x_3) = (0, 0, 0, 0)$ for all choices of the parameter.*
- *the point $(x_0, 0, 0, x_3) = (x_s, 0, 0, \alpha_0^{-1}\beta_0 x_s)$, with $x_s := \sqrt[m]{d_0^{-1}(l_0 - d_0)}$, provided $l_0 - d_0 > 0$ holds.*
- *the point $(0, x_1, 0, 0) = (0, y_s^{(1)}, 0, 0)$, with $y_s^{(1)} := \sqrt[m]{d_1^{-1}(l_1 - d_1)}$, provided $l_1 - d_1 > 0$ holds, and the point $(0, 0, x_2, 0) = (0, 0, y_s^{(2)}, 0)$, with $y_s^{(2)} := \sqrt[m]{d_2^{-1}(l_2 - d_2)}$, provided $l_2 - d_2 > 0$ holds.*
- *the points $(x_0, x_1, 0, x_3) = (x_f^{(1)}, y_f^{(1)}, 0, \alpha_0^{-1}\beta_0 x_f^{(1)})$, and $(x_0, 0, x_2, x_3) = (x_f^{(2)}, 0, y_f^{(2)}, \alpha_0^{-1}\beta_0 x_f^{(2)})$, where*

$$x_f^{(i)} := \frac{x_s - w_{0,i} y_s^{(i)}}{1 - w_{0,i} w_{i,0}}, \quad \textit{and} \quad y_f^{(i)} := \frac{y_s^{(i)} - w_{i,0} x_s}{1 - w_{0,i} w_{i,0}},$$

provided that $x_f^{(i)}, y_f^{(i)} > 0$, $w_{0,i} w_{i,0} \neq 1$ *and* $x_s, y_s^{(i)}$ $(i = 1, 2)$ *exist as defined above.*

- *the line of equilibria* $(x_0, x_1, 0, x_3) = (x_0, w_{1,0}(x_s - x_0), 0, \alpha_0^{-1}\beta_0 x_0)$, *if* $w_{0,1} w_{1,0} = 1$, *and the line of equilibria* $(x_0, 0, x_2, x_3) = (x_0, 0, w_{2,0}(x_s - x_0), \alpha_0^{-1}\beta_0 x_0)$, *if* $w_{0,2} w_{2,0} = 1$, *provided* x_s *exists as defined above, for all* x_0 *such that* $x_s > x_0$.
- *the point* $(x_0, x_1, x_2, x_3) = (y_1, y_2, y_3, \alpha_0^{-1}\beta_0 y_1)$, *provided that the following conditions hold:*

  (i) $x_s, y_s^{(1)}, y_s^{(2)} > 0$ *exist as defined above,*

  (ii) $1 - w_{1,0} w_{0,1} \neq 0$ *and* $y_s^{(2)} - w_{2,0} x_s - \frac{y_s^{(1)} - w_{1,0} x_s}{1 - w_{1,0} w_{0,1}} \neq 0$ *together with* $1 - w_{2,0} w_{0,2} - \frac{w_{1,2} - w_{1,0} w_{0,2}}{1 - w_{1,0} w_{0,1}} \neq 0$, *or*

  (ii') $1 - w_{1,0} w_{0,1} = 0$ *and* $w_{1,2} - w_{1,0} w_{0,2} \neq 0$ *and* $w_{2,1} - w_{2,0} w_{0,1} \neq 0$,

  (iii) $y = (y_1, y_2, y_3)^T$ *is the unique solution of* $Wy = (x_s, y_s^{(1)}, y_s^{(2)})^T$, *where* $W = (w_{i,j})$ *with* $w_{i,i} = 1$ $(i = 1, 2, 3)$, *and finally*

  (iv) $y_1, y_2, y_3 > 0$. *Note, due to condition (ii) or (ii') the system* $Wy = (x_s, y_s^{(1)}, y_s^{(2)})^T$ *has an unique solution.*

- *the one-parameter family* $(x_0, x_1, x_2, x_3) = (y_0(z), y_0(z), z, \alpha_0^{-1}\beta_0 y_0(z))$ *of equilibria, where* $z > 0$ *is such that*

$$y_1(z) = \frac{y_s^{(2)} - w_{2,0} x_s - (1 - w_{2,0} w_{0,2}) z}{w_{2,1} - w_{2,0} w_{0,1}} > 0$$

and

$$y_0(z) = x_s - w_{0,1} y_1(z) - w_{0,2} z > 0$$

hold, provided that $1 - w_{0,1} w_{1,0} = 0$, $w_{1,2} - w_{1,0} w_{0,2} = 0$, $y_s^{(1)} - w_{1,0} x_s = 0$, *together with* $w_{2,1} - w_{2,0} w_{0,1} \neq 0$.

- *the one-parameter family* $(x_0, x_1, x_2, x_3) = (y_0(z), z, y_2, \alpha_0^{-1}\beta_0 y_0(z))$ *of equilibria, where* $z > 0$ *is such that*

$$y_0(z) = x_s - w_{0,1} z - w_{0,2} \frac{y_s^{(1)} - w_{1,0} x_s}{w_{1,2} - w_{1,0} w_{0,2}} > 0$$

holds, provided that

(i) $1 - w_{0,1} w_{1,0} = 0$, $w_{1,2} - w_{1,0} w_{0,2} \neq 0$ *together with* $w_{2,1} - w_{2,0} w_{0,1} = 0$,

(ii) $(1 - w_{2,0} w_{0,2})(y_s^{(1)} - w_{1,0} x_s) = (w_{1,2} - w_{1,0} w_{0,2})(y_s^{(2)} - w_{2,0} x_s)$, *and*

(iii) $y_2 := \frac{y_s^{(1)} - w_{1,0} x_s}{w_{1,2} - w_{1,0} w_{0,2}} > 0$.

- *the one-parameter family* $(x_0, x_1, x_2, x_3) = (y_0(z), z, y_2, \alpha_0^{-1}\beta_0 y_0(z))$ *of equilibria, where* $z > 0$ *is such that*

$$y_0(z) = x_s - w_{0,1}z - w_{0,2}\frac{y_s^{(2)} - w_{2,0}x_s}{1 - w_{2,0}w_{0,2}} > 0$$

holds, provided that

(i) $y_2 = \frac{y_s^{(2)} - w_{2,0}x_s}{1 - w_{2,0}w_{0,2}} > 0$,

(ii) $1 - w_{0,1}w_{1,0} = 0$, $w_{1,2} - w_{1,0}w_{0,2} = 0$ *together with* $y_s^{(1)} - w_{1,0}x_s = 0$, *and*

(iii) $w_{2,1} - w_{2,0}w_{0,1} = 0$ *with* $1 - w_{2,0}w_{0,2} > 0$.

- *the two-parameter family* $(x_0, x_1, x_2, x_3) = (y_0(z), z_1, z_2, \alpha_0^{-1}\beta_0 y_0(z))$ *of equilibria, where* $z_1, z_2 > 0$ *are such that* $y_0(z_1, z_2) = x_s - w_{0,1}z_1 - w_{0,2}z_2 > 0$, *provided that*

(i) $1 - w_{0,1}w_{1,0} = 0$, $w_{1,2} - w_{1,0}w_{0,2} = 0$ *together with* $y_s^{(1)} - w_{1,0}x_s = 0$, *as well as*

(ii) $w_{2,1} - w_{2,0}w_{0,1} = 0$, $1 - w_{2,0}w_{0,2} = 0$ *together with* $y_s^{(2)} - w_{2,0}x_s = 0$.

Moreover, these are the only candidates for equilibria in the first orthant $\mathscr{P} := \{(x_0, x_1, x_2, x_3)^T \in \mathbb{R}^4 : x_i \geq 0 \text{ for all } i = 0, 1, 2, 3\}$.

Proof. Theorems 11.1 and 11.2 show the existence of the postulated equilibria if at least one of the four components x_0, x_1, x_2, or x_3 vanishes.

Let us now assume that $x_0, x_1, x_2, x_3 \neq 0$, and denote the coordinates of an equilibrium by $(y_0, y_1, y_2, y_3)^T$. The x_0- and x_3-coordinates at an equilibrium are coupled via $y_3 = \alpha_0^{-1}\beta_0 y_0$. Thus, we only have to care about the first three coordinates $y := (y_0, y_1, y_2)^T$ of an equilibrium point: at an equilibrium, we have with $i = 0, 1, 2$

$$\frac{l_i}{1 + (w_i^T y)^m} - d_i = 0 \Leftrightarrow \delta_i := \sqrt[m]{\frac{l_i - d_i}{d_i}} = w_i^T y\,,$$

where $w_0 := (1, w_{0,1}, w_{0,2})^T$, $w_1 := (w_{1,0}, 1, w_{1,2})^T$, and $w_2 := (w_{2,0}, w_{2,1}, 1)^T$. Thus, we are left with solving the linear system $Wy = \delta$ with $W := (w_0^T, w_1^T, w_2^T) \in \mathbb{R}^{3\times 3}$ and $\delta = (\delta_0, \delta_1, \delta_2)^T = (x_s, y_s^{(1)}, y_s^{(2)})^T \in \mathbb{R}^3$. By Gaussian elimination we have

$$\left(\begin{array}{ccc|c} 1 & w_{0,1} & w_{0,2} & x_s \\ w_{1,0} & 1 & w_{1,2} & y_s^{(1)} \\ w_{2,0} & w_{2,1} & 1 & y_s^{(2)} \end{array}\right)$$

$$\xrightarrow{(1)} \left(\begin{array}{ccc|c} 1 & w_{0,1} & w_{0,2} & x_s \\ 0 & 1 - w_{1,0}w_{0,1} & w_{1,2} - w_{1,0}w_{0,2} & y_s^{(1)} - w_{1,0}x_s \\ 0 & w_{2,1} - w_{2,0}w_{0,1} & 1 - w_{2,0}w_{0,2} & y_s^{(2)} - w_{2,0}x_s \end{array}\right).$$

If first $1 - w_{1,0}w_{0,1} \neq 0$, we continue to obtain

$$\xrightarrow{(2)} \left(\begin{array}{cc|c} 1 - w_{1,0}w_{0,1} & w_{1,2} - w_{1,0}w_{0,2} & y_s^{(1)} - w_{1,0}x_s \\ 0 & 1 - w_{2,0}w_{0,2} - \frac{w_{1,2} - w_{1,0}w_{0,2}}{1 - w_{1,0}w_{0,1}} & y_s^{(2)} - w_{2,0}x_s - \frac{y_s^{(1)} - w_{1,0}x_s}{1 - w_{1,0}w_{0,1}} \end{array} \right),$$

and thus, if second $y_s^{(2)} - w_{2,0}x_s - \frac{y_s^{(1)} - w_{1,0}x_s}{1 - w_{1,0}w_{0,1}} \neq 0$ together with $1 - w_{2,0}w_{0,2} - \frac{w_{1,2} - w_{1,0}w_{0,2}}{1 - w_{1,0}w_{0,1}} \neq 0$, the three coordinates of an equilibrium follow as

$$\begin{aligned} y_2 &= \frac{y_s^{(2)} - w_{1,0}w_{0,1}y_s^{(2)} - w_{2,0}x_s + w_{1,0}w_{0,1}w_{2,0}x_s - y_s^{(1)} + w_{1,0}x_s}{1 - w_{1,0}w_{0,1} - w_{2,0}w_{0,2} + w_{1,0}w_{0,1}w_{2,0}w_{0,2} - w_{1,2} + w_{1,0}w_{0,2}} \\ y_1 &= \frac{y_s^{(1)} - w_{1,0}x_s - (w_{1,2} - w_{1,0}w_{0,2})y_2}{1 - w_{1,0}w_{0,1}} \\ y_0 &= x_s - w_{0,1}y_1 - w_{0,2}y_2 \,. \end{aligned}$$

Before we consider further singular cases, we have to check for positivity, i.e., $y_0, y_1, y_2 > 0$. This is guaranteed by the conditions imposed via the theorem. Note, as the entries of W are positive and independent of the entries of δ, we can construct parameter values l_i, d_i $(i = 0, 1, 2)$ such that for any choice of positive entries of y the system $Wy = \delta$ has an unique solution.

Next, let $1 - w_{0,1}w_{1,0} = 0$, then the Gaussian elimination procedure proceeds according to

$$\xrightarrow{(2')} \left(\begin{array}{cc|c} 0 & w_{1,2} - w_{1,0}w_{0,2} & y_s^{(1)} - w_{1,0}x_s \\ w_{2,1} - w_{2,0}w_{0,1} & 1 - w_{2,0}w_{0,2} & y_s^{(2)} - w_{2,0}x_s \end{array} \right).$$

If, additionally, $w_{1,2} - w_{1,0}w_{0,2} \neq 0$ and $w_{2,1} - w_{2,0}w_{0,1} \neq 0$, then

$$\begin{aligned} y_2' &= \frac{y_s^{(1)} - w_{1,0}x_s}{w_{1,2} - w_{1,0}w_{0,2}} \\ y_1' &= \frac{y_s^{(2)} - w_{2,0}x_s - (1 - w_{2,0}w_{0,2})y_s^{(1)}}{w_{2,1} - w_{2,0}w_{0,1}} \\ y_0' &= x_s - w_{0,1}y_1' - w_{0,2}y_2' \,. \end{aligned}$$

Again, we moreover require $y_2', y_1', y_0' > 0$.

If $1 - w_{0,1}w_{1,0} = 0$ and $w_{1,2} - w_{1,0}w_{0,2} = 0$, then $y_s^{(1)} - w_{1,0}x_s = 0$ is required to allow for a solution. Under these conditions let $z := y_2 > 0$ be arbitrary, for $w_{2,1} - w_{2,0}w_{0,1} \neq 0$ we get a one-parameter family of equilibria as

$$y_1(z) = \frac{y_s^{(2)} - w_{2,0}x_s - (1 - w_{2,0}w_{0,2})z}{w_{2,1} - w_{2,0}w_{0,1}}$$
$$y_0(z) = x_s - w_{0,1}y_1 - w_{0,2}z\,.$$

provided $y_1, y_0 > 0$.

If $1 - w_{0,1}w_{1,0} = 0$, $w_{1,2} - w_{1,0}w_{0,2} \neq 0$ and $w_{2,1} - w_{2,0}w_{0,1} = 0$, then

$$y_2 \;=\; \frac{y_s^{(1)} - w_{1,0}x_s}{w_{1,2} - w_{1,0}w_{0,2}}\,,$$

and

$$(1 - w_{2,0}w_{0,2})(y_s^{(1)} - w_{1,0}x_s) \;=\; (w_{1,2} - w_{1,0}w_{0,2})(y_s^{(2)} - w_{2,0}x_s)\,,$$

has to hold. Finally, let $z := y_1 > 0$ be arbitrary, then

$$y_0(z) \;=\; x_s - w_{0,1}z - w_{0,2}\frac{y_s^{(1)} - w_{1,0}x_s}{w_{1,2} - w_{1,0}w_{0,2}}\,.$$

Again, for all combinations we further require $y_2, y_1 = z, y_0(z) > 0$.

Let $1 - w_{0,1}w_{1,0} = 0$, $w_{1,2} - w_{1,0}w_{0,2} = 0$ together with $y_s^{(1)} - w_{1,0}x_s = 0$, and $w_{2,1} - w_{2,0}w_{0,1} = 0$ with $1 - w_{2,0}w_{0,2} > 0$, then

$$y_2 \;=\; \frac{y_s^{(2)} - w_{2,0}x_s}{1 - w_{2,0}w_{0,2}}\,.$$

With $z := y_1 > 0$ it follows that

$$y_0(z) \;=\; x_s - w_{0,1}z - w_{0,2}\frac{y_s^{(2)} - w_{2,0}x_s}{1 - w_{2,0}w_{0,2}}\,.$$

For all combinations we additionally require $y_2, y_1 = z, y_0(z) > 0$.

Finally, let $1 - w_{0,1}w_{1,0} = 0$, $w_{1,2} - w_{1,0}w_{0,2} = 0$ together with $y_s^{(1)} - w_{1,0}x_s = 0$, and $w_{2,1} - w_{2,0}w_{0,1} = 0$, $1 - w_{2,0}w_{0,2} = 0$ together with $y_s^{(2)} - w_{2,0}x_s = 0$. Then, with $z_1 > 0$ and z_2 we get the two-parameter family of feasible solutions

$$y_0(z_1, z_2) \;=\; x_s - w_{0,1}z_1 - w_{0,2}z_2\,,$$

provided $y_0(z_1, z_2) > 0$.

No further options that would lead to equilibria are given. □

Not unsurprisingly, we recover essentially the same types of isolated and connected equilibria as in Theorem 11.2, though with one additional dimension

for the additional leukaemic species. In principle, the stability of each of the 13 equilibrium structures of Theorem 11.3 can now be obtained analogously to those discussed in Theorem 11.2.

11.4 Summary and Outlook

Starting from first principles with the assumption of Hill-function-like signaling, we derived the approximated Fokker–Planck and its associated Langevin equation for the dynamics of a four-dimensional system consisting of cycling and noncycling normal hematopoietic stem cells together with the wild-type and one *imatinib*-resistant cycling CML stem cell clone (model C). Since the equilibria of the corresponding deterministic system are known to have a non-negligible influence on the stochastic dynamics, we determined the location of all those equilibria in the full four-dimensional setting. Moreover, a two-dimensional (model A) and a three-dimensional (model B) deterministic subsystem were studied with respect to stability of all equilibria and complete bifurcation behavior.

The discussion of the bifurcation behavior of the solutions of models A, B, and C confirms that it is possible to direct the equation systems into single-state equilibrium situations characterized by elimination of CML stem cells within certain parameter regimes. Treatment is assumed to have the effect of reducing and increasing the growth and death rate, respectively, of CML stem cells by appropriate medication.

Our study was motivated by current research on *imatinib*-resistant CML strains that appear after treatment of the wild-type leukaemic clone with *imatinib* [20, 21, 53]. The models proposed here will hopefully serve, in further studies, to better understand the emergence of *imatinib*-resistant CML mutations and allow for an effective medication strategy. A relevant question on the modeling side is the best form of the signaling function, as there are some indicators that a signaling model employing Tsallis-functions may resemble observed statistical patterns better. An interesting biological aspect that will be addressed in future theoretical work is the issue of the involvement of quiescent stem cells in the development of CML, in particular with respect to the mechanisms that drive the emergence of *imatinib*-resistant CML clones after successful elimination of the wild-type leukaemic strain by this drug.

Acknowledgements C.W. would like to thank the Mohn Foundation, the Centre for Theoretical and Computational Chemistry (CTCC) at the University of Tromsø and the Research Council of Norway (Grant Nr. 177558/V30) for continued support.

We also acknowledge computational resources provided by Norwegian High Performance Computing (NOTUR).

Special thanks go to Jürgen Scheurle, who brought F.R. back to academia.

Appendix: Milstein Scheme for the Stochastic Itô-/Langevin-Model Presented in Sect. 11.2.4

With suitable functions $f_0, \ldots, f_q$ for the drift and $g_0, \ldots, g_q$ for the diffusion our four-dimensional system of stochastic differential Itô equations from Sect. 11.2.4 reads as

$$\mathrm{d}X_t = \begin{pmatrix} f_0(X_t) \\ f_1(X_t) \\ f_2(X_t) \\ f_q(X_t) \end{pmatrix} \mathrm{d}t + \begin{pmatrix} g_0(X_t) \\ 0 \\ 0 \\ 0 \end{pmatrix} \mathrm{d}W_t^{(0)} + \begin{pmatrix} 0 \\ g_1(X_t) \\ 0 \\ 0 \end{pmatrix} \mathrm{d}W_t^{(1)} + \begin{pmatrix} 0 \\ 0 \\ g_2(X_t) \\ 0 \end{pmatrix} \mathrm{d}W_t^{(2)} + \begin{pmatrix} -g_q(X_t) \\ 0 \\ 0 \\ g_q(X_t) \end{pmatrix} \mathrm{d}W_t^{(q)}.$$

As discussed in [28, p. 345], a numerical scheme with strong order of convergence one is the so-called (explicit) Milstein-method. The order one of strong convergence depends, of course, on the right-hand side of the stochastic differential equation considered, cf. Theorem 10.3.5, [28, p. 350]. In particular, for our square-root diffusion functions this order of convergence will hold true only outside the origin. Due to Itô's chain rule, some additional terms and, in particular in our multi-noise setting, multiple stochastic integrals occur compared to the familiar deterministic Euler-method. Nevertheless, the Milstein-method as a first order scheme, like the deterministic Euler-method, can be considered as the proper stochastic analogue to the deterministic Euler-method.

For our system of stochastic differential equations Milstein's method leads to the following time-discretization with $t_{n+1} - t_n =: \Delta t$ is the spacing between two time-points of an equidistant time grid, and $W_{t_{n+1}}^{(i)} - W_{t_n}^{(i)} =: \Delta W^{(i)}$ the corresponding increment of the Wiener process $W_t^{(i)}$ $(i = 0, 1, 2, q)$. First, we get for the normal stem cells

$$\begin{aligned} x_{s,0}(t_{n+1}) = x_{s,0}(t_n) &+ f_0(X(t_n))\Delta t + g_0(X(t_n))\Delta W^{(0)} \\ &+ \sum_{i=0}^{2} \left(g_i(X(t_n)) \left. \frac{\partial}{\partial x_{s,i}} g_0(X) \right|_{t_n} \right) I(i,0) \\ &+ \left(g_q(X(t_n))I(q,0) - g_q(X(t_n))I(0,0) \right) \left. \frac{\partial}{\partial x_{s,q}} g_0(X) \right|_{t_n}, \end{aligned}$$

where $X := (x_{s,0}, x_{s,1}, x_{s,2}, x_q)$, and $I(i,j)$ $(i,j \in \{0,1,2,q\})$ are multiple stochastic integrals given as

$$I(i,j) \;=\; \int_{t_n}^{t_{n+1}} \int_{t_n}^{s_1} \mathrm{d}W_{s_2}^{(i)} \mathrm{d}W_{s_1}^{(j)} \,,$$

and, in particular,

$$I(i,i) \;=\; \tfrac{1}{2} \left(\left(\Delta W^{(i)} \right)^2 - \Delta t \right) .$$

For model A, we obtain with this notation

$$\begin{aligned} x_{s,1}(t_{n+1}) = x_{s,1}(t_n) + f_1(X(t_n))\Delta t + g_1(X(t_n))\Delta W^{(1)} \\ + \sum_{i=0}^{2} \left(g_i(X(t_n)) \left. \frac{\partial}{\partial x_{s,i}} g_1(X) \right|_{t_n} \right) I(i,1) \,, \end{aligned}$$

and

$$\begin{aligned} x_{s,2}(t_{n+1}) = x_{s,2}(t_n) + f_2(X(t_n))\Delta t + g_2(X(t_n))\Delta W^{(1)} \\ + \sum_{i=0}^{2} \left(g_i(X(t_n)) \left. \frac{\partial}{\partial x_{s,i}} g_2(X) \right|_{t_n} \right) I(i,2) \,. \end{aligned}$$

Finally, the Milstein discretization scheme for the quiescent stem cell population is

$$\begin{aligned} x_q(t_{n+1}) = x_q(t_n) + f_q(X(t_n))\Delta t + g_q(X(t_n))\Delta W^{(q)} \\ + g_0(X(t_n)) \left. \frac{\partial}{\partial x_{s,0}} g_q(X) \right|_{t_n} I(q,0) \\ + \left(g_q(X(t_n)) I(q,q) - g_q(X(t_n)) I(0,q) \right) \left. \frac{\partial}{\partial x_{s,q}} g_q(X) \right|_{t_n} . \end{aligned}$$

The efficient numerical evaluation of multiple stochastic integrals is somewhat challenging, cf. [28, 29, p. 198] or [17]: For one of our standard Wiener processes $W_t^{(i)}$ ($i,j \in \{0,1,2,q\}$), the starting point for the computation of $I(i,j)$, $i \neq j$, is the Brownian bridge process

$$W_t^{(i)} \;-\; \frac{t}{s} W_s^{(i)} \,, \qquad \text{for } 0 \leq t \leq s := \Delta t \,.$$

The associated Fourier series reads as

$$W_t^{(i)} \;-\; \frac{t}{s} W_s^{(i)} \;=\; \tfrac{1}{2} a_{i,0} \;+\; \sum k = 1^{\infty} \left(a_{i,k} \cos\left(\frac{2\pi k t}{s} \right) + a_{i,k} \sin\left(\frac{2\pi k t}{s} \right) \right) ,$$

which is equivalent to

$$W_t^{(i)} \;=\; \frac{1}{s} W_s^{(i)} t + \tfrac{1}{2} a_{i,0} + \sum^{\infty} k = 1 \left(a_{i,k} \cos\left(\frac{2\pi k t}{s} \right) + a_{i,k} \sin\left(\frac{2\pi k t}{s} \right) \right) , \tag{11.17}$$

where $a_{i,0} = -2 \sum_{k=0}^{\infty} a_{i,k}$ by setting $t = 0$ and the coefficients $b_{i,j}$ and $a_{i,j}$ are $\mathcal{N}(0, (2\pi^2 k^2) s)$-distributed pairwise independent random variables. As outlined in [29], this Fourier series can be used to successively derive a hierarchy of multiple stochastic integrals. In particular, it can be shown by first integrating equation (11.17) with respect to t over $[0, s]$ and then with respect to $W_t^{(j)}$ over $[0, s]$, that the following relation is true:

$$I(i,j) \;=\; \tfrac{1}{2} W_s^{(i)} W_s^{(j)} \;-\; \tfrac{1}{2} \left(a_{j,0} W_s^{(i)} - a_{i,0} W_s^{(j)} \right) \;+\; s A_{i,j} \, ,$$

where

$$A_{i,j} \;=\; \tfrac{\pi}{s} \sum_{k=1}^{\infty} k \left(a_{i,k} b_{j,k} - a_{j,k} b_{i,k} \right) \, .$$

To handle these infinite series on a computer, a truncation method is required. First, we observe that

$$\xi_i \;:=\; \tfrac{1}{\sqrt{s}} W_s^{(i)} \, , \qquad \xi_{i,k} \;:=\; \sqrt{\tfrac{2}{s}} \pi k a_{i,k} \, , \qquad \eta_{i,k} \;:=\; \sqrt{\tfrac{2}{s}} \pi k b_{i,k}$$

are independent Gaussian random variables which can be conveniently sampled prior to computation. Let $p \in \mathbb{N}$ denote a truncation index such that $I(i,j) \approx I(i,j)^p$, where the approximation $I(i,j)^p$ of $I(i,j)$ is given as

$$I(i,j)^p \;=\; \tfrac{1}{2} s \xi_i \xi_j \;-\; \tfrac{\sqrt{s}}{2} \left(a_{j,0}^p \xi_i - a_{i,0}^p \xi_j \right) \;+\; s A_{i,j}^p \, ,$$

with

$$A_{i,j}^p \;=\; \tfrac{1}{2\pi} \sum_{k=1}^{p} \tfrac{1}{k} \left(\xi_{i,k} \eta_{j,k} - \xi_{j,k} \eta_{i,k} \right) \quad \text{and} \quad a_{i,0}^p \;=\; -\frac{\sqrt{2s}}{\pi} \sum_{k=1}^{p} \tfrac{1}{k} \xi_{i,k} \, .$$

The mean-square error of this approximation is discussed in [29], and in order to achieve a strong order of convergence one for the Milstein scheme with this approximation of the multiple stochastic integrals. Ref. [17] suggests that p should be chosen of order $\mathcal{O}(s^{-1})$. A method to reduce the number p is for instance proposed in [51].

A very efficient alternative to the rather cumbersome Fourier series ansatz to simulate double/multiple stochastic integrals is to apply the Euler–Mayurama

scheme on each sub-interval with a very fine step size, cf. [26]. For instance, the integral $I_{1,2}$ is the solution of the two-dimensional stochastic differential equation

$$\mathrm{d}X_t \;=\; 1\,\mathrm{d}W_t^{(1)}\,, \qquad \mathrm{d}Y_t \;=\; X_t\,\mathrm{d}W_t^{(2)}\,,$$

such that $I_{1,2} = Y_\Delta$ on $t_n \le t \le t_{n+1}$ with step size $\Delta = t_{n+1} - t_n$. This two-dimensional stochastic differential equation can now, for instance, be solved with the Euler–Mayurama scheme with step size Δ^2.

References

1. Ackleh, A.S., Hu, S.: Comparison between stochastic and deterministic selection-mutation models. Math. Biosci. Eng. **4**, 133–157 (2007)
2. Ackleh, A.S., Hu, S.: Global dynamics of hematopoietic stem cells and differentiated cells in a chronic myeloid leukemia model. J. Math. Biol. **62**, 975–997 (2011)
3. Alon, U.: An Introduction to Systems Biology—Design Principles of Biological Circuits. Mathematical and Computational Biology. Chapman and Hall/CRC/Taylor and Francis Group, Boca Raton (2007)
4. Arnold, L.: Random Dynamical Systems. Springer, Berlin (2003)
5. Bai, F., Wu, Z., Jin, J., Hochendoner, P., Xing, J.: Slow protein conformational change, allostery and network dynamics. In: Cai, W., Hong, H. (eds.) Protein-Protein Interactions - Computational and Experimental Tools. InTech, Shanghai (2012)
6. Buchdunger, E., Zimmermann, J., Mett, H., Meyer, T., Müller, M., Druker, B.J., Lydon, N.B.: Inhibition of the abl protein-tyrosine kinase in vitro and in vivo by a 2-phenylaminopyrimidine derivative. Cancer Res. **56**, 100 (1996)
7. Devys, A., Goudon, T., Lafitte, P.: A model describing the growth and the size distribution of multiple metastatic tumors. Discrete Contin. Dyn. B **12**, 731–767 (2009)
8. Doumic-Jauffet, M., Kim, P.S., Perthame, B.: Stability analysis of a simplified yet complete model for chronic myelogenous leukemia. Bull. Math. Biol. **72**, 1732–1759 (2010)
9. Flå, T., Ahmed, S.H.: Evolution of cold adapted protein sequences. In: Fung, G. (ed.) Sequence and Genome Analysis: Methods and Application. II. iConcept Press Ltd, Brisbane (2011)
10. Foo, J., Drummond, M.W., Clarkson, B., Holyoake, T., Michor, F.: Eradication of chronic myeloid leukemia stem cells: a novel mathematical model predicts no therapeutic benefit of adding g-csf to imatinib. PLoS Comput. Biol. **5**(9), e1000503 (2009). URL http://view.ncbi.nlm.nih.gov/pubmed/19749982
11. Foo, J., Michor, F.: Evolution of resistance to anti-cancer therapy during general dosing schedules. J. Theor. Biol. **263**(2), 179–88 (2010). URL http://view.ncbi.nlm.nih.gov/pubmed/20004211
12. Freidlin, M.I., Wentzell, A.D.: Random Perturbations of Dynamical Systems. Springer, Berlin (1984)
13. Friedman, A.: A hierarchy of cancer models and their mathematical challenges. Discrete Contin. Dyn. B **4**, 147–159 (2004)
14. Friedman, A.: Stochastic Differential Equations and Applications. Dover Publications, New York (2006)
15. Friedman, A., Kim, Y.: Tumor cells proliferation and migration under the influence of their microenvironment. Math. Biosci. Eng. **8**, 371–383 (2011)
16. Fujarewicz, K., Kimmel, M., Swierniak, A.: On fitting of mathematical models of cell signaling pathways using adjoint systems. Math. Biosci. Eng. **2**, 527–534 (2005)

17. Gilsing, H., Shardlow, T.: Sdelab: A package for solving stochastic differential equations in matlab. J. Comput. Appl. Math. **205**, 1002–1018 (2007)
18. Glauche, I., Lorenz, R., Hasenclever, D., Roeder, I.: A novel view on stem cell development: analysing the shape of cellular genealogies. Cell Prolif. **42**(2), 248–263 (2009). URL http://view.ncbi.nlm.nih.gov/pubmed/19254328
19. Glauche, I., Moore, K., Thielecke, L., Horn, K., Loeffler, M., Roeder, I.: Stem cell proliferation and quiescence—two sides of the same coin. PLoS Comput. Biol. **5**(7), e1000447 (2009). URL http://view.ncbi.nlm.nih.gov/pubmed/19629161
20. Gruber, F.X.: Towards a quantitative understanding of cml resistance. Ph.D. Thesis. Department of Pharmacology, University of Tromsø (2009)
21. Gruber, F.X., Ernst, T., Porkka, K., Engh, R., Mikkola, I., Maier, J., Lange, T., Hochhaus, A.: Dynamics of the emergence of dasatinib and nilotinib resistance in imatinib resistant cml patients. Leukemia **26**, 172 (2012). http://dx.doi.org/10.1038/leu.2011.187. URL http://www.nature.com/leu/journal/v26/n1/full/leu2011187a.html
22. Hlavacek, W.S., Faeder, J.R.: The complexity of cell signaling and the need for a new mechanics. Sci. Signal. **2**(81), pe46 (2009). DOI 10.1126/scisignal.281pe46. URL http://stke.sciencemag.org/cgi/content/abstract/sigtrans;2/81/pe46
23. Horsthemke, W., Lefever, R.: Noise-Induced Transitions – Theory ans Applications in Physics, Chemistry, and Biology. Springer, Berlin (1984)
24. Kimura, M.: Diffusion models in population genetics. J. Appl. Probab. **1**(2), 177–232 (1964). URL http://www.jstor.org/stable/3211856
25. Kimura, M., Crow, J.F.: The measurement of effective population number. Evolution **17**(3), 279–288 (1963). URL http://www.jstor.org/stable/2406157
26. Kloeden, P.E.: The systematic derivation of higher order numerical methods for stochastic differential equations. Milan J. Math. **70**, 187–207 (2002)
27. Kloeden, P.E., Neukirch, A.: Convergence of numerical methods for stochastic differential equations in mathematical finance (2012). arXiv:1204.6620v1 [math.NA]
28. Kloeden, P.E., Platen, E.: Numerical Solution of Stochastic Differential Equations. Springer, Berlin (1999)
29. Kloeden, P.E., Platen, E., Wright, I.W.: The approximation of multiple stochastic integrals. Stoch. Anal. Appl. **10**, 431–441 (1992)
30. Komarova, N.L.: Mathematical modeling of cyclic treatment of chronic myeloid leukemia. Math. Biosci. Eng. **8**, 289–306 (2011)
31. Komarova, N.L., Katouli, A.A., Wodarz, D.: Combination of two but not three current targeted drugs can improve therapy of chronic myeloid leukemia. PLoS One **4**(2), e4423 (2009). URL http://view.ncbi.nlm.nih.gov/pubmed/19204794
32. Komarova, N.L., Wodarz, D.: Drug resistance in cancer: principles of emergence and prevention. Proc. Natl. Acad. Sci. USA **102**, 9714 (2005)
33. Komarova, N.L., Wodarz, D.: Effect of cellular quiescence on the success of targeted cml therapy. PLoS One **2**(10), e990 (2007). URL http://view.ncbi.nlm.nih.gov/pubmed/17912367
34. Komarova, N.L., Wodarz, D.: Stochastic modeling of cellular colonies with quiescence: an application to drug resistance in cancer. Theor. Popul. Biol. **72**(4), 523–538 (2007). URL http://view.ncbi.nlm.nih.gov/pubmed/17915274
35. Komarova, N.L., Wodarz, D.: Combination therapies against chronic myeloid leukemia: short-term versus long-term strategies. Cancer Res. **69**(11), 4904–4910 (2009). URL http://view.ncbi.nlm.nih.gov/pubmed/19458080
36. Li, W., Wolynes, P.G., Takada, S.: Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins. Proc. Natl. Acad. Sci. **108**(9), 3504–3509 (2011). DOI 10.1073/pnas.1018983108. URL http://www.pnas.org/content/108/9/3504.abstract
37. Lin, Y.T., Kim, H., Doering, C.R.: Features of fast living: on the weak selection for longevity in degenerate birth-death processes. J. Stat. Phys. **148**, 646–662 (2012)

38. Michor, F., Hughes, T.P., Iwasa, Y., Branford, S., Shah, N.P., Sawyers, C.L., Nowak, M.A.: Dynamics of chronic myeloid leukaemia. Nature **435**(7046), 1267–1270 (2005). URL http://view.ncbi.nlm.nih.gov/pubmed/15988530
39. Mjolsness, E.: On cooperative quasi-equilibrium models of transcriptional regulation. J. Bioinform. Comput. Biol. **5**(2b), 467–490 (2007)
40. Mjolsness, E.: Towards a calculus of biomolecular complexes at equilibrium. Brief. Bioinform. **8**(4), 226–233 (2007)
41. Paquin, D., Kim, P.S., Lee, P.P., Levy, D.: Strategic treatment interruptions during imatinib treatment of chronic myelogenous leukemia. Bull. Math. Biol. **73**, 1082 (2010). URL http://view.ncbi.nlm.nih.gov/pubmed/20532990
42. Roeder, I., Herberg, M., Horn, M.: An age-structured model of hematopoietic stem cell organization with application to chronic myeloid leukemia. Bull. Math. Biol. **71**(3), 602–626 (2009). URL http://view.ncbi.nlm.nih.gov/pubmed/19101772
43. Roeder, I., Horn, M., Glauche, I., Hochhaus, A., Mueller, M.C., Loeffler, M.: Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. Nat. Med. **12**(10), 1181–4 (2006). URL http://view.ncbi.nlm.nih.gov/pubmed/17013383
44. Strook, D.W., Varadhan, S.R.S.: Multidimensional Diffusion Processes. Springer, Berlin (2006)
45. Tomasetti, C., Levy, D.: Role of symmetric and asymmetric division of stem cells in developing drug resistance. Proc. Natl. Acad. Sci. USA **107**, 16766–16771 (2010)
46. Vallee-Belisle, A., Ricci, F., Plaxco, K.W.: Thermodynamic basis for the optimization of binding-induced biomolecular switches and structure-switching biosensors. Proc. Natl. Acad. Sci. **106**(33), 13802–13807 (2009). DOI 10.1073/pnas.0904005106. URL http://www.pnas.org/content/106/33/13802.abstract
47. Walczak, A.M., Tkačik, G.c.v., Bialek, W.: Optimizing information flow in small genetic networks. ii. feed-forward interactions. Phys. Rev. E **81**, 041905 (2010). DOI 10.1103/PhysRevE.81.041905. URL http://link.aps.org/doi/10.1103/PhysRevE.81.041905
48. Wang, G., Zaman, M.H.: Communications: Hamiltonian regulated cell signaling network. J. Chem. Phys. **132**(12), 121103 (2010) DOI 10.1063/1.3357980. URL http://link.aip.org/link/?JCP/132/121103/1
49. Wang, J., Huang, B., Xia, X., Sun, Z.: Funneled landscape leads to robustness of cell networks: yeast cell cycle. PLoS Comput. Biol. **2**(11), e147 (2006). DOI 10.1371/journal.pcbi.0020147. URL http://dx.plos.org/10.1371%2Fjournal.pcbi.0020147
50. Waxman, D.: A unified treatment of the probability of fixation when population size and the strength of selection change over time. Genetics **188**(4), 907–913 (2011). DOI 10.1534/genetics.111.129288. URL http://www.genetics.org/content/188/4/907.abstract
51. Wiktorsson, M.: Joint characteristic function and simultaneous simulation of iterated Ito integrals for multiple independent Brownian motions. Ann. Appl. Probab. **11**, 470–487 (2001)
52. Wodarz, D.: Stem cell regulation and the development of blast crisis in chronic myeloid leukemia: implications for the outcome of imatinib treatment and discontinuation. Med. Hypotheses **70**, 128 (2008)
53. Woywod, C., Gruber, F., Engh, R., Flå, T.: Dynamical models of mutated chronic myelogenous leukaemia cells for a post-imatinib treatment scenario: response to dasatinib or nilotinib therapy. PLoS Comput. Biol. (2013, submitted)

Part II
Hamiltonian Dynamics, Geometric Mechanics and Control Theory

Chapter 12
Singular Solutions of Euler–Poincaré Equations on Manifolds with Symmetry

D.D. Holm, J. Munn, and S.N. Stechmann

In honor of Jürgen Scheurle's 60th birthday

Abstract The Euler–Poincaré equation EPDiff governs geodesic flow on the diffeomorphisms with respect to a chosen metric, which is typically a Sobolev norm on the tangent space of vector fields. For a strong enough norm, EPDiff admits singular solutions, called "diffeons," whose momenta are supported on embedded subspaces of the ambient space. Diffeons are true solitons for some choices of the norm. The diffeon solution itself is a momentum map. Consequently, the diffeons evolve according to canonical Hamiltonian equations.

This paper examines diffeon solutions on Einstein spaces that are "mostly" symmetric, i.e., whose quotient by a subgroup of the isometry group is one-dimensional. An example is the two-sphere, whose isometry group $\mathrm{SO}3$ contains S^1. In this situation, the singular diffeons are supported on latitudes of the sphere. For this S^1 symmetry of the two-sphere, the canonical Hamiltonian dynamics for diffeons reduces from integral partial differential equations to a dynamical system of ordinary differential equations for their co-latitudes. Explicit examples are computed numerically for the motion and interaction of the Puckons on the sphere with respect to the H^1 norm. We analyze this case and several other two-dimensional examples.

D.D. Holm (✉)
Mathematics Department, Imperial College London, SW7 2AZ London, UK
e-mail: d.holm@ic.ac.uk

J. Munn
Eltham College, Grove Park Road, Mottingham, SE9 4QF London, UK
e-mail: jmm@eltham-college.org.uk

S.N. Stechmann
Mathematics Department, University of Wisconsin, Madison, WI 53706, USA
e-mail: stechmann@wisc.edu

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_12,

From consideration of these two-dimensional spaces, we outline the theory for reduction of diffeons on a general manifold possessing a metric equivalent to the warped product of the line with the bi-invariant metric of a Lie group.

12.1 Introduction

12.1.1 Motivation and Problem Statement

The Euler–Poincaré equation EPDiff arises in continuum mechanics, where for the L^2 norm it governs incompressible fluid flow [2]. EPDiff also arises in shape analysis, where it is used to measure the "distance" between two images with respect to a chosen metric [16, 17]. When the metric is a Sobolev norm, $(2, s)$ with $s \geq 1$, EPDiff admits singular solutions, called ***diffeons***, whose momenta are supported on embedded subspaces of the ambient space.

Finally, EPDiff also arises in integrable systems theory, as the dispersionless case of the Camassa–Holm (CH) equation for strongly nonlinear shallow water waves [4]. In this case, the metric is the H^1 norm of the fluid velocity. The momentum for each solitary wave of the CH equation in one spatial dimension is concentrated on a point moving with the flow, at which the velocity profile has a jump in derivative at its peak. These singular soliton solutions are called ***peakons***.

Previously, EPDiff has been solved analytically for the interactions of two solitary waves on the real line in one dimension [4], as radially symmetric rotating concentric circles on the two-dimensional plane [10] and numerically in two dimensions and three dimensions, for a variety of initial value problems [12]. In cases with one-dimensional linear, or radial symmetry, these dynamics reduce to canonical Hamiltonian ordinary differential equations (ODEs). Here we consider the corresponding reduction for singular wave motion on the surface of the sphere for EPDiff with respect to the H^1 norm and we discuss its generalizations for other surfaces of constant curvature.

In each case, we derive the symmetry which reduces the dynamics of the singular solutions of the EPDiff equation to a system of canonical Hamiltonian ODEs. We also provide explicit numerical results for the interactions of these singular EPDiff solutions in the case of Puckon motion as concentric latitudes moving on the sphere. In showing a variety of examples of how EPDiff can be reduced to interesting systems of canonical Hamiltonian ODEs for multi-diffeon solutions, we hope to inspire ideas for further explorations and applications of these solutions. To facilitate this purpose, most of the calculations will be done explicitly.

12.1.2 The Camassa–Holm Equation on a Riemannian Manifold

Variational Formulation of 1D CH.

The dispersionless Camassa–Holm (CH) equation [4] for one-dimensional shallow water waves arises from stationarity of the kinetic energy functional given by the Sobolev (2,1)-norm

$$\ell \,:\, W^{2,1}(\mathbb{R}) \longrightarrow \mathbb{R}\,,$$
$$\ell[u] = \int_{\mathbb{R}} \left(u(x)^2 + u_x(x)^2\right) \mathrm{d}x\,,$$

subject to velocity variations in the Euler–Poincaré form [9]

$$\delta u = \dot{\zeta} - \mathrm{ad}_u\zeta = \dot{\zeta} + [u,\zeta]\,,$$

where $u, \zeta \in \mathfrak{g}$, and $\mathfrak{g}$ consists of the vector fields on the real line, $\mathfrak{X}(\mathbb{R})$, with Lie bracket $[\cdot,\cdot]$ given by

$$[u,\zeta] = u\zeta_x - \zeta u_x = -\mathrm{ad}_u\zeta\,.$$

The dispersionless CH equation itself is given by the following system for velocity u and its dual momentum m,

$$\begin{aligned} \dot{m} &= -\mathrm{ad}^*_u m = -(m\partial_x + \partial_x m)u\,,\\ \text{with}\quad m &= \frac{\delta \ell}{\delta u} = u - u_{xx} = u + \Delta u\,,\\ \text{so that}\quad u &= G * m = \int_{\mathbb{R}} G(x-y)m(y)dy\,, \end{aligned} \tag{12.1}$$

where $G(x) = \frac{1}{2}e^{-|x|}$ is the Green's function for the Helmholtz operator $1 - \partial_x^2$ on the real line. We are using the convention that the Laplacian in 1D is given by

$$\Delta u = -\,u_{xx}.$$

The minus sign comes from regarding the Laplacian as $\partial_x^*\partial_x$, where ∂_x^* is the L^2 adjoint of the derivative operator ∂_x.

Legendre Transforming 1D CH to the Hamiltonian Side.

The Legendre transform yields the following invertible relations between momentum and velocity,

$$m = (1 - \partial_x^2)u \quad \text{and} \quad u = G * m\,, \tag{12.2}$$

where $G(x) = \frac{1}{2}e^{-|x|}$ is the ***Green's function*** for the Helmholtz operator $(1 - \partial_x^2)$, assuming homogeneous boundary conditions (on u) that allow inversion of the Helmholtz operator to determine u from m.

The associated ***Hamiltonian*** is,

$$h[m] = \langle m\,,\, u\rangle - \frac{1}{2}\|u\|^2 = \frac{1}{2}\int m \cdot G * m \,\mathrm{d}x =: \frac{1}{2}\|m\|^2\,, \tag{12.3}$$

which also defines a norm $\|m\|$ via a convolution kernel G, which is symmetric and positive, when the Lagrangian $\ell[u]$ is a norm. As expected, the norm $\|m\|$ given by the Hamiltonian $h[m]$ specifies the velocity u in terms of its Legendre-dual momentum m by the variational operation,

$$u = \frac{\delta h}{\delta m} = G * m \equiv \int G(x-y)\, m(y)\,\mathrm{d}y\,. \tag{12.4}$$

The kernel $G(x-y) = \frac{1}{2}e^{-|x-y|}$ for the Helmholtz operator is translation-invariant, so Noether's theorem implies that the total momentum $M = \int m\,\mathrm{d}^n x$ is conserved. It is also symmetric under spatial reflections, so u and m have the same parity under spatial reflections.

After the Legendre transformation (12.3), Eq. (12.1) appears in its equivalent ***Lie–Poisson Hamiltonian form***,

$$\frac{\partial}{\partial t}m = \{m, h\} = -\,\mathrm{ad}^*_{\delta h/\delta m} m\,. \tag{12.5}$$

Here the operation $\{\cdot\,,\,\cdot\}$ denotes the Lie–Poisson bracket dual to the (right) action of vector fields among themselves by vector-field commutation. That is,

$$\{f\,,\,h\} = -\left\langle m,\, \left[\frac{\delta f}{\delta m},\, \frac{\delta h}{\delta m}\right]\right\rangle. \tag{12.6}$$

Variational Formulation of CH on a Riemannian Manifold.

The dispersionless CH equation (12.1) may be put onto a general Riemannian manifold $(M, \langle\cdot,\cdot\rangle)$ with $\dim(M) = n$ and Levi–Civita connection ∇, by introducing the functional on the space of weakly differentiable square-integrable vector fields,

$$\ell \,:\, W^{2,1}\mathscr{X}(M) \longrightarrow \mathbb{R}$$
$$\ell[u] = \int_M \left(|u|^2 + |\nabla u|^2\right)\mathrm{d}vol.$$

Again, we consider the stationary points of ℓ with respect to variations of the Euler–Poincaré form

$$\delta u = \dot{\zeta} + [\zeta, u]\,.$$

This requirement yields the following n-dimensional EPDiff equation on M,

$$\dot{m} = -\mathrm{ad}^*_u m\,, \quad \text{where} \quad m = \frac{\delta \ell}{\delta u} = u + \Delta^\nabla u\,, \tag{12.7}$$

with which we shall work in the remainder of the paper. Here we denote

$$\Delta^\nabla = \nabla^* \nabla$$

for the connection Laplacian with respect to the metric and we assume homogeneous boundary conditions for the velocity u.

The same type of Legendre transformation and Hamiltonian formulation as before also applies to the EPDiff equation (12.7) for the CH equation on a Riemannian manifold. For more details and additional background concerning the relation of classical EP theory to Lie–Poisson Hamiltonian equations, see [9, 15].

Diffeons: Singular Solutions for EPDiff.

The Hamiltonian formulation of the EPDiff equation (12.7) is Lie–Poisson in the momentum variable m, which possesses a remarkable set of singular solutions, given by

$$m(t,x) = \sum_{i=1}^{N} \int_{S_i} P_i(t,s_i)\delta(x - R_i(t,s_i))\, ds_i \quad \text{for} \quad x \in M. \tag{12.8}$$

These singular solutions emerge dynamically from smooth confined initial conditions and are called ***diffeons***. The diffeon singular solutions of EPDiff are vector-valued functions supported in M on a set of N surfaces (or curves) of codimension $(n-k)$ for $s \in M^k \subset M$ with $\dim M^k = k < n$. For example, diffeons may be supported on sets of points ($k = 0$), curves ($k = 1$), or surfaces ($k = 2$) in three dimensions. These support sets of m move with the fluid velocity $u = G * m$; so the coordinates $s \in M^k \subset M$ are Lagrangian fluid labels. The Green's function G for the operator $1 + \Delta^\nabla$ is continuous, but it has a jump in its derivative on the support set that advects with the velocity $u = G * m$ under the evolution of EPDiff. In fluid dynamics, a jump in derivative which moves with the flow is called a "contact discontinuity" [14]. The relationship of the singular solutions of EPDiff equation (12.7) to fluid dynamics is illuminated by rewriting equation (12.7) in "Riemann-invariant form,"

$$\frac{d}{dt}(m \cdot dx \otimes dvol) = 0, \quad \text{along} \quad \frac{dx}{dt} = u = G * m.$$

The diffeon singular solution ansatz (12.8) was first discovered as the "peakon" solutions for CH motion on the real line in [4]. This was generalized to motion in higher dimensions in [11] and was shown to be a momentum map in [8]. As a result of the singular solution ansatz being a momentum map, the $2N$ variables P_i, R_i,satisfy Hamilton's canonical equations. In general, these are integro-partial differential equations for the canonically conjugate diffeon parameters $P_i(t, s_i), R_i(t, s_i)$ in (12.8) with $i = 1, \ldots, N$.

12.1.3 Main Results of the Paper

This paper will study diffeon solutions of the EPDiff equations (12.7) on Einstein manifolds that are "mostly" symmetric, i.e., that have a group acting by isometries so that the orbits have co-dimension 1 in the manifold except on a set of measure zero, and hence the quotient of the manifold by the group is one-dimensional.

Thus, we are interested in identifying and analyzing cases where imposing an additional translation symmetry on the solution reduces the canonical Hamiltonian dynamics of the singular solutions of the EPDiff on Einstein manifolds from integro-partial differential equations to Hamiltonian ordinary differential equations for canonical variables $P_i(t), R_i(t)$. We shall then analyze some of the properties of those canonical equations and examine their numerical solutions.

We begin by noting some simplifications of the EPDiff equations when restricted to Einstein spaces. These simplifications take advantage of the relation between the connection Laplacian and the Hodge Laplacian, the latter of which is easier to compute. Knowing the EPDiff equations in one dimension, we move up to the Einstein spaces in two dimensions.

The first manifold we study is the 2-sphere, which has the familiar group action formed by rotations about a fixed axis. Thus we construct "Puckons," which are singular solutions of the EPDiff equations supported on concentric circular latitudes of the 2-sphere. The name "Puckon" (as opposed to "peakon") arises from the famous boast by Shakespeare's character, Puck, in *A Midsummer Night's Dream*, that he would, "put a girdle round about the earth in forty minutes."[1] The case of rotationally symmetric singular solutions of EPDiff in the Euclidean plane has already been discussed in [10], so we proceed to examine the hyperbolic plane, which has a rich isometry group. However, the diffeon dynamics on the unbounded hyperbolic plane affords little opportunity for multiple diffeon interactions and does not admit periodic behavior. Even richer opportunities for diffeon dynamics and

[1]Thanks to J.D. Gibbon for reminding us of this quote and suggesting the name, "Puckon" for these solutions.

interactions in hyperbolic spaces are offered in the context of Teichmüller theory. However, the latter is beyond the scope of the present work, and we refer to [13].

From consideration of these two-dimensional spaces, we sketch how one might develop the theory for a general manifold that possesses a metric equivalent to the warped product of the line with the bi-invariant metric of a Lie group.

12.1.4 Plan of the Paper

After briefly recalling the essentials needed from the theory of Einstein manifolds in Sect. 12.2, we begin in Sect. 12.3 by reducing the singular solutions of EPDiff to a canonical Hamiltonian dynamical system for the simplest Einstein manifold, the 2-sphere. We study the motion on a sphere of singular EPDiff solutions supported on concentric circular latitudes, or "girdles." These new singular solutions girdling the sphere are the "Puckons." Although the canonical reduction is *guaranteed* by the momentum map property of EPDiff, the reduction of its singular diffeon solutions for Puckons is given in detail, so it may be used to confirm the numerical results for the interactions of Puckons on the sphere described in Sect. 12.4. Section 12.5 generalizes the Puckons to other surfaces that are Einstein manifolds with a translation symmetry. Section 12.6 incorporates these ideas into the theory of warped products.

12.2 EPDiff Equations on Einstein Spaces

The operator

$$\Delta^{\nabla} + \mathbb{1}$$

arises in the following context in Riemannian geometry.

Definition 12.1. On any given Riemannian manifold, the musical isomorphisms are defined to be the Riesz representation and inverse maps with respect to the metric:

$$\begin{aligned} \sharp : \mathrm{T}^*M &\longrightarrow \mathrm{T}M \\ \langle \alpha^\sharp, w \rangle &= \alpha(w) \end{aligned}$$

$$\begin{aligned} \flat : \mathrm{T}M &\longrightarrow \mathrm{T}^*M \\ v_\flat(w) &= \langle v, w \rangle \,. \end{aligned}$$

In Riemannian geometry, the Levi–Civita connection respects the musical isomorphisms, namely for $v, w \in \mathscr{X}(M)$ and $\alpha \in \Omega^1(M)$:

$$(\nabla_v w)_\flat = \nabla_v (w_\flat)$$
$$(\nabla_v \alpha)^\sharp = \nabla_v (\alpha^\sharp).$$

Thus, the musical isomorphisms identify vector fields with 1-forms and allow them to be differentiated in the same way.

The connection Laplacian on 1-forms satisfies the Bochner–Weitzenböck formula.

Theorem 12.1 (Bochner–Weitzenböck). *On a Riemannian manifold with Levi–Civita connection ∇, any 1-form α satisfies*

$$\Delta^{\mathrm{d}} \alpha = \Delta^{\nabla} \alpha + \mathrm{Ric}(\alpha)$$

where $\Delta^{\mathrm{d}} = \mathrm{d}^*\mathrm{d} + \mathrm{d}\mathrm{d}^*$ is the Hodge Laplacian formed from the exterior derivative d on forms and Ric is the Ricci curvature operator.

From now on, we restrict our attention to a special type of manifold, namely the Einstein manifolds.

Definition 12.2. An Einstein manifold is a Riemannian manifold that satisfies

$$\mathrm{Ric} = k\, \mathbb{1}$$

for some constant k.

On an Einstein manifold, one may scale the metric so that k can be replaced by $-1, 0, 1$ depending on whether k is negative, zero, or positive, respectively. Thus on an Einstein manifold, the Bochner–Weitzenböck formula becomes

$$\Delta^{\mathrm{d}} = \Delta^{\nabla} + \mathrm{sign}(k)\, \mathbb{1}.$$

We restrict our study further to Einstein manifolds with positive k (we shall call these positive Einstein manifolds) and scale to $k = 1$. Thus we have

$$\Delta^{\mathrm{d}} = \Delta^{\nabla} + \mathbb{1}.$$

The implication of this for the EPDiff Lagrangian on vector fields is as follows, for homogeneous boundary conditions:

$$\ell[\boldsymbol{u}] = \int_M \left(|\boldsymbol{u}|^2 + |\nabla \boldsymbol{u}|^2 \right) \mathrm{d}vol$$
$$= \int_M \left(|\boldsymbol{u}_\flat|^2 + |\nabla \boldsymbol{u}_\flat|^2 \right) \mathrm{d}vol$$

$$
\begin{aligned}
&= \int_M \left\langle \boldsymbol{u}_\flat, \boldsymbol{u}_\flat + \Delta^\nabla \boldsymbol{u}_\flat \right\rangle \mathrm{d}vol \\
&= \int_M \left\langle \boldsymbol{u}_\flat, \Delta^\mathrm{d} \boldsymbol{u}_\flat \right\rangle \mathrm{d}vol \\
\Big(&= \int_M \left(|\mathrm{d}\boldsymbol{u}_\flat|^2 + |\mathrm{d}^* \boldsymbol{u}_\flat|^2 \right) \mathrm{d}vol \Big).
\end{aligned}
$$

Hence, finding the stationary points of Lagrangian ℓ with the usual Euler–Poincaré constraints on the variations implies the EPDiff equation system of Eq. (12.7) in the form

$$
\dot{\boldsymbol{m}} = -\mathrm{ad}^*_{\boldsymbol{u}} \boldsymbol{m} \,, \quad \text{where} \quad \boldsymbol{m} = \Delta^\mathrm{d} \boldsymbol{u}_\flat \,. \tag{12.9}
$$

These equations generalize the EPDiff equations to Einstein manifolds.

12.3 The EPDiff Equation on the Sphere

In two dimensions, the only manifold with positive Einstein constant is the standard round sphere which we shall regard as the Riemann sphere. We use stereographic projections to identify the complex plane with the sphere whose North Pole is removed. This is equivalent to putting the metric

$$
g = \frac{4}{(1 + x^2 + y^2)^2} \left(\mathrm{d}x^2 + \mathrm{d}y^2 \right) = \frac{4}{(1 + |z|^2)^2} \mathrm{d}z \mathrm{d}\bar{z}
$$

on the plane.

12.3.1 Rotationally Invariant Solutions

We shall examine vector-field solutions $\boldsymbol{u}$ of EPDiff in (12.9) with

$$
\int_{\mathrm{S}^2} \left\langle \Delta^\mathrm{d} \boldsymbol{u}_\flat, \beta \right\rangle = \int_C \left\langle p \mathrm{d}r + q \mathrm{d}\theta, \beta \right\rangle
$$

for any smooth 1-form β and constant p, q, and where C is a circle of latitude on the sphere. The quantity $\boldsymbol{m} = \Delta^\mathrm{d} \boldsymbol{u}_\flat$ is a distribution, defined by its integration against a smooth function. Thus, we seek weak, or singular, solutions of EPDiff on the sphere. We change coordinates on the sphere to assist us in this search. The metric on $\mathbb{R}^2 \backslash 0$ in polar coordinates is $\mathrm{d}r^2 + r^2 \mathrm{d}\theta$. Thus we can regard the sphere minus both poles as $\mathbb{R}^2 \backslash 0$ with the metric

$$g_{\mathbb{S}} = \rho^2 \left(\mathrm{d}r^2 + r^2 \mathrm{d}\theta^2 \right) , \quad \text{where} \quad \rho = \frac{2}{(1+r^2)} .$$

Green's Function for the Helmholtz Operator on the Riemann Sphere.

We now seek solutions to the equation

$$\Delta^{\mathrm{d}} \boldsymbol{u}_\flat = \delta(r-R)(k_1 \mathrm{d}r + k_2 \mathrm{d}\theta) \tag{12.10}$$

where R, k_1, k_2 are constants. Let us assume that the velocity one-form $\boldsymbol{u}_\flat$ is invariant under rotations of the sphere about the axis joining North and South Poles and is radial, i.e., we must solve

$$\boldsymbol{u}_\flat = a(r)\mathrm{d}r, \tag{12.11}$$

$$\Delta^{\mathrm{d}} \boldsymbol{u}_\flat = P \frac{\delta(r-R)}{R\rho(R)^2} \mathrm{d}r. \tag{12.12}$$

Let us first solve the Green's function equation

$$\Delta^{\mathrm{d}} (G(r,R)\mathrm{d}r) = \frac{\delta(r-R)}{R} \mathrm{d}r ,$$

for one recognizes that the single diffeon velocity is proportional to the Green's function, i.e.,

$$a(r) = PG(r,R) .$$

To solve explicitly for the Green's function in our present case, we begin by recalling

$$\Delta^{\mathrm{d}} G(r,R)\mathrm{d}r = -\frac{\partial}{\partial r} \left(\frac{(1+r^2)^2}{4r} \frac{\partial (rG(r,R))}{\partial r} \right) \mathrm{d}r .$$

Integrating the Green's function equation using this expression yields

$$G(r,R)\mathrm{d}r = \frac{1}{R} \left(\frac{A_1}{r} + \frac{2A_2}{r(1+r^2)} + \frac{2}{r(1+\max(r,R)^2)} \right) \mathrm{d}r ,$$

for constants A_1, A_2. In particular, we can remove the singularities at $r = 0$ and $r = \infty$ by setting $A_1 = 0$ and

$$A_2 = -\frac{1}{R(1+R^2)} ,$$

whence

$$G(r,R)\mathrm{d}r = \frac{2\min(r,R)^2}{rR(1+r^2)(1+R^2)}\mathrm{d}r.$$

is continuous over the sphere, but has a jump in derivative at $r = R$.

Thus, we have proved the following.

Proposition 12.1 (Radial Green's Function on the Sphere). *The solution to (12.11, 12.12) for the radial Green's function on the sphere is*

$$\boldsymbol{u}_\flat = \frac{PG(r,R)}{\rho(R)^2}\mathrm{d}r = \frac{PR(1+R^2)\min(r,R)^2}{2r(1+r^2)}\mathrm{d}r\,.$$

Solution Ansatz for EPDiff on the Sphere.

Following [11], we propose a solution ansatz for EPDiff velocity on the sphere as the following superposition of Green's functions,

$$\eta = \boldsymbol{u}_\flat = \sum_{i=1}^{N} \frac{P_i G_{R_i}}{R_i \rho(R_i)^2}\mathrm{d}r \tag{12.13}$$

where we denote

$$G_R(r) = RG(r,R).$$

and P_i, R_i are $2N$ functions of time. The corresponding vector field dual to η is

$$\boldsymbol{u} = \sum_{i=1}^{N} \frac{P_i G_{R_i}}{R_i \rho(R_i)^2 \rho(r)^2}\partial_r \tag{12.14}$$

We pair the arbitrary smooth vector field $w = f\partial_r + g\partial_\theta$ on S^2 with the EPDiff equation using the solution ansatz given by the one-form η in Eq. (12.13).

A direct calculation yields the following equations for the diffeon parameters,

Proposition 12.2 (Diffeon Parameter Equations on the Sphere). *The equations for diffeon parameter evolution on the sphere are:*

$$\dot{R}_i = \sum_{j=1}^{N}\left(\frac{P_j}{\rho(R_i)^2\rho(R_j)^2 R_j}G(R_i,R_j)\right) \tag{12.15}$$

$$\dot{P}_i = -\sum_{j=1}^{N}\frac{P_iP_j}{\rho(R_i)^2\rho(R_j)^2}\left(\frac{\partial}{\partial r}G(r,R_j)\Big|_{r=R_i} + 2R_i\rho(R_i)G(R_i,R_j)\right). \tag{12.16}$$

Proof. By direct calculation,

$$
\begin{aligned}
0 &= \frac{\partial}{\partial t}\left(\int_{\mathrm{S}^{2i}} \Delta^{\mathrm{d}}\eta(w)\mathrm{d}vol\right) + \int_{\mathrm{S}^2} \Delta^{\mathrm{d}}\eta([w,\boldsymbol{u}])\mathrm{d}vol \\
&= \frac{\partial}{\partial t}\left(\sum_{i=1}^{N}\int_0^{2\pi}\int_0^{\infty} \delta(r-R_i)f(r,\theta)P_i\rho^2 r\mathrm{d}r\mathrm{d}\theta\right) \\
&+ \int_0^{2\pi}\sum_{i=1}^{N}\int_0^{\infty}\delta(r-R_i)P_i\mathrm{d}r\left(\left[f(r,\theta)\partial_r + g(r,\theta)\partial_\theta, \sum_{j=1}^{N}\rho^{-2}P_jG_{R_j}(r)\partial_r\right]\right)\rho^2 r\mathrm{d}r\mathrm{d}\theta \\
&= \frac{\partial}{\partial t}\left(\sum_{i=1}^{N} P_i\int_0^{2\pi} f(R_i,\theta)\mathrm{d}\theta\right) \\
&+ \sum_{i,j=1}^{N} R_i\int_0^{2\pi}\Big(P_jG'_{R_j}(R_i)f(R_i,\theta) + 2P_jR_i\rho(R_i)G_{R_j}(R_i)f(R_i,\theta) \\
&\qquad\qquad -P_jG_{R_j}(R_i)f_r(R_i,\theta)\Big)P_i\mathrm{d}\theta \\
&= \sum_{i=1}^{N}\int_0^{2\pi}(f_r(R_i,\theta)\dot{R}_iP_i + \dot{P}_if(R_i,\theta))\mathrm{d}\theta \\
&+ \sum_{i,j=1}^{N} R_i\int_0^{2\pi}\Big(P_jG'_{R_j}(R_i)f(R_i,\theta) + 2P_jR_i\rho(R_i)G_{R_j}(R_i)f(R_i,\theta) \\
&\qquad\qquad -P_jG_{R_j}(R_i)f_r(R_i,\theta)\Big)P_i\mathrm{d}\theta
\end{aligned}
$$

Comparing coefficients of f and f_r implies

$$
\begin{aligned}
0 &= \dot{R}_iP_i - \sum_{j=1}^{N}\left(\frac{P_iP_j}{\rho(R_i)^2\rho(R_j)^2}G(R_i,R_j)\right) \\
0 &= \dot{P}_i + \sum_{j=1}^{N}\frac{P_iP_j}{\rho(R_i)^2\rho(R_j)^2R_j}\left(G'_{R_j}(R_i) + 2R_i\rho(R_i)G_{R_j}(R_i)\right)
\end{aligned}
$$

This finishes the calculation of the diffeon parameter evolution equations (12.15, 12.16). □

Proposition 12.3 (Canonical Hamiltonian Form of Diffeon Parameter Equations). *Evolution equations (12.15, 12.16) for the diffeon parameters are equivalent to Hamilton's canonical equations,*

$$\pi\dot{R}_i = \frac{\partial H}{\partial R_i}\,, \quad \pi\dot{P}_i = -\frac{\partial H}{\partial P_i}\,, \tag{12.17}$$

with Hamiltonian function of $\mathbf{P} = (P_1, \ldots, P_N)$ *and* $\mathbf{R} = (R_1, \ldots, P_N)$ *given by,*

$$H(\mathbf{P}, \mathbf{R}) = \pi \sum_{i,j=1}^{N} \frac{P_i P_j G(R_j, R_i)}{\rho(R_i)^2 \rho(R_j)^2}\,. \tag{12.18}$$

Proof. We Legendre transform the Lagrangian ℓ into the Hamiltonian H by setting

$$H(\boldsymbol{m}) = \int_{\mathrm{S}^2} \boldsymbol{m}(\boldsymbol{u})\mathrm{d}vol - \ell[\boldsymbol{u}]$$

whence we find that

$$H(\boldsymbol{m}) = \frac{1}{2}\int_{\mathrm{S}^2} \boldsymbol{m}(\boldsymbol{u})\mathrm{d}vol$$

with

$$\boldsymbol{m} = \Delta^{\mathrm{d}}\boldsymbol{u}_\flat.$$

We then express the velocity as the superposition of Green's functions,

$$\boldsymbol{u}_\flat = \eta = \sum_{i=1}^{N} \frac{1}{\rho(R_i)^2} \frac{2P_i \min(r, R_i)^2}{rR_i(1+r^2)(1+R_i^2)}\mathrm{d}r\,.$$

We set $k_i = P_i/(\rho(R_i)^2 R_i)$, for $i = 1, \ldots, N$, and we evaluate the Hamiltonian on this solution with $\boldsymbol{m} = \Delta^{\mathrm{d}}\eta$ as

$$\begin{aligned}
H(\boldsymbol{m}) &= \frac{1}{2}\int_{\mathrm{S}^2} \langle \eta, \boldsymbol{m}\rangle \,\mathrm{d}vol \\
&= \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} \sum_{i,j=1}^{N} \left\langle \frac{2k_i \min(r, R_i)^2}{r(1+r^2)(1+R_i^2)}\mathrm{d}r, \Delta^{\mathrm{d}} \frac{2k_j \min(r, R_j)^2}{r(1+r^2)(1+R_j^2)}\mathrm{d}r \right\rangle \rho^2 r\mathrm{d}r \wedge \mathrm{d}\theta \\
&= \frac{1}{2}\int_0^{2\pi}\int_0^{\infty} \sum_{i,j=1}^{N} \left\langle k_i G_{R_i}(r)\mathrm{d}r, k_j \delta(r - R_j)\mathrm{d}r \right\rangle \rho^2 r\mathrm{d}r \wedge \mathrm{d}\theta \\
&= \frac{1}{2}\int_0^{2\pi} \sum_{i,j=1}^{N} k_i G_{R_i}(R_j) k_j R_j \mathrm{d}\theta \\
&= \pi \sum_{i,j=1}^{N} \frac{P_i P_j G(R_i, R_j)}{\rho(R_i)^2 \rho(R_j)^2}.
\end{aligned}$$

The canonical equations for this Hamiltonian now recover the diffeon parameter evolution equations (12.15, 12.16). □

Thus, we see that we have

$$H(\boldsymbol{m}) = 2\pi H(\mathbf{P}, \mathbf{R})$$

Remark 12.1. As explained in [8], the reduction of EPDiff to canonical Hamiltonian form for the diffeons was guaranteed, because the singular solution ansatz (12.8) is a momentum map. However, we shall require the explicit results here for the numerical solutions in Sect. 12.4.

Definition 12.3 (N-Puckon). The singular solution of EPDiff

$$\frac{\partial}{\partial t}\Delta^{\mathrm{d}}\eta = -\mathrm{ad}^*_{\eta^\sharp}\Delta^{\mathrm{d}}\eta\,, \tag{12.19}$$

is given on the Riemann sphere by a vector field η satisfying

$$\Delta^{\mathrm{d}}\eta = \sum_{i=1}^{N} \frac{P_i}{\rho(R_i)^2 R_i}\delta(r - R_i)\mathrm{d}r\,. \tag{12.20}$$

The support set of this (weak) solution of EPDiff on the Riemann sphere is a set of circular latitudes (girdles) at radii $R_i(t)$ with conjugate radial momenta $P_i(t)$, where $i = 1, \ldots, N$. Equation (12.20) constitutes a solution ansatz for the velocity vector field η, which will be called an N-Puckon solution.

12.3.2 The Basic Irrotational Puckon

The Single Irrotational Puckon.

Let us consider the case where $N = 1$ and examine the motion of the basic Puckon without rotation. For $N = 1$, the Hamiltonian (12.18) is given by

$$H(P, R) = \frac{1}{2}\frac{P^2 G(R, R)}{\rho(R)^4} = \frac{P^2}{8}(1 + R^2)^2\,,$$

which follows because the Green's function in this case is

$$G(R, R) = 2\frac{\min(R, R)^2}{R^2(1 + R^2)^2} = \frac{2}{(1 + R^2)^2}\,.$$

Thus, upon restricting ourselves to the constant level set of the Hamiltonian defined by

$$H(P,R) = \frac{K^2}{2}\,,$$

for constant K, we may solve for the momentum variable,

$$P = K\rho(R) = \frac{2K}{(1+R^2)}\,.$$

Next, we note that the canonical coordinate equation

$$\dot{R} = \frac{\partial H}{\partial P} = \frac{P}{4}(1+R^2)^2 = \frac{K}{2}(1+R^2)\,,$$

integrates to

$$R = \tan\left(\frac{K}{2}t + \frac{\varepsilon}{2}\right)\,. \tag{12.21}$$

Consequently, the single Puckon momentum is found as

$$P = K + K\cos\left(Kt+\varepsilon\right)\,. \tag{12.22}$$

So far we have been using the stereographic projection to provide charts for the sphere. However we may easily pass from the coordinates (r,θ) obtained from stereographic projection to latitudinal–longitudinal coordinates (ϕ,θ) where r and ϕ are related by

$$r = \tan\left(\frac{\phi}{2}\right)\,.$$

By this token, the co-latitude $\phi = \Phi(t)$ of the peak of the Puckon evolves linearly in time as

$$\Phi(t) = Kt + \varepsilon.$$

Proposition 12.4 (Irrotational Puckon Solution). *The velocity vector field $v = \eta^\sharp$ generated by the irrotational Puckon motion is given in stereographic coordinates by*

$$\boldsymbol{u} = \frac{P}{4}\frac{\min(r,R)}{\max(r,R)}(1+r^2)\partial_r\,.$$

This appears in latitudinal–longitudinal coordinates on the Riemann sphere, as follows:

$$\boldsymbol{u} = \frac{K(1+\cos(Kt+\varepsilon))}{2} \frac{\min\left(\left|\tan\left(\frac{\phi}{2}\right)\right|, \left|\tan\left(\frac{K}{2}t+\frac{\varepsilon}{2}\right)\right|\right)}{\max\left(\left|\tan\left(\frac{\phi}{2}\right)\right|, \left|\tan\left(\frac{K}{2}t+\frac{\varepsilon}{2}\right)\right|\right)} \frac{\partial}{\partial\phi}.$$

Proof. This result is obtained by direct substitution of solutions (12.21) and (12.22) for the diffeon parameters into the solution ansatz (12.14). □

Remark 12.2 (Puckon Peak). The peak of the Puckon occurs when

$$\phi = \Phi(t) = Kt + \varepsilon,$$

(modulo behavior at the poles), whence

$$|\boldsymbol{u}|_{S^2} = \frac{|K|(1+\cos(Kt+\varepsilon))}{2}.$$

We still need to examine precisely what happens to the dynamics of a Puckon as it collides with itself at the poles.

Proposition 12.5 (Puckon Behavior at the Poles). *The Puckon bounces elastically, when it collides with itself at the poles.*

Proof. We notice that $\dot{R} = K/2 \neq 0$ at $R = 0$ and by setting $U = 1/R$ we see that at $U = 0$, $\dot{U} = K/2 \neq 0$. Similarly, one can show that $\dot{P} = 0$ at both poles. Thus, the Puckon velocity is *finite* at the poles. Since Hamiltonian motion is time-reversible, the Puckon must bounce elastically, when it collides with itself at the poles. □

12.3.3 Rotating Puckons

So far we have concentrated on singular EPDiff solutions on the sphere moving with only a radial component. Now we turn to examine the rotating N-Puckon, i.e., the solution of the following system of equations for η, cf. Eq. (12.10),

$$\Delta^{\mathrm{d}}\eta = \sum_{i=1}^{N} \frac{\delta(r-R_i)}{\rho(R_i)^2 R_i} \left(P_i \mathrm{d}r + M_i \mathrm{d}\theta\right), \tag{12.23}$$

$$\frac{\partial}{\partial t}\Delta^{\mathrm{d}}\eta = -\mathrm{ad}^*_{\eta^\sharp}\Delta^{\mathrm{d}}\eta, \tag{12.24}$$

in which $M_i(t)$ with $i = 1, \ldots, N$, are the **angular momenta** of the N Puckons.

From the definition of G that

$$\Delta^{\mathrm{d}}\left(AG(r,R)\mathrm{d}r + BrG(r,R)\mathrm{d}\theta\right) = \delta(r-R)\left(\frac{A}{R}\mathrm{d}r + B\mathrm{d}\theta\right). \tag{12.25}$$

Proposition 12.6 (Canonical Hamiltonian Equations for the Rotating N-Puckon).
The equations for the rotating N-Puckon may be expressed as:

$$\dot{R}_i = \sum_{j=1}^{N} P_j F_j(R_i) R_j R_i = \frac{\partial H}{\partial P_i}\,, \tag{12.26}$$

$$\dot{P}_i = -\sum_{j=1}^{N}\left(P_i P_j F_j'(R_i) R_j R_i + P_i P_j F_j(R_i) R_j + M_i M_j F_j'(R_i)\right) = -\frac{\partial H}{\partial R_i}\,, \tag{12.27}$$

$$\dot{M}_i = 0\,, \tag{12.28}$$

where $F_i(R_j) = F_j(R_i)$ is defined in terms of the Green's function G as

$$F_i(R_j) = \frac{G(R_i,R_j)}{\rho(R_i)^2 R_i \rho^2(R_j) R_j}\,.$$

The last equation (12.28) expresses the conservation of the angular momentum of each Puckon. With M_i constant, the first two equations (12.26) and (12.27) comprise Hamilton's canonical equations with symmetry-reduced Hamiltonian

$$H(\mathbf{P},\mathbf{R};\mathbf{M}) = \frac{1}{2}\sum_{i,j=1}^{N}\left(P_i P_j R_i R_j + M_i M_j\right) F_i(R_j)\,. \tag{12.29}$$

Proof. First, we note from (12.25) that the velocity one-form corresponding to the momentum singular solution ansatz (12.23) is given by,

$$\eta = \sum_{i=1}^{N} \frac{G(r,R_i)}{\rho(R_i)^2 R_i}\left(R_i P_i \mathrm{d}r + M_i r \mathrm{d}\theta\right)$$

Thus the velocity vector field $\boldsymbol{u}$ dual to η is given by

$$\boldsymbol{u} = \sum_{i=1}^{N} \frac{G(r,R_i)}{\rho(R_i)^2 R_i \rho^2 r}\left(R_i P_i r \partial_r + M_i \partial_\theta\right).$$

To make our calculations more transparent, we collect terms and define notation as

$$F_i(r) = \frac{G(r,R_i)}{\rho(R_i)^2 R_i \rho^2 r},$$

so that

$$\boldsymbol{u} = \sum_{i=1}^{N} F_i(r)\left(R_i P_i r\partial_r + M_i\partial_\theta\right)$$

and

$$\eta = \sum_{i=1}^{N} F_i(r)\rho^2 r\left(R_i P_i \mathrm{d}r + M_i r\mathrm{d}\theta\right).$$

Note that the expression for F_i is symmetric in r and R_i, which implies the permutation symmetry,

$$F_i(R_j) = F_j(R_i).$$

Now let $w = f\partial_r + g\partial_\theta$ be any smooth vector field on S^2. Then, one computes the pairing

$$\int_{\mathrm{S}^2} \Delta^{\mathrm{d}}\eta(w)\mathrm{d}vol = \sum_{i=1}^{N}\int_0^{2\pi}\left(P_i f(R_i,\theta) + M_i g(R_i,\theta)\right)\mathrm{d}\theta.$$

Hence, the left-hand side of the EPDiff equation becomes

$$\begin{aligned}\tfrac{\partial}{\partial t}\left(\textstyle\int_{\mathrm{S}^2}\Delta^{\mathrm{d}}\eta(w)\mathrm{d}vol\right) = \textstyle\sum_{i=1}^{N}\int_0^{2\pi}\Big(&\dot{P}_i f(R_i,\theta) + P_i\dot{R}_i f_r(R_i,\theta) + \dot{M}_i g(R_i,\theta)\\ &+ M_i\dot{R}_i g_r(R_i,\theta)\Big)\mathrm{d}\theta.\end{aligned}$$

Now we need to calculate its right-hand side.

$$\int_{\mathrm{S}^2}\Delta^{\mathrm{d}}\eta([w,\boldsymbol{u}])\mathrm{d}vol.$$

For this, we write the commutator,

$$\begin{aligned}[w,\boldsymbol{u}] = &\sum_{i=1}^{N}\left(F_i' R_i P_i r f + F_i R_i P_i f - F_i R_i P_i r f_r - F_i M_i f_\theta\right)\partial_r\\ &+\sum_{i=1}^{N}\left(f F_i' M_i - F_i R_i P_i r g_r - F_i M_i g_\theta\right)\partial_\theta.\end{aligned}$$

Now, by Stokes' theorem, when we integrate over the whole sphere, the contribution due to f_θ, g_θ will be zero because f, g, the only components of the integration to depend on θ, satisfy $f(r,0) = f(r,2\pi)$ etc. Thus

$$
\begin{aligned}
&\int_{\mathrm{S}^2} \Delta^{\mathrm{d}}\eta([w,\boldsymbol{u}])\mathrm{d}vol \\
&= \sum_{i,j=1}^{N} \int_0^{2\pi} \big(P_iF_j'(R_i)R_jP_jR_if(R_i,\theta) + P_iF_j(R_i)R_jP_jf(R_i,\theta) \\
&\qquad\qquad -P_iF_j(R_i)R_jP_jR_if_r(R_i,\theta)\big)\,\mathrm{d}\theta \\
&\quad + \sum_{i,j=1}^{N} \int_0^{2\pi} \big(M_iF_j'(R_i)M_jf(R_i,\theta) - M_iF_j(R_i)R_jP_jR_ig_r(R_i,\theta)\big)\,\mathrm{d}\theta.
\end{aligned}
$$

Now to find the evolution equations for P_i, R_i, M_i we need to only compare the coefficients of f, f_r, g, g_r occurring in

$$
0 = \frac{\partial}{\partial t}\left(\int_{\mathrm{S}^2} \Delta^{\mathrm{d}}\eta(w)\mathrm{d}vol\right) + \int_{\mathrm{S}^2} \Delta^{\mathrm{d}}\eta([w,\boldsymbol{u}])\mathrm{d}vol\,.
$$

From this comparison, one finds

$$
\begin{aligned}
0 &= \dot{P}_i + \sum_{j=1}^{N}\Big(P_iF_j'(R_i)R_jP_jR_i + P_iF_j(R_i)R_jP_j + M_iF_j'(R_i)M_j\Big)\,, \\
0 &= P_i\dot{R}_i - \sum_{j=1}^{n}\Big(P_iF_j(R_i)R_jP_jR_i\Big)\,, \\
0 &= \dot{M}_i\,, \\
0 &= M_i\dot{R}_i - \sum_{j=1}^{N}\Big(M_iF_j(R_i)R_jP_jR_i\Big)\,,
\end{aligned}
$$

in which the second and fourth equations provide the *same* information. We therefore obtain the evolution equations (12.26–12.28) for P_i, R_i, M_i given in the statement of Proposition 12.6. □

Remark 12.3 (Verifying the Hamiltonian). A simple check following the same pattern as Proposition 12.3 shows that if the Hamiltonian

$$
H(\boldsymbol{m}) = \int_{\mathrm{S}^2} \boldsymbol{m}(\boldsymbol{u})\mathrm{d}vol - \ell[\boldsymbol{u}]
$$

is the Legendre transform of the Lagrangian ℓ then

$$
\begin{aligned}
H(\boldsymbol{m}) &= \frac{1}{2}\int_{\mathrm{S}^2} \big\langle \Delta^{\mathrm{d}}\eta, \eta\big\rangle\,\mathrm{d}vol \\
&= 2\pi H(\mathbf{P},\mathbf{R};\mathbf{M})\,,
\end{aligned}
$$

which is the Hamiltonian for the rotating N-Puckon in formula (12.29) of Proposition 12.6. As explained in [8], the reduction to canonical Hamiltonian form is guaranteed, since the singular solution ansatz (12.8) is a momentum map.

12.3.4 The Basic Rotating Puckon

Proposition 12.7 (Extremal Radii of the Basic Rotating Puckon). *The motion of the basic rotating Puckon lies on the Riemann plane between the maximum and minimum values of R given by,*

$$R_{max/min} = \frac{2K \pm \sqrt{4K^2 - M^2}}{M}\,, \tag{12.30}$$

assuming that $2K \geq M > 0$.

Proof. We $N = 1$ in Hamiltonian (12.29) to find

$$H(P,R) = \frac{1}{16R^2}(1+R^2)^2(P^2R^2 + M^2).$$

We notice that for a rotating Puckon with $H = K^2 = const$, the girdle of the Puckon cannot have zero radius unless $M = 0$, in which case we return to the irrotational Puckon. Likewise, the girdle radius cannot be infinite unless again $M = 0$. Thus, a rotating Puckon with $M \neq 0$ is constrained to lie between a maximum and a minimum radius. To find these extremal radii, we must solve

$$\frac{\partial H}{\partial P} = \frac{P}{8}(1+R^2)^2 = 0\,,$$
$$H = \frac{1}{16R^2}(1+R^2)^2(P^2R^2 + M^2) = K^2.$$

The first equation can only be solved by $P = 0$; for which the second becomes

$$\frac{M^2}{16R^2}(1+R^2)^2 = K^2.$$

Assuming that K and M are both positive, we find that the maximum and minimum values of R are the roots,

$$R_{max/min} = \frac{2K \pm \sqrt{4K^2 - M^2}}{M}$$

This proves the proposition. □

Remark 12.4 (A Single Critical Point). We observe that if $M = 2K$, then the Puckon is radially static at the equator and P is identically zero. The solution $(P, R) = (0, 1)$ is the only critical point of the Hamiltonian H unless $M = 0$ in which case the set defined by $P = 0$ is the critical manifold.

Proposition 12.8 (Periodic motion of the basic rotating Puckon). *The motion of the rotating Puckon is periodic, with period $2\pi/\sqrt{2H}$ determined by the constant value of the Hamiltonian, H.*

Proof. By using the co-latitude representation of periodic rotating Puckon motion to investigate the motion of the rotating Puckon, we pass from stereographic coordinates to longitude-co-latitude coordinates in which we put

$$R = \tan\frac{\Phi}{2}.$$

As we will see later in the general case, this does not alter the situation a great deal. For example, the momentum is given later in Eq. (12.39), after substituting $\sin$ for ψ. The Hamiltonian for the rotating Puckon on the sphere in longitude-co-latitude coordinates is

$$H(P, \Phi) = \frac{1}{2}\left(P^2 + \frac{M^2}{\sin^2\Phi}\right), \tag{12.31}$$

in which Φ and P are canonically conjugate variables. Along the motion of the Puckon, H is constant, say $H = K^2$. Thus, taking the positive branch for P

$$P = \sqrt{2K^2 - \frac{M^2}{\sin^2\Phi}}$$

yields the equation for the co-latitude

$$\dot{\Phi} = P = \sqrt{2K^2 - \frac{M^2}{\sin^2\Phi}}.$$

Integration of this ODE implies that

$$\begin{aligned} t &= \int_{\Phi(0)}^{\Phi(t)} \frac{\sin\Phi\,\mathrm{d}\Phi}{\sqrt{2K^2\sin^2\Phi - M^2}} \\ &= \frac{1}{K\sqrt{2}}\left(\arccos\left(\frac{K\sqrt{2}\cos\Phi(t)}{\sqrt{2K^2 - M^2}}\right) - \arccos\left(\frac{K\sqrt{2}\cos\Phi(0)}{\sqrt{2K^2 - M^2}}\right)\right). \end{aligned}$$

Thus the co-latitude for the single rotating Puckon is given as a function of time by

$$\Phi(t) = \sqrt{\frac{2K^2\cos^2\Phi(0)}{2K^2 - M^2}}\cos(Kt\sqrt{2}) - \sqrt{\frac{2K^2\sin^2\Phi(0) - M^2}{2K^2 - M^2}}\sin(Kt\sqrt{2}).$$

These rotating Puckon solutions are periodic with period $2\pi/(K\sqrt{2})$, determined by the constant value K^2 of the Hamiltonian. □

12.3.5 Puckons and Geodesics

Proposition 12.9 (Geodesic Motion of a Point on the Girdle of the Rotating Puckon). *A point on the girdle of the rotating Puckon moves along a great circle at constant speed equal to $K\sqrt{2}$ determined by the value $H = K^2$ of the Hamiltonian, H. The normal to its plane of motion is inclined to the South Pole at the angle $\pi - \arctan(2R_{max})$ with R_{max} given in (12.30).*

Proof. For a rotationally symmetric surface $\mathscr{M}$ with metric

$$g = \mathrm{d}r^2 + \psi(r)^2\mathrm{d}\theta,$$

the equations for a curve $\boldsymbol{x} = (R(t), \Theta(t))$ to be a geodesic are equivalent to

$$\dot{R}^2 + \psi^2\dot{\Theta}^2 = 1 \tag{12.32}$$

$$\psi^2\dot{\Theta} = const. \tag{12.33}$$

So in the case of the sphere with metric $g = \mathrm{d}\phi^2 + \sin^2\phi\,\mathrm{d}\theta$, if we define

$$\dot{\Theta} = u_\theta(\Phi) = \frac{M^2}{\sin^2\Phi} \tag{12.34}$$

then we obtain a curve on the sphere with coordinates $\boldsymbol{x} = (\Phi, \Theta)$ and tangent vector

$$\dot{\boldsymbol{x}}(t) = (\dot{\Phi}\,,\,\dot{\Theta}).$$

Now, the speed is given by

$$\begin{aligned}
|\dot{\boldsymbol{x}}|^2 &= g(\dot{\boldsymbol{x}}, \dot{\boldsymbol{x}}) \\
&= \dot{\Phi}^2 + (\sin^2\Phi)\,\dot{\Theta}^2 \\
&= P^2 + \frac{M^2}{\sin^2\Phi} \\
&= 2H(P, \Phi) = constant = 2K^2.
\end{aligned}$$

Thus a point on the geodesic has constant speed equal to $\sqrt{2H}$ and by (12.34) we have

$$(\sin^2 \Phi)\dot{\Theta} = M^2 = constant.$$

Consequently, $\dot{\boldsymbol{x}}/(K\sqrt{2})$ satisfies equations (12.32) and (12.33), thereby determining a geodesic on the sphere. This means that a point on the girdle of the Puckon moves along a great circle at constant speed equal to $\sqrt{2H}$ and the normal to its plane of motion is inclined to the South Pole at the angle $\pi - \arctan(2R_{max})$, with R_{max} given in Eq. (12.30). □

12.3.6 Further Hamiltonian Aspects of Radial Solutions of EPDiff on the Riemann Sphere

Proposition 12.10 (Lie–Poisson Hamiltonian form of EPDiff on the Riemann Sphere). *The radially symmetric solutions of EPDiff on the Riemann sphere may be written as*

$$\frac{\partial}{\partial t}\begin{pmatrix} m_r \\ rm_\theta \end{pmatrix} = \mathscr{D}\begin{pmatrix} u_r \\ \frac{u_\theta}{r} \end{pmatrix} = \mathscr{D}\begin{pmatrix} \delta H/\delta m_r \\ \delta H/\delta(rm_\theta) \end{pmatrix}, \tag{12.35}$$

where $\mathscr{D}$ is the skew-symmetric Hamiltonian operator given by

$$\mathscr{D} = \begin{pmatrix} -\left(m_r\partial_r + \partial_r m_r - 2m_r\rho r + \frac{m_r}{r}\right) & -rm_\theta\partial_r \\ \left(-\partial_r rm_\theta + 2m_\theta r^2\rho - m_\theta\right) & 0 \end{pmatrix}, \tag{12.36}$$

and the velocities u_r and u_θ/r are given by the variational derivatives of the Hamiltonian,

$$u_r = \frac{\delta H}{\delta m_r}, \quad \frac{u_\theta}{r} = \frac{\delta H}{\delta rm_\theta}.$$

Equations (12.35) and (12.36) provide the Lie–Poisson Hamiltonian form of the EPDiff equation for radially symmetric dynamics on the Riemann sphere.

Proof. In the case that

$$\begin{aligned} \boldsymbol{u} &= \sum_{i=1}^{N} \frac{G(r,R_i)}{\rho(R_i)^2 R_i \rho^2 r}\left(R_i P_i r\partial_r + M_i\partial_\theta\right) \\ &=: u_r\partial_r + \frac{u_\theta}{r}\partial_\theta. \end{aligned}$$

where the P_i and R_i satisfy (12.26) and (12.27), the associated momentum density is

$$\boldsymbol{m} = \Delta^{\mathrm{d}}\boldsymbol{u}_\flat = m_r \mathrm{d}r + r m_\theta \mathrm{d}\theta,$$

where

$$m_r = \sum_{i=1}^{N} \frac{\delta(r-R_i)}{\rho(R_i)^2 R_i} P_i\,, \quad \text{and} \quad r m_\theta = \sum_{i=1}^{N} \frac{\delta(r-R_i)}{\rho(R_i)^2 R_i} M_i.$$

Let $w_1 = \frac{f}{\rho^2 r}\partial_r$ and $w_2 = \frac{g}{\rho^2 r}\partial_\theta$ be two vector fields on S^2. These satisfy the commutator relations,

$$\begin{aligned}
[w_1, \boldsymbol{u}] &= \left[\frac{f}{\rho^2 r}\partial_r, u_r\partial_r + \frac{u_\theta}{r}\partial_\theta\right] \\
&= \left(\frac{f u_r'}{\rho^2 r} - \frac{u_r \partial_r f}{\rho^2 r} - 2\frac{f u_r}{\rho} + \frac{f u_r}{\rho^2 r^2} - \frac{u_\theta \partial_\theta f}{\rho^2 r^2}\right)\partial_r + \left(\frac{f u_\theta'}{\rho^2 r^2} - \frac{f u_\theta}{\rho^2 r^3}\right)\partial_\theta,
\end{aligned}$$

and

$$\begin{aligned}
[w_2, \boldsymbol{u}] &= \left[\frac{g}{\rho^2 r}\partial_\theta, u_r\partial_r + \frac{u_\theta}{r}\partial_\theta\right] \\
&= -\left(\frac{u_r \partial_r g}{\rho^2 r} + 2\frac{g u_r}{\rho} - \frac{g u_r}{\rho^2 r^2} + \frac{u_\theta \partial_\theta g}{\rho^2 r^2}\right)\partial_\theta.
\end{aligned}$$

Thus, the EPDiff equations become, for w_1,

$$\begin{aligned}
0 &= \int_{\mathrm{S}^2} \left(\frac{\partial}{\partial t}\boldsymbol{m}(w_1) + \boldsymbol{m}([w_1, \boldsymbol{u}])\right) \mathrm{d}vol \\
&= \int_0^{2\pi}\int_0^{\infty} \left(\frac{\partial m_r}{\partial t}\frac{f}{\rho^2 r} + m_r\left(\frac{f u_r'}{\rho^2 r} - \frac{u_r\partial_r f}{\rho^2 r} - 2\frac{f u_r}{\rho} + \frac{f u_r}{\rho^2 r^2} - \frac{u_\theta \partial_\theta f}{\rho^2 r^2}\right)\right)\rho^2 r \mathrm{d}r\mathrm{d}\theta \\
&\quad + \int_0^{2\pi}\int_0^{\infty} r m_\theta \left(\frac{f u_\theta'}{\rho^2 r^2} - \frac{f u_\theta}{\rho^2 r^3}\right)\rho^2 r \mathrm{d}r\mathrm{d}\theta \\
&= \int_0^{2\pi}\int_0^{\infty} \left(\frac{\partial m_r}{\partial t} f + m_r\left(f u_r' - u_r\partial_r f - 2 f u_r \rho r + \frac{f u_r}{r}\right)\right)\mathrm{d}r\mathrm{d}\theta \\
&\quad + \int_0^{2\pi}\int_0^{\infty} r^2 m_\theta \left(\frac{f u_\theta'}{r^2} - \frac{f u_\theta}{r^3}\right)\mathrm{d}r\mathrm{d}\theta
\end{aligned}$$

From this we obtain the radial equation

$$\frac{\partial m_r}{\partial t} = -\left(m_r\partial_r + \partial_r m_r - 2 m_r \rho r + \frac{m_r}{r}\right) u_r - r m_\theta \partial_r \frac{u_\theta}{r}. \quad (12.37)$$

Similarly, for w_2,

$$
\begin{aligned}
0 &= \int_{\mathrm{S}^2} \left(\frac{\partial}{\partial t} \boldsymbol{m}(w_2) + \boldsymbol{m}([w_2, \boldsymbol{u}]) \right) \mathrm{d}vol \\
&= \int_0^{2\pi} \int_0^{\infty} \left(\frac{\partial r m_\theta}{\partial t} \frac{g}{\rho^2 r} - r m_\theta \left(\frac{u_r \partial_r g}{\rho^2 r} + 2\frac{g u_r}{\rho} - \frac{g u_r}{\rho^2 r^2} + \frac{u_\theta \partial_\theta g}{\rho^2 r^2} \right) \right) \rho^2 r \mathrm{d}r\mathrm{d}\theta \\
&= \int_0^{2\pi} \int_0^{\infty} \left(\frac{\partial r m_\theta}{\partial t} g - r m_\theta \left(u_r \partial_r g + 2 g u_r r \rho - \frac{g u_r}{r} \right) \right) \mathrm{d}r\mathrm{d}\theta
\end{aligned}
$$

Hence we have the azimuthal equation

$$
\frac{\partial r m_\theta}{\partial t} = \left(-\partial_r r m_\theta + 2 m_\theta r^2 \rho - m_\theta \right) u_r . \tag{12.38}
$$

Equations (12.37) and (12.38) provide the Lie–Poisson form of EPDiff on the Riemann Sphere. □

12.4 Numerical Solutions for EPDiff on the Sphere

12.4.1 Overview

We present numerical solutions to both the EPDiff partial differential equations (12.35, 12.36) and the corresponding ordinary differential equations (12.26, 12.27) for Puckons. Instead of using the stereographic projection, the numerical solutions were calculated on the sphere with co-latitude–longitude (ϕ, θ) coordinates (and canonical variables Φ, P instead of R, P).

The equations in (ϕ, θ) coordinates on the sphere are obtained from those in Sect. 12.3 using $\psi = \sin\phi$, so that $g = \mathrm{d}\phi^2 + \sin^2\phi\, \mathrm{d}\theta^2$.

We also introduce a length scale α for our numerical solutions by effectively changing the radius of the sphere from 1 to α.

In this case $\alpha^2 \Delta^{\mathrm{d}} = 1 + \alpha^2 \Delta^{\nabla}$ so that the Green's function is

$$
G(\phi, \Phi) = \frac{\alpha}{2} \begin{cases} \left(\tan\frac{\phi}{2} \cot\frac{\Phi}{2} \right)^{1/\alpha}, & \phi < \Phi , \\ \left(\tan\frac{\Phi}{2} \cot\frac{\phi}{2} \right)^{1/\alpha}, & \phi > \Phi . \end{cases}
$$

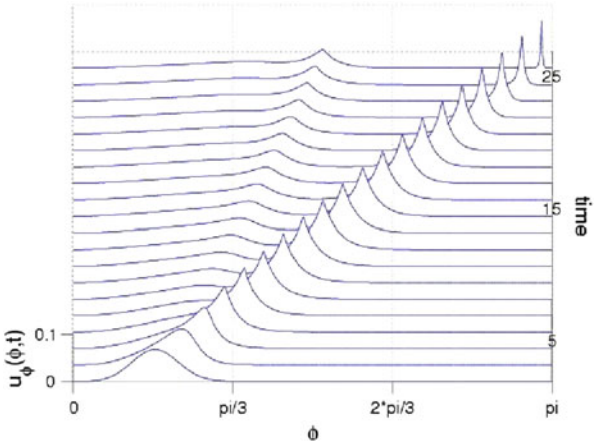

Fig. 12.1 The PDE simulations show an initially Gaussian distribution of velocity breaking up into Puckons, two of which are evident for the times shown

12.4.2 Numerical Specifications

Numerical simulations for diffeons on the sphere were performed using both PDEs and ODEs. In simulating the Lie–Poisson partial differential equations (12.35, 12.36) in co-latitude–longitude coordinates in Proposition 12.10, fourth-order finite differences were used to calculate spatial derivatives, and the momentum was advanced in time using a fourth-order Runge–Kutta scheme. The Hamiltonian was fond to be conserved in the PDE simulations to within 10^{-4} of its initial value for the simulations with smooth initial velocity distributions, but only to within 5–10% for the simulations when using one or two Puckons as the initial velocity distribution. The ordinary differential equations (12.26, 12.27) for the canonical variables in co-latitude–longitude coordinates were advanced in time using a fourth-order Runge–Kutta scheme. For these ODE simulations, the Hamiltonian was conserved to within 10^{-8}. The PDE simulation results will be shown as a velocity distribution changing in time, whereas the results of the ODE simulations will be shown as plots of the time evolution of the canonical variables Φ, P. The length scale α in the simulations was set to 0.1 unless otherwise noted.

Irrotational Puckons.

We first consider irrotational Puckons ($u_\theta = 0$). Figure 12.1 shows the evolution when the initial meridional velocity $u_\phi(\theta, 0)$ is a Gaussian. The Gaussian is chosen as a typical smooth confined initial condition on which to demonstrate the general

property of emergence of the singular Puckons. As time elapses ($t = 0$ is at the bottom of the figure, and later times are shown above it), a Puckon emerges from the Gaussian in Fig. 12.1, and a second Puckon begins to emerge. In agreement with the prediction of Eq. (12.31), the first Puckon retains its height (and thus retains its velocity and canonical momentum P) as it approaches the South Pole.

An overtaking collision between two irrotational Puckons is shown in Fig. 12.2. As the plot of co-latitudes shows, these two Puckons do not pass through each other. Instead, they bounce and exchange momenta. However, their momenta are not exchanged exactly, as the plot of $P(t)$ shows. This behavior contrasts with 2-soliton collisions for completely integrable PDEs in which momentum is exactly exchanged.

A head-on collision between two irrotational Puckons is shown in Fig. 12.3. In the PDE simulation, a vertical slope appears to form in finite time. Note that the Puckon velocities (i.e., the Puckon heights in the PDE simulation or the slopes of $\Phi(t)$ in the ODE simulation) remain finite and actually decrease to zero, whereas the equal and opposite canonical radial Puckon momenta diverge as the collision takes place.

Rotating Puckons.

Now we consider numerical simulations of rotating Puckons ($u_\theta \neq 0$). Figure 12.4 shows the PDE evolution when the initial meridional velocity u_ϕ is zero and the initial azimuthal velocity u_θ is a Gaussian. Two rotating Puckons have emerged from the Gaussian by the time shown, and they are moving toward opposite poles. The Puckons each have small but nonzero azimuthal velocities.

Finally, Fig. 12.5 shows an ODE simulation of a single rotating Puckon when the length scale is $\alpha = 1$ and the canonical variables are initially $\Phi = 1.5$, $P = -1$, $M = -2$. As discussed in Sect. 12.3.4, the basic rotating Puckon moves between minimum and maximum co-latitudes. The Puckon is initially moving toward the North Pole, and one can see that its meridional velocity vanishes as it changes its direction and moves toward the South Pole. Its azimuthal velocity is negative throughout, i.e., it does not change its sense of rotation. Numerically, the Puckon moves between minimum and maximum co-latitudes of (to five significant digits) 1.1033 and 2.0384, in comparison with the corresponding values of 1.1031 and 2.0385 predicted by the formula given in Sect. 12.3.4. Also, the numerical result for the period of the Puckon's motion (averaged over the first five periods) agrees to six significant digits (5.60858) with the value given in Sect. 12.3.4. Thus, these numerical results for the ODE dynamics are accurate to between four and six significant figures.

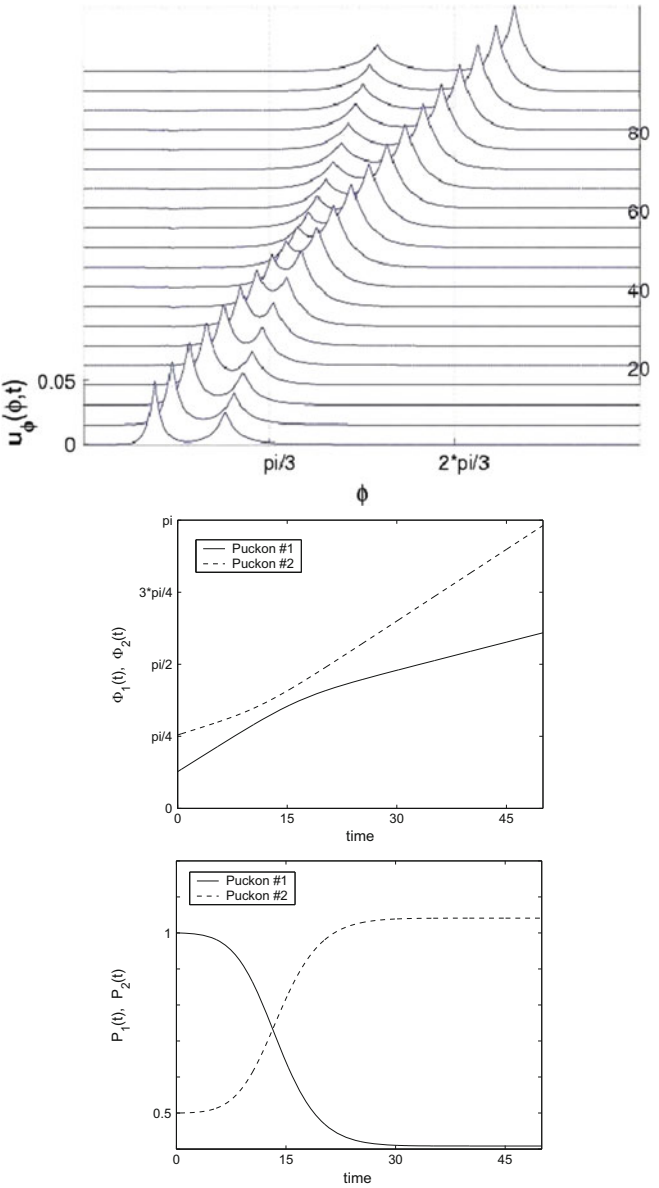

Fig. 12.2 The *upper panel* shows two Puckons undergoing an overtaking collision. The *lower panels* show the evolution of the co-latitudes (Φ_1, Φ_2) and canonical momenta (P_1, P_2) of the Puckons. The PDE simulation agrees with the co-latitudes Φ of the ODE simulation to within 1 % and with the canonical momenta P to within 3 %

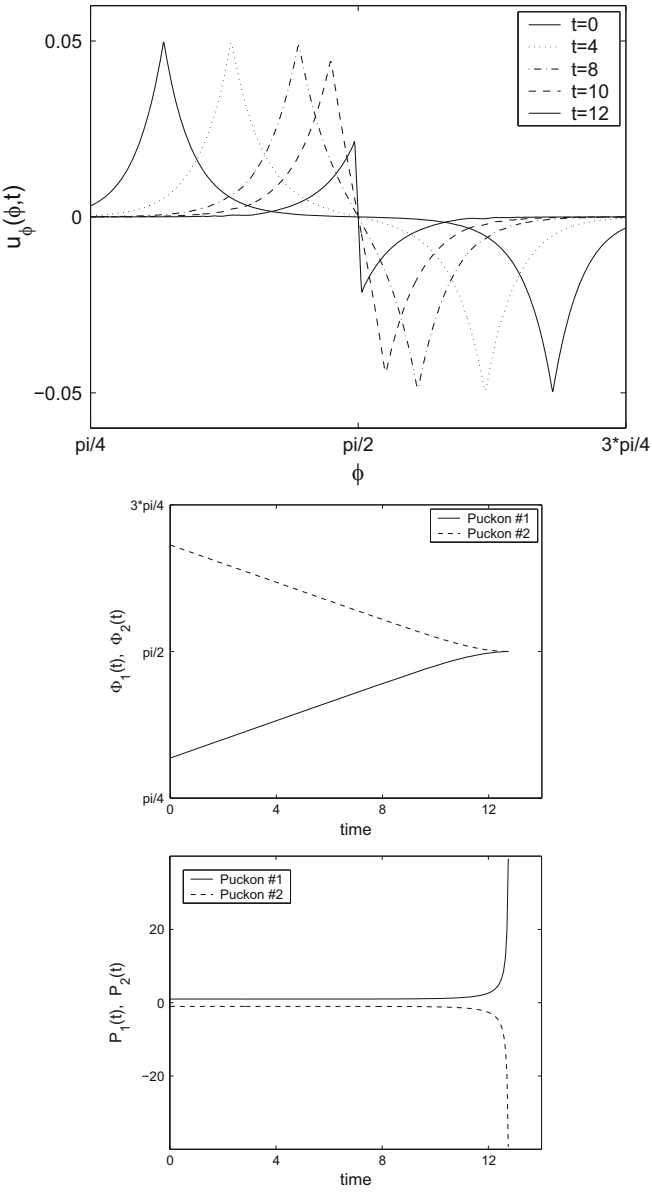

Fig. 12.3 The *upper panel* shows the velocity profile of two Puckons undergoing a head-on collision. The *lower panels* show the evolution of the co-latitudes (Φ_1, Φ_2) and canonical momenta (P_1, P_2) of the Puckons. The Puckon velocities remain finite and tend to zero during the collision, whereas the equal and opposite canonical momenta of the Puckons diverge. In the PDE simulation, a vertical slope appears to form in finite time

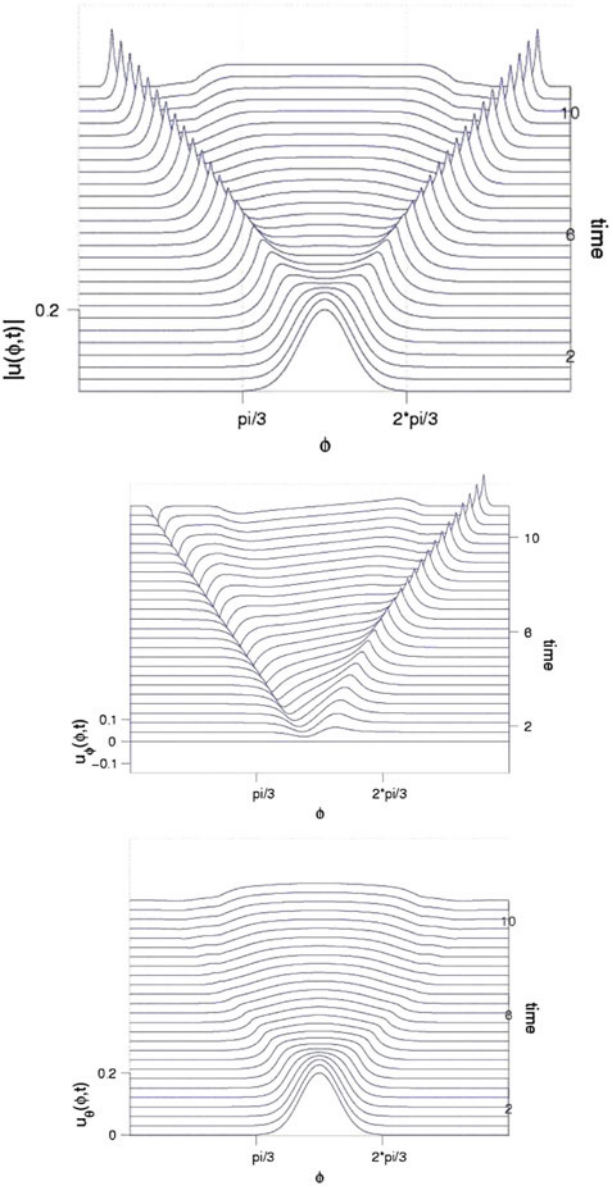

Fig. 12.4 The evolution when the initial meridional velocity u_ϕ is zero and the initial azimuthal velocity u_θ is a Gaussian. The *larger picture* shows the magnitude of the velocity. Two rotating Puckons have emerged at the time shown, and two more are in the process of emerging from the Gaussian

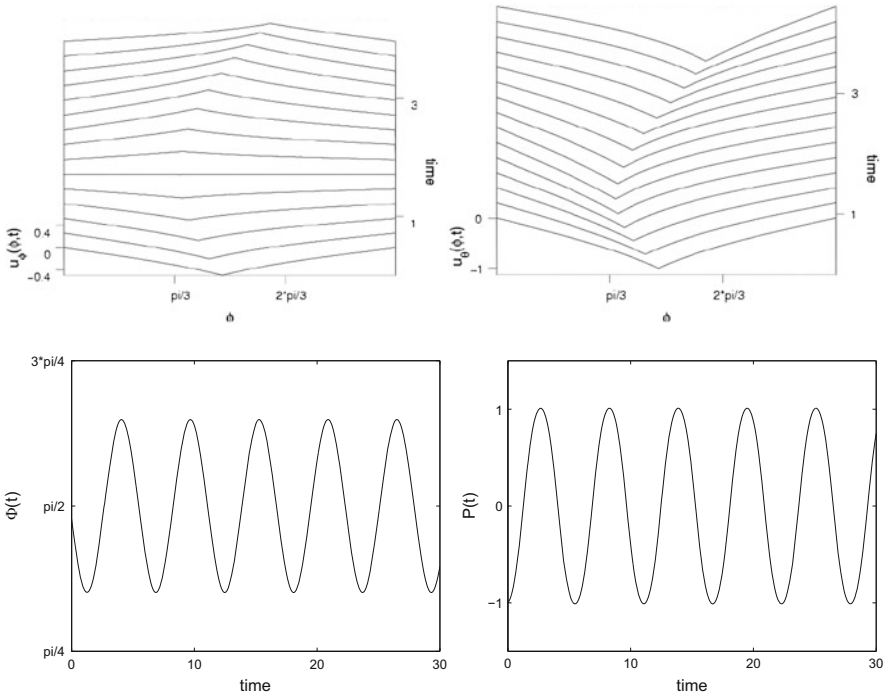

Fig. 12.5 The evolution of a single rotating Puckon is shown for length scale $\alpha = 1$. The initial parameters are $\Phi = 1.5$, $P = -1$, $M = -2$. The PDE simulation shown in the *upper panels* contains about half of the period of the Puckon's motion. The ODE simulation shown in the *lower panels* contains about five periods of the Puckon's motion. The meridional and azimuthal velocity components u_ϕ and u_θ are shown for the PDE simulation

12.5 Generalizing to Other Surfaces

We now wish to extend these Hamiltonian reductions for diffeons on the sphere to other surfaces. This will require us to summon considerably more resources from the differential geometry of Riemannian surfaces than in the previous sections. These resources from differential geometry will be used to identify conditions for which the singular momentum map property of EPDiff and the corresponding reduction to canonically Hamiltonian ODEs persist. In particular, these properties will be shown to persist for Riemannian manifolds Σ with a free isometric action of a Lie group G, so that Σ/G is a one-dimensional manifold N, and Σ may be taken as $\Sigma \cong G \times N$. For the moment, we retain the rotational symmetry. We will pass to the more general case in Sect. 12.6.

12.5.1 Rotationally Symmetric Surfaces

Any surface Σ with an isometric action of S^1 has rotationally invariant coordinate charts on which the metric is given by

$$g = \mathrm{d}r^2 + \psi(r)^2 \mathrm{d}\theta^2$$

where $(r, \theta) \in (r_{min}, r_{max}) \times (0, 2\pi)$. We may take $r_{min} = 0$ to describe a fixed point of the motion. Let ∇ be the Levi–Civita connection with respect to this metric and set

$$\mathbb{I} = (\mathbb{1} + \Delta^{\nabla}).$$

By the Bochner–Weitzenböck theorem on 1-forms we compute

$$\begin{aligned} \mathbb{I} &= \mathbb{1} + \Delta^d - \mathrm{Ric} \\ &= \left(\frac{\psi + \psi''}{\psi} \right) \mathbb{1} + \Delta^{\mathrm{d}}. \end{aligned}$$

We also have the explicit relation,

$$\Delta^{\mathrm{d}}(a(r)\mathrm{d}r + b(r)\mathrm{d}\theta) = -\frac{\partial}{\partial r}\left(\frac{1}{\psi}\frac{\partial}{\partial r}(\psi a(r))\right)\mathrm{d}r - \psi\frac{\partial}{\partial r}\left(\frac{1}{\psi}\frac{\partial}{\partial r}(b(r))\right)\mathrm{d}\theta.$$

Suppose that $\boldsymbol{u}$ is a rotationally invariant vector field on Σ. That is, suppose

$$\boldsymbol{u} = u^r(r)\partial_r + u^\theta(r)\partial_\theta \,,$$

then the singular momentum solution ansatz for EPDiff,

$$\mathbb{I}\boldsymbol{u}_\flat = \frac{\delta(r-R)}{\psi(R)}\mathrm{d}r + \psi(R)\delta(r-R)\mathrm{d}\theta \,, \tag{12.39}$$

is a system of ordinary differential equations in r. We may solve this system to obtain two rotationally invariant, continuous functions G^r, G^θ of r and R, within our coordinate chart, which are symmetric in the two variables and which satisfy

$$\mathbb{I}\left(G^r(r,R)\mathrm{d}r + G^\theta(r,R)\mathrm{d}\theta\right) = \frac{\delta(r-R)}{\psi(R)}\mathrm{d}r + \psi(R)\delta(r-R)\mathrm{d}\theta \,.$$

The two Green's functions G^r, G^θ are related by

$$G^\theta(r,R) = \psi(r)\psi(R)G^r(r,R).$$

They also have a jump in the derivative along the diagonal $r = R$. Thus the solution of (12.39) for the velocity $\boldsymbol{u}$ is

$$\begin{aligned}\boldsymbol{u} &= (G^r(r,R)\mathrm{d}r + \psi(r)\psi(R)G^r(r,R)\mathrm{d}\theta)^\sharp \\ &= G^r(r,R)\frac{\partial}{\partial r} + \psi(R)\frac{G^r(r,R)}{\psi(r)}\frac{\partial}{\partial\theta}\,.\end{aligned}$$

The EPDiff equations on the rotationally symmetric surface Σ are

$$\boldsymbol{m} = \mathbb{I}\boldsymbol{u}_\flat = \sum_{i=1}^{N}\frac{\delta(r-R_i)}{\psi(R_i)}\left(P_i\mathrm{d}r + M_i\mathrm{d}\theta\right) \tag{12.40}$$

$$0 = \frac{\partial \boldsymbol{m}}{\partial t} + \mathrm{ad}^*_{\boldsymbol{u}}\boldsymbol{m}\,, \tag{12.41}$$

and these equations extremize

$$\begin{aligned}\ell[\boldsymbol{u}] &= \frac{1}{2}\int_\Sigma (|\boldsymbol{u}|_g^2 + |\nabla\boldsymbol{u}|_g^2)\mathrm{d}vol_g \\ &= \frac{1}{2}\int_\Sigma \langle \boldsymbol{u}, \mathbb{I}\boldsymbol{u}\rangle\,\mathrm{d}vol_g \\ &= \frac{1}{2}\int_\Sigma \boldsymbol{m}(\boldsymbol{u})\mathrm{d}vol_g\,.\end{aligned}$$

After the Legendre transformation we arrive at the Hamiltonian

$$\begin{aligned}H[\boldsymbol{m}] &= \int_\Sigma \boldsymbol{m}(\boldsymbol{u})\mathrm{d}vol - \ell[\boldsymbol{u}] \\ &= \frac{1}{2}\int_\Sigma \boldsymbol{m}(\boldsymbol{u})\mathrm{d}vol.\end{aligned}$$

We already know that the solution to (12.40) is

$$\boldsymbol{u}_\flat = \sum_{i=1}^{N}\frac{G^r(r,R_i)}{\psi(R_i)}\Big(P_i\psi(R_i)\mathrm{d}r + M_i\psi(r)\mathrm{d}\theta\Big),$$

that is,

$$\begin{aligned}\boldsymbol{u} &= \sum_{i=1}^{N}\frac{G^r(r,R_i)}{\psi(r)\psi(R_i)}\left(P_i\psi(R_i)\psi(r)\frac{\partial}{\partial r} + M_i\frac{\partial}{\partial\theta}\right) \\ &=: u_r\frac{\partial}{\partial r} + u_\theta\frac{\partial}{\partial\theta}\,.\end{aligned}$$

Because $\boldsymbol{m}$ is only supported within our coordinate chart and is invariant under the action of S^1, we have

$$H(\boldsymbol{m}) = \frac{1}{2} 2\pi \sum_{i=1}^{N} \left(P_i u_r(R_i) + M_i u_\theta(R_i) \right)$$

where 2π arises as the volume of S^1. Choose a vector field $w = f(r,\theta)\frac{\partial}{\partial r} + g(r,\theta)\frac{\partial}{\partial \theta}$. Again since $\boldsymbol{m}$ is supported within our chart, we know that

$$\int_\sigma \boldsymbol{m}(w) \mathrm{d}vol_g = \sum_i \int_0^{2\pi} \left(P_i f(R_i,\theta) + M_i g(R_i,\theta) \right) \mathrm{d}\theta.$$

We also have

$$[w, \boldsymbol{u}] = \left(f u_r' - u_r \partial_r f - u_\theta \partial_\theta f \right) \frac{\partial}{\partial r} + \left(f u_\theta' - u_r \partial_r g - u_\theta \partial_\theta g \right) \frac{\partial}{\partial \theta}.$$

Consequently, the EPDiff equations yield

$$\begin{aligned}
0 &= \int_\Sigma \left(\frac{\partial}{\partial t}(\boldsymbol{m}(w)) + \boldsymbol{m}([w,\boldsymbol{u}]) \right) \mathrm{d}vol_g \\
&= \int_0^{2\pi} \frac{\partial}{\partial t} \left(P_i f(R_i,\theta) + M_i g(R_i,\theta) \right) \mathrm{d}\theta \\
&+ \int_0^{2\pi} \left(P_i \left(f u_r' - u_r \partial_r f - u_\theta \partial_\theta f \right) + M_i \left(f u_\theta' - u_r \partial_r g - u_\theta \partial_\theta g \right) \right) \big|_{r=R_i} \mathrm{d}\theta \\
&= \int_0^{2\pi} \left(\dot{P}_i f(R_i,\theta) + P_i \dot{R}_i \partial_r f(R_i,\theta) + \dot{M}_i g(R_i,\theta) + M_i \dot{R}_i \partial_r g(R_i,\theta) \right) \mathrm{d}\theta \\
&+ \int_0^{2\pi} \left(P_i \left(f u_r' - u_r \partial_r f \right) + M_i \left(f u_\theta' - u_r \partial_r g \right) \right) \big|_{r=R_i} \mathrm{d}\theta.
\end{aligned}$$

Comparing coefficients of $f, \partial_r f, g \partial_r g$ now yields

$$\begin{aligned}
0 &= \dot{P}_i + P_i u_r'(R_i) + M_i u_\theta'(R_i) \\
0 &= P_i \dot{R}_i - P_i u_r(R_i) \\
0 &= \dot{M}_i \\
0 &= M_i \dot{R}_i - M_i u_r(R_i)
\end{aligned}$$

Now since

$$u_r(R_i) = \sum_{j=1}^{N} G^r(R_i, R_j)$$

and

$$u_r'(R_i) = \sum_{j=1}^{N} \frac{\partial G^r}{\partial r}(R_i, R_j)$$

we see that

$$\begin{aligned}
\frac{\partial}{\partial R_i} \sum_{j=1}^{N} u_r(R_j) &= \frac{\partial}{\partial R_i} \sum_{j,k=1}^{N} G^r(R_j, R_k) \\
&= \sum_{j,k=1}^{N} \left(\frac{\partial R_j}{\partial R_i} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_j, R=R_k} + \frac{\partial R_k}{\partial R_i} \left. \frac{\partial G^r(r,R)}{\partial R} \right|_{r=R_j, R=R_k} \right) \\
&= \sum_{j,k=1}^{N} \left(\frac{\partial R_j}{\partial R_i} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_j, R=R_k} + \frac{\partial R_k}{\partial R_i} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_k, R=R_j} \right) \\
&\quad \text{since } G^r \text{ is symmetric in its arguments} \\
&= \sum_{j,k=1}^{N} \left(\delta_{ij} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_j, R=R_k} + \delta_{ki} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_k, R=R_j} \right) \\
&= \sum_{k=1}^{N} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_i, R=R_k} + \sum_{j=1}^{N} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_i, R=R_j} \\
&= 2 \sum_{j=1}^{N} \left. \frac{\partial G^r(r,R)}{\partial r} \right|_{r=R_i, R=R_j} \\
&= 2u_r'(R_i).
\end{aligned}$$

Similarly

$$u_\theta'(R_i) = \frac{1}{2} \frac{\partial}{\partial R_i} \sum_{j=1}^{N} u_\theta(R_j).$$

This finally yields

$$\dot{P}_i = -\frac{1}{2} \frac{\partial}{\partial R_i} \sum_{j=1}^{N} (P_j u_r(R_j) + M_j u_\theta(R_j)) \tag{12.42}$$

$$\dot{R}_i = u_r(R_i) \tag{12.43}$$

$$\dot{M}_i = 0. \tag{12.44}$$

Thus, we have proved the following realization of the momentum map in [8].

Proposition 12.11 (Canonical Equations for Rotationally Symmetric Diffeons). *The parameters P_i, R_i for diffeons on a rotationally symmetric surface Σ with constant positive curvature and metric given by $g = \mathrm{d}r^2 + \psi(r)^2\mathrm{d}\theta^2$ satisfy Hamilton's canonical equations with Hamiltonian given by*

$$\begin{aligned} H(\mathbf{P},\mathbf{R};\mathbf{M}) &= \frac{1}{2\pi}\left(\int_\Sigma \boldsymbol{m}(\boldsymbol{u})\mathrm{d}vol - \ell[\boldsymbol{u}]\right) \\ &= \frac{1}{2}\sum_{i=1}^{N}\left(P_i u_r(R_i) + M_i u_\theta(R_i)\right) \\ &= \frac{1}{2}\sum_{i,j=1}^{N}\frac{G^r(R_i,R_j)}{\psi(R_i)\psi(R_j)}\left(P_iP_j\psi(R_i)\psi(R_j) + M_iM_j\right). \end{aligned}$$

A solution to (12.42, 12.43, 12.44) is an example of an N-diffeon on a rotationally symmetric surface Σ.

We notice that the critical points of H are those $\mathbf{R} = (R_1,\ldots,R_n)$ such that $G^r(R_i,R_j) = 0$ for all i,j. But the value of H at these critical points is always 0 since H depends on G^r.

Remark 12.5. The only one-dimensional Lie groups are $\mathbb{R}$ and S^1, and while we have considered the isometric action of S^1 on a surface, we have to be more careful with translation invariance because the group ceases to be compact. However we can consider θ, which parameterizes each orbit $r = const$, taking values in $(-L, L)$ rather than $(-\infty,\infty)$ for the full translation group. This makes the integration finite.

12.5.2 Rotationally Invariant Diffeons on Hyperbolic Space

Hyperbolic space has a richer structure than either the plane, or the sphere. Indeed, the isometry group of the sphere is SO3; so all isometries are rotations. We have already examined the case of rotationally invariant diffeons. (These are the Puckons.) The isometry group of the plane $\mathbb{R}^2$ is $\mathrm{S}^1 \ltimes \mathbb{R}^2$, the semi-direct product of rotations and translations. Translational invariance yields the direct product of the original one-dimensional peakons, whereas rotational invariance yields the rotating circular peakons developed in [10]. However, the isometry group of hyperbolic space is $\mathbb{P}\mathrm{SL}(2,\mathbb{R})$, which is not compact and contains three different types of isometry. These are the rotational, translational, and horolational subgroups.

The Hamiltonian for Hyperbolic N-Diffeons with Rotational Symmetry.

We have already done most of the work for the rotationally symmetric case. All that remains is to write the hyperbolic metric in a conformally flat way and from it

deduce the Green's functions G^r and G^θ. The model is familiar: we use the Poincaré disc $D = \{(x, y) \in \mathbb{R}^2 |\ x^2 + y^2 < 1\}$ in polar coordinates with the metric

$$g_{\mathscr{H}} = \frac{4}{(1-r^2)^2}\left(\mathrm{d}r^2 + r^2\mathrm{d}\theta^2\right),$$

i.e., the conformal parameter is

$$\rho(r) = \frac{2}{1-r^2}\,.$$

The Bochner–Weitzenböck formula states that on $\mathrm{T}D$

$$\Delta^{\mathrm{d}} = \Delta^{\nabla} - \mathbb{1}\,,$$

so that

$$\mathbb{I} = \mathbb{1} + \Delta^{\nabla} = 2\,\mathbb{1} + \Delta^{\mathrm{d}}. \tag{12.45}$$

We wish to find Green's functions G^r and G^θ such that

$$\mathbb{I}(G^r(r,R)\mathrm{d}r + G^\theta(r,R)\mathrm{d}\theta) = \delta(r-R)\left(\frac{1}{R}\mathrm{d}r + R\mathrm{d}\theta\right).$$

We also ask that these Green's functions be finite at the origin, continuous and symmetric in their variables. Explicitly, we have

$$\Delta^{\mathrm{d}}\left(a(r)\mathrm{d}r + b(r)\mathrm{d}\theta\right) = -\frac{\partial}{\partial r}\left(\frac{1}{\rho^2 r}\frac{\partial}{\partial r}(ra(r))\right)\mathrm{d}r - r\frac{\partial}{\partial r}\left(\frac{1}{\rho^2 r}\frac{\partial b}{\partial r}\right)\mathrm{d}\theta.$$

Upon setting

$$F(r) = 2 - 2r^4 + 8\log(r)r^2 + r^2$$

we write

$$G^r(r,R) = \frac{\rho(r)^2\rho(R)^2}{16}\frac{\min(r,R)}{\min(r,R)}F(\max(r,R)),$$

and hence

$$G^\theta(r,R) = RrG^r(r,R).$$

Thus, we have shown the following.

Proposition 12.12 (Hamiltonian for Rotationally Invariant Hyperbolic N-Diffeons). *The Hamiltonian for rotationally invariant N-diffeons on hyperbolic space is given by*

$$H(\mathbf{P},\mathbf{R};\mathbf{M}) = \frac{1}{2}\sum_{i,j=1}^{N} \frac{1}{\rho(R_i)^2 R_i \rho(R_j)^2 R_j}\left(P_i P_j R_i R_j G^r(R_i,R_j) + \frac{M_i M_j}{R_i R_j} G^\theta(R_i,R_j)\right)$$
$$= \frac{1}{2}\sum_{i,j=1}^{N} \frac{G^r(R_i,R_j)}{\rho(R_i)^2 R_i \rho(R_j)^2 R_j}\left(P_i P_j R_i R_j + M_i M_j\right).$$

The Hyperbolic Diffeon.

Let us examine the behavior of the basic rotationally symmetric hyperbolic diffeon. The Hamiltonian is

$$H(P,R) = \frac{1}{2}\frac{G^r(R,R)}{\rho(R)^4 R^2}\left(P^2R^2 + M^2\right)$$
$$= \frac{F(R)}{32R^2}\left(P^2R^2 + M^2\right).$$

The function $F(R)/R^2$ is positive and continuous on $(0,1)$, and it has a singularity at $R=0$. Its derivative is negative on $(0,1)$ and its image is $(1,\infty)$. Consequently, $F(R)/R^2$ is an invertible function $(0,1) \longrightarrow (1,\infty)$.

Assuming $M \neq 0$, on the level set $H(P,R) = K^2$ we find that

$$P^2 = \frac{32K^2}{F(R)} - \frac{M^2}{R^2}.$$

So $\dot{R} = 0$, whenever

$$\frac{F(R)}{R^2} = \frac{32K^2}{M^2}.$$

Hence, whenever $32K^2 > M^2 > 0$, precisely one turning point exists for R. This proves the following.

Proposition 12.13. *Rotating diffeons with $M^2 > 0$ on a rotationally symmetric hyperbolic surface do not exhibit periodic behavior.*

For the irrotational diffeon, $M = 0$ and we have

$$P^2 = \frac{32K^2}{F(R)}.$$

Since $F(R) > 0$ for all $R \in [0,1]$, the irrotational diffeon has no turning points of R at all.

12.5.3 Horolationally Invariant Diffeons on Hyperbolic Space

We now turn to a subtly different problem. So far we have been using solely the rotation group (the circle) to produce symmetric diffeons. Now we consider a different subgroup of hyperbolic isometries. This time it is expedient to use the upper-half-plane model of hyperbolic geometry. That is,

$$\mathscr{H} = \{(x, y) \in \mathbb{R}^2 | y > 0\}$$

with the metric

$$g_{\mathscr{H}} = \frac{\mathrm{d}x^2 + \mathrm{d}y^2}{y^2}\,.$$

The group we consider is the horolation group $(x, y) \mapsto (x + a, y)$. This is a unique type of isometry in planar geometries and is, figuratively speaking, a "rotation about infinity." We can see immediately that the orbits of this group action are the lines $y = const$, and that we seek diffeons which are independent of x. In this situation by (12.45),

$$\mathbb{I}(a(y)\mathrm{d}x + b(y)\mathrm{d}y) = \frac{1}{y}\left(2ya - \frac{\partial^2 a}{\partial y^2}y^3 - 2\frac{\partial a}{\partial y}y^2\right)\mathrm{d}x + \left(2b - \frac{\partial^2 b}{\partial y^2}y^2 - 2\frac{\partial b}{\partial y}y\right)\mathrm{d}y$$

Thus the solution to

$$\mathbb{I}(G(y, Y)\mathrm{d}x) = \delta(y - Y)\mathrm{d}x$$

is

$$G(y, Y) = \frac{\min(y, Y)}{3\max(y, Y)^2},$$

and this solution also solves

$$\mathbb{I}(G(y, Y)\mathrm{d}y) = \delta(y - Y)\mathrm{d}y.$$

Next, we consider the compactness issue. Since the coordinate x ranges from $-\infty$ to ∞ we know that functions that only involve y cannot be integrated across all of $\mathscr{H}$. Thus the Diffeon Hamiltonian cannot exist over the whole space and the theory of Sect. 12.5.1 does not apply. Let us restrict ourselves to the strip $B_L = \{(x, y)|y > 0, |x| < L\} \subset \mathscr{H}$. By doing this we can apply similar calculations to those used in Sect. 12.5.1 to B_L, but we must apply them only in the case of vector fields which are tangent to the boundary $x = \pm L$.

Thus if

$$\boldsymbol{m} = \sum_{i=1}^{N} \delta(y - Y_i) Y_i^2 (M_i \mathrm{d}x + P_i \mathrm{d}y)$$

then we arrive at the Hamiltonian

$$\begin{aligned} H(\mathbf{P}, \mathbf{Y}) &= L \sum_{i,j=1}^{N} G(Y_i, Y_j) Y_i^2 Y_j^2 (P_i P_j + M_i M_j) \\ &= L \sum_{i,j=1}^{N} \frac{\min(Y_i, Y_j)}{3 \max(Y_i, Y_j)^2} Y_i^2 Y_j^2 (P_i P_j + M_i M_j) \\ &= L \sum_{i,j=1}^{N} \frac{\min(Y_i, Y_j)^3}{3} (P_i P_j + M_i M_j) \end{aligned}$$

and the condition that $\dot{M}_i = 0$.

For the 1-diffeon, the Hamiltonian is

$$H(P, Y) = L \frac{Y^3}{3} (P^2 + M^2)$$

so for $H(P, Y) = const = K$, the only turning point of Y is at

$$Y = \sqrt[3]{\frac{3K}{LM^2}}$$

or at 0, if $M = 0$.

Remark 12.6. Joining the edges of B_L together yields a surface called Gabriel's Horn, whereby the translation invariant diffeons on B_L become rotationally invariant diffeons on Gabriel's horn.

12.5.4 Translation Invariant Diffeons on Hyperbolic Space

Translations on the hyperbolic plane are characterized by having two ideal fixed points. The version of the metric we use in this case for the upper half plane is

$$g_{\mathscr{H}} = \mathrm{d}r + \cosh^2 r \mathrm{d}\theta^2 \,,$$

where the subsets formed by $r = const$ are the orbits of the translation. Again, the orbits are not compact, so we limit ourselves to the strip $B_L = \{(r, \theta) \in \mathbb{R}^2 | r > 0, |\theta| < L\} \subset \mathscr{H}$.

In this situation, the Green's function $G(r, R)$ solving

$$\mathbb{I}G(r, R)\mathrm{d}r = \frac{\delta(r - R)}{\cosh(R)}\mathrm{d}r$$

is given by

$$G(r, R) =, \frac{1}{2}\tanh(\min(r, R))\cosh(\max(r, R)) + \cosh(r)\cosh(R)\arctan(e^{\min(r,R)}).$$

Due to the complicated form of the Green's function, we will not write down the explicit form of the Hamiltonian for the N-diffeon. However, the Hamiltonian for the 1-diffeon is,

$$\begin{aligned} H(P, R; M) &= \frac{1}{2}\frac{G(R, R)}{\cosh^2 R}\left(P^2\cosh^2 R + M^2\right) \\ &= \frac{1}{4}\left(\tanh R\,\mathrm{sech}\, R + 2\arctan(e^R)\right)\left(P^2\cosh^2 R + M^2\right). \end{aligned}$$

In this case

$$\begin{aligned} \frac{\partial H}{\partial P} &= \frac{1}{2}\left(\tanh R\,\mathrm{sech}\, R + 2\arctan(e^R)\right) P\cosh^2 R\,, \\ \frac{\partial H}{\partial R} &= \frac{1}{2}\left(P^2\left(\cosh R + \sinh 2R\arctan(e^R)\right) + M^2\,\mathrm{sech}^3 R\right)\,. \end{aligned}$$

It can be shown that there are no equilibria for the motion unless $M = 0$, where the line $P = 0$ is the critical point set for H. However,

$$H(0, R) = 0\,,$$

for each value of R. So the line P consists entirely of stationary points, and since away from this line no points exist for which $\dot{R} = 0$, there can be no periodic behavior of 1-diffeons in this case.

Extending EPDiff on Hyperbolic Space.

Our study of EPDiff on hyperbolic spaces so far shows that the behavior of hyperbolic diffeons seems less interesting than the rich dynamical structure of

multiple Puckons interacting on the sphere. This is because hyperbolic diffeons will bounce only once before heading off to infinity, in the cases we have studied so far.

However, hyperbolic space is related intimately with the study of curves of genus at least 2: it forms the universal cover for all such Riemann surfaces. So far, we have been unable to examine the case of curves of genus ≥ 2 simply because their isometry groups are small (usually discrete) which means that no 1-parameter subgroup exists by which we may reduce the EPDiff equations to ODEs.

Any high genus Riemann surface can be realized as the quotient of the hyperbolic plane by a tiling, or hyperbolic lattice [19]. The complex structure of the Riemann surface depends upon the shape of the lattice. A surface Σ of genus $\gamma \geq 2$ is determined topologically by the symmetry group of the tiling, but a variety of complex structures may be put upon Σ to turn it into a Riemann Surface. If Γ is the symmetry group of the tiling, then the space of complex structures which may be put on Σ is given by the space

$$T_\Gamma = \frac{\Gamma\text{-invariant quasiconformal maps of } \mathbb{D}}{\text{Hyperbolic isometries of } \mathbb{D}}.$$

Here $\mathbb{D}$ is the Poincaré disc (the unit disc in the complex plane). In the space T_Γ a quasiconformal map is a generalization of a conformal map in which although angles are not preserved, the angular dilation is uniformly bounded, see [1, 5].

The inclusion of a subgroup Γ' in Γ induces a natural inclusion of T_Γ in $T_{\Gamma'}$, so it makes sense to consider the universal Teichmüller space as the Teichmüller space T_1 associated with the trivial group. The universal Teichmüller space is endowed with the Weil–Petersson metric, thus providing the framework essential for the theory of EPDiff. Furthermore, the restriction of the maps given in the definition of Teichmüller space to the boundary of the disc provides us with the essential method by which we reduce the EPDiff PDEs to ODEs. See [6, 7, 13] for some recent progress in this framework.

12.6 EPDiff on Warped Product Spaces

In the previous sections, we have studied rotationally symmetric solutions. That is, the solutions have been invariant under a circle action, and this effectively reduced $\mathbb{I}$ in (12.45) to an ordinary differential operator. As a result, we were able to find G^r and G^θ which are invariant under the group action, continuous and have a jump in the derivative along an orbit. This motivates the study of higher dimensional Riemannian manifolds (Σ, g) which possess an isometric action of the Lie group K such that the quotient space Σ/K is one-dimensional. In particular, we shall consider a Riemannian manifold Σ with a free isometric action of a Lie group K, so that Σ/K is a one-dimensional manifold N, and hence $\Sigma \cong K \times N$.

12.6.1 Warped Products

Let K be a compact semi-simple Lie group of dimension $n-1$ and $\Sigma \longrightarrow I \subset \mathbb{R}$ a principal K-bundle over the open interval I. (In what follows, one may also take I to be the circle S^1.) Suppose Σ has a Riemannian metric g preserved by K, and that coordinates exist such that the metric has the form of a warped product

$$g = \mathrm{d}r^2 + \psi(r)^2 g_K$$

where g_K is a bi-invariant inner product on $\mathrm{T}K$ and ψ is a K-invariant function on Σ, i.e., a function of r alone. The coordinate r parameterizes the orbit of K in Σ. Let ∇ be the Levi–Civita connection with respect to g.

This situation is rich enough to produce interesting results. In what follows $\{\xi_i\}_{i=1}^n$ will be a g_K orthonormal basis of $\mathfrak{k}$ and X_ξ the left-invariant vector field generated by $\xi \in \mathfrak{k}$. We write $X_i = X_{\xi_i}$, assume that these form an orthonormal frame of $\mathrm{T}K$ and define θ^i to be the coframe dual to the X_i. Let ∇^K be the Levi–Civita connection of K with respect to g_K.

Proposition 12.14. *We have*

$$\Delta^{\nabla^K} X_i = k X_i.$$

Proof. This follows from the identity on K,

$$\nabla^K_{X_i} X_j = \frac{1}{2}[X_i, X_j]\,.$$

□

Thus, Δ^{∇^K} will be the Casimir operator associated with the adjoint representation of K on $\mathfrak{k}$ applied to the frame $\{X_i\}$. The Casimir however is just a constant multiple of the identity, the constant k being

$$k = 1 - \frac{\dim \mathfrak{z}(K)}{\dim \mathfrak{k}}$$

where $\mathfrak{z}(K)$ is the Lie algebra of the center of the group K.

The geometry of warped products is reasonably well known [3, 18]. In particular, we have the following (see pp 58–59 of [3]).

Proposition 12.15. *Let f be a K-invariant function on Σ and suppose $\psi(r) = e^{\lambda(r)}$. The connection Laplacian associated with g satisfies*

$$\Delta^\nabla f \mathrm{d}r = -\left(\frac{\partial^2 f}{\partial r^2} + (n-1)\frac{\partial \lambda}{\partial r}\frac{\partial f}{\partial r} - (n-1)\left(\frac{\partial \lambda}{\partial r}\right)^2 f\right)\mathrm{d}r\,,$$

$$\Delta^\nabla f e^\lambda \theta^i = -\left(\frac{\partial^2 f}{\partial r^2} + (n-1)\frac{\partial \lambda}{\partial r}\frac{\partial f}{\partial r} - \left(\frac{\partial \lambda}{\partial r}\right)^2 f - e^{-\lambda}kf\right)e^\lambda \theta^i\,.$$

These two propositions imply that the K-invariant equations

$$\mathbb{I}a^0(r)\mathrm{d}r = \frac{\delta(r-R)}{\psi(R)}\mathrm{d}r\,, \tag{12.46}$$

$$\mathbb{I}a^i(r)\psi(r)\theta^i = \delta(r-R)\theta^i\,, \tag{12.47}$$

are n ordinary differential equations in r. Thus, there are n functions $G^0(r,R)$, $G^i(r,R)$ that solve (12.46) and (12.47), respectively, are continuous on $r=R$, and are symmetric in the sense that

$$G^0(r,R) = G^0(R,r) \quad \text{and} \quad G^i(r,R) = G^i(R,r).$$

Indeed, the second identity in Proposition 12.15 shows that G^i is *independent* of the index i. Hence, instead of G^i, we shall write G^K.

Now let $\boldsymbol{m}$ be the measure-valued one-form

$$\boldsymbol{m}^j = \sum_{i=1}^N \frac{\delta(r-R_i)}{\psi(R_i)}\left(P_i\mathrm{d}r + M_i\theta^j\right). \tag{12.48}$$

The solution to $\mathbb{I}(\boldsymbol{u}_j)_\flat = \boldsymbol{m}^j$ is

$$\begin{aligned}\boldsymbol{u}_j &= \sum_{i=1}^N \left(P_i G^0(r,R_i)\frac{\partial}{\partial r} + \frac{M_i}{\psi(r)\psi(R_i)}G^K(r,R_i)X_j\right)\\ &=: u_j^0(r)\frac{\partial}{\partial r} + u_j(r)X_j \quad \text{(no sum on } j\text{)}.\end{aligned}$$

Note, we are not using the summation convention here. Also denote

$$w = f^0\frac{\partial}{\partial r} + f^\zeta X_\zeta$$

for any $\zeta \in \mathfrak{k}$ and arbitrary scalar functions of r, denoted f^0 and f^ζ. Any vector field on Σ will be the sum of such vector fields w. Then we may write the vector field commutation relation,

$$[w, \boldsymbol{u}_j] = \left(f^0 u_j^{0\prime} - u_j^0\partial_r f^0 - u_j X_j(f^0)\right)\frac{\partial}{\partial r} - \left(u_j^0\partial_r f^\zeta + u_j X_j(f^\zeta)\right)X_\zeta + f^0 u_j' X_j\,.$$

Thus

$$\begin{aligned}\int_\Sigma \boldsymbol{m}^j(w)\mathrm{d}vol_P &= \sum_{i=1}^N \int_K \int_I \frac{\delta(r-R_i)}{\psi(R_i)} \left(P_i f^0 + M_i f^\zeta g_K(X_\zeta, X_j)\right) \psi(r)\mathrm{d}r\mathrm{d}vol_K \\ &= \sum_{i=1}^N \int_{h\in K} \left(P_i f^0(R_i,h) + M_i f^\zeta(R_i,h)\zeta^j\right) \mathrm{d}vol_K\,,\end{aligned}$$

where $\zeta = \sum_{j=1}^{n-1} \zeta^j \xi_j$. In particular, this yields the ad* relation needed for expressing EPDiff in this setting,

$$\begin{aligned}&\int_\Sigma \boldsymbol{m}^j([w,\boldsymbol{u}_j])\mathrm{d}vol_P \\ &= \sum_{i=1}^N \int_{h\in K} P_i \left(f^0(R_i,h)u_j^{0\prime}(R_i) - u_j^0(R_i)\partial_r f^0(R_i,h) - u_j(R_i)X_j(f^0)(R_i,h)\right) \mathrm{d}vol_K \\ &- \sum_{i=1}^N \int_{h\in K} M_i \left(u_j^0(R_i)\partial_r f^\zeta(R_i,h)\zeta^j + u_j X_j(f^\zeta)(R_i,h)\zeta^j - f^0(R_i,h)u_j'(R_i)\right) \mathrm{d}vol_K\,.\end{aligned}$$

Remark 12.7. When we were dealing in Sect. 12.5 with isometric action of S^1, terms such as $u_\theta \partial_\theta f$ disappeared when integrated over the circle, by the Stokes theorem. Here, however, we have the terms $u_j X_j(f^0)$ and $u_j X_j(f^\zeta)$ (again no sum on j). As we shall see, these terms will vanish when we integrate over the group K.

From standard Riemannian geometry, one has

Proposition 12.16. *For any compact Riemannian manifold N, and any vector field X and function ϕ on N we have*

$$\int_N X(\phi)\mathrm{d}vol = \int_N \phi \,\mathrm{div}\, X \mathrm{d}vol\,.$$

From this formula we see that, in the present situation,

Proposition 12.17. *Given any $\kappa \in \mathfrak{k}$ and any function ϕ on K.*

$$\int_K X_\kappa(\phi)\mathrm{d}vol_K = 0\,.$$

This relation follows because X_κ generates volume preserving (and metric preserving) diffeomorphisms of K.

Hence, upon integrating over K, the terms $u_j X_i(f^0)$ and $u_j X_j(f^\zeta)$ will vanish because u_j is K-invariant. Thus, we have

$$\int_\Sigma \boldsymbol{m}^j([w,\boldsymbol{u}_j])\mathrm{d}vol_P = \sum_{i=1}^N \int_{h\in K} P_i\left(f^0(R_i,h)u_j^{0\prime}(R_i) - u_j^0(R_i)\partial_r f^0(R_i,h)\right)\mathrm{d}vol_K$$
$$- \sum_{i=1}^N \int_{h\in K} M_i\left(u_j^0(R_i)\partial_r f^\zeta(R_i,h)\zeta^j - f^0(R_i,h)u_j'(R_i)\right)\mathrm{d}vol_K$$

This formula implies the EPDiff equations,

$$\begin{aligned}
0 &= \int_\Sigma \left(\frac{\partial \boldsymbol{m}^j}{\partial t}(w) + \boldsymbol{m}^j([w,\boldsymbol{u}_j])\right)\mathrm{d}vol_P \\
&= \sum_{i=1}^N \int_{h\in K} \Big(\dot{P}_i f^0(R_i,h) + P_i\dot{R}_i\frac{\partial}{\partial R_i}f^0(R_i,h) + \dot{M}_i f^\zeta(R_i,h)\zeta^j \\
&\qquad\qquad + M_i\dot{R}_i\frac{\partial}{\partial R_i}f^\zeta(R_i,h)\zeta^j\Big)\mathrm{d}vol_K \\
&+ \sum_{i=1}^N \int_{h\in K} P_i\left(f^0(R_i,h)u_j^{0\prime}(R_i) - u_j^0(R_i)\partial_r f^0(R_i,h)\right)\mathrm{d}vol_K \\
&- \sum_{i=1}^N \int_{h\in K} M_i\left(u_j^0(R_i)\partial_r f^\zeta(R_i,h)\zeta^j - f^0(R_i,h)u_j'(R_i)\right)\mathrm{d}vol_K
\end{aligned}$$

Again, we compare coefficients of $f^0, f^\zeta, \frac{\partial}{\partial R_i}f^0$ and $\frac{\partial}{\partial R_i}f^\zeta$ to find the dynamical equations

$$\begin{aligned}
0 &= \dot{P}_i + P_i u_j^{0\prime}(R_i) + M_i u_j'(R_i), \\
0 &= P_i\dot{R}_i - P_i u_j^0(R_i)\,, \\
0 &= \dot{M}_i\zeta^j\,, \\
0 &= M_i\dot{R}_i\zeta^j - M_i u_j^0(R_i)\zeta^j\,.
\end{aligned}$$

Therefore, we have proved the following.

Theorem 12.2 (Hamiltonian Diffeon Reduction for K-Invariant Warped Product Spaces). *The warped product diffeon parameters in Eq. (12.48) satisfy*

$$\dot{P}_i = -\frac{1}{2}\frac{\partial}{\partial R_i}\sum_{k=1}^N\left(P_k u_j^0(R_k) + M_k u_j(R_k)\right)\,, \tag{12.49}$$

$$\dot{R}_i = u_j^0(R_i)\,, \tag{12.50}$$

$$\dot{M}_i = 0\,. \tag{12.51}$$

As guaranteed by the momentum map property of the diffeon solution (12.8) for EPDiff, these are Hamilton's equations with collectivized Hamiltonian $H_j/(\mathrm{vol}K)$ given by

$$\begin{aligned} H_j(\mathbf{P},\mathbf{R};\mathbf{M}) &= \frac{1}{2}\int_\Sigma \boldsymbol{m}^j(\boldsymbol{u}_j)\mathrm{d}vol_P \\ &= \frac{1}{2}\mathrm{vol}\,(K)\sum_{i=1}^{N}\left(P_i u_j^0(R_i)+M_i u_j(R_i)\right) \\ &= \frac{1}{2}\mathrm{vol}\,(K)\sum_{i=1}^{N}\left(P_i u_j^0(R_i)+M_i u_j(R_i)\right) \\ &= \frac{1}{2}\mathrm{vol}\,(K)\sum_{i,k=1}^{N}\left(P_iP_kG^0(R_i,R_k)+\frac{M_iM_k}{\psi(R_i)\psi(R_k)}G^K(R_i,R_k)\right). \end{aligned}$$

Notice that H_j is independent of j; so we may write H^K for H_j.

Thus, given any $\zeta\in\mathfrak{k}$ a diffeon exists that moves on Σ with motion described by Hamilton's equations with the Hamiltonian given by H. The diffeon itself is "rotating" in the direction determined by ζ with conserved angular momentum determined by $\mathbf{M}=(M_1,\ldots,M_N)$.

Remark 12.8. One may repeat the whole procedure replacing K with a compact symmetric space K/H for some closed Lie subgroup H of K. The only significant changes would be that the θ^i would become local on K/H rather than global. However, all the propositions will remain true because of the intimate relationship between symmetric spaces and Lie groups. Thus, for example, we would expect diffeon behavior on S^n to be similar to the Puckon behavior on S^2 since

$$g_{\mathrm{S}^n}=\mathrm{d}r^2+(\cos^2 r)\,g_{\mathrm{S}^{n-1}}.$$

12.6.2 Singular Fibers

We have so far dealt with a free action of a group on a manifold such that the quotient space is a 1-manifold. Thinking back to the case of $\Sigma=\mathrm{S}^2$, we see that the action of the circle was not completely free, because there are precisely two points (the poles) which are fixed under the group. Away from these fixed points, the sphere is a principal S^1 fibration, and the Puckons "bounce" off the poles (provided they are not rotating). For the general manifold Σ, the situation could become much more complex.

For example, in the situation of the previous section, we see that problems arise if we choose a diffeon which is rotating in the direction $\zeta\in\mathfrak{k}$ and find that X_ζ vanishes

at a point p. However, if the diffeon is not rotating, and the vector field $\frac{\partial}{\partial r}$ vanishes nowhere, then all the previous theory holds. The theory also holds for the warped product of the line (or circle) with the flat metric with an Einstein space, upon using harmonic coordinate charts (page 285 of [18]).

12.7 Conclusions

We provided a canonical Hamiltonian framework for exploring the solutions of the EPDiff equations on surfaces of constant curvature. This framework used symmetry and the momentum map for singular solutions to reduce the EPDiff integro-partial differential equation to canonical Hamiltonian ODEs in time. We specialized to the case of the sphere and provided both numerical integrations and qualitative analysis of the solutions, which we called "Puckons."

The main conclusions from our numerical study were:

- Momentum plays a key role in the dynamics of Puckons. Radial momentum drives Puckons to collapse onto one of the poles, and angular momentum prevents this collapse from occurring. Puckons were found to exhibit elastic collision behavior (with its associated exchanges of momentum and angular momentum, but with no excitation of any internal degrees of freedom) just as occurs in soliton dynamics.
- Puckons without rotation may collapse onto one of the poles. This collapse occurs with bounded canonical momentum and the radial slope in velocity appears to become vertical at the instant of collapse.
- For nonzero rotation, Puckon collapse onto one of the poles cannot occur and the radial slope in velocity never becomes infinite.
- Head-on collisions between two Puckons may be accompanied by an apparently vertical radial slope in velocity which forms in finite time.

The main theoretical questions that remain are:

- Numerical simulations show that near vertical or vertical slope occurs at head-on collision between two Puckons of nearly equal height. A rigorous proof of this fact is still missing.
- It remains to discover whether a choice of Green's function exists for which the reduced motion is integrable on our $2N$ dimensional Hamiltonian manifold of concentric rotating Puckons for $N > 1$.
- It also remains to determine the number and speeds of the rotating Puckons that emerge from a given initial condition.

All of these challenging theoretical problems are beyond the scope of this paper and we will leave them as potential subjects for future work.

We applied these ideas to hyperbolic spaces, as well. This led to rather simple reduced dynamics with only a limited number of possible collisions. We discussed

a new departure for hyperbolic space, based on Teichmüller theory, whose investigation has already begun elsewhere [6, 7, 13].

Finally, we answered an outstanding question by generalizing the momentum map to the case of diffeons with $(n-1)$-dimensional internal degrees of freedom by using the theory of warped product spaces.

In summary, we identified and analyzed cases where imposing an additional translation symmetry on the solution reduced the canonical Hamiltonian dynamics of the singular solutions of EPDiff on Einstein surfaces from (integral) partial differential equations to Hamilton's canonical ordinary differential equations, in time. We extended our methods for surfaces to "mostly symmetric" manifolds in higher dimensions by using warped products.

All calculations were done explicitly, in hopes of encouraging further applications of these ideas.

Acknowledgements DDH is grateful for partial support by an Advanced Grant from the European Research Council. The research of JM was partially supported by an EPSRC postdoctoral fellowship at Imperial College London. SNS was supported by a US Department of Energy Computational Science Graduate Fellowship under grant number DE-FG02-97ER25308. The authors wish to thank Simon Donaldson, John Gibbon, François Gay-Balmaz, Sergey Kushnarev, Tudor Ratiu, and Richard Thomas for their thoughts and advice.

References

1. Ahlfors, L.V.: Lectures on Quasiconformal Mappings. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey (1987). With the assistance of Clifford J. Earle, Jr., Reprint of the 1966 original
2. Arnold, V.I.: Sur la géométrie differentielle des groupes de Lie de dimenson infinie et ses applications à l'hydrodynamique des fluids parfaits. Ann. Inst. Fourier Grenoble **16**, 319–361 (1966)
3. Besse, A.L.: Einstein Manifolds. Springer, Berlin (1987)
4. Camassa, R., Holm, D.D. An integrable shallow water equation with peaked solitons, Phys. Rev. Lett. **71**, 1661–1664 (1993)
5. Gardiner, F.P., Harvey, W.J.: Universal Teichmüller space. In: Handbook of Complex Analysis: Geometric Function Theory, vol. 1, pp. 457–492. North-Holland, Amsterdam (2002)
6. Gay-Balmaz, F.: Infinite dimensional geodesic flows and the universal Teichmüller space. PhD Thesis. Ecole Polytechnique Federale de Lausanne (2008)
7. Gay-Balmaz, F., Marsden, J.E., Ratiu, T.S.: The geometry of the universal Teichmüller space and the Euler-Weil-Petersson equation (2010)
8. Holm, D.D., Marsden, J.E.: Momentum maps and measure-valued solutions (peakons, filaments and sheets) for the EPDiff equation. In: Marsden, J.E., Ratiu, T.S. (eds.) The Breadth of Symplectic and Poisson Geometry. Progr. Math., vol. 232, pp. 203–235. Birkhäuser Boston, Boston (2004). Preprint at arxiv.org/abs/nlin.CD/0312048
9. Holm, D.D., Marsden, J.E., Ratiu, T.S.: The Euler–Poincaré equations and semidirect products with applications to continuum theories. Adv. Math. **137**, 1–81 (1998)
10. Holm, D.D., Putkaradze, V., Stechmann, S.N.: Rotating concentric circular peakons. Nonlinearity **17**, 2163–2186 (2004). Preprint at arxiv.org/abs/nlin.SI/0312012
11. Holm, D.D., Staley, M.F.: Wave structures and nonlinear balances in a family of evolutionary PDEs. SIAM J. Appl. Dyn. Syst. **2**, 323–380 (2003)

12. Holm, D.D., Staley, M.F.: Interaction dynamics of singular wave fronts. Unpublished MS (2004)
13. Kushnarev, S. Teichons: soliton-like geodesics on universal Teichmüller space. Exp. Math. **18**(3), 325–336 (2009)
14. LeVeque, R.J. Numerical Methods for Conservation Laws. Birkhäuser, Boston (1992)
15. Marsden, J.E., Ratiu, T.S.: Introduction to Mechanics and Symmetry, 2nd edn. Texts in Applied Mathematics, vol. 17. Springer, New York (1999)
16. Miller, M.I., Trouvé, A., Younes, L.:, On the metrics and Euler-Lagrange equations of computational anatomy. Annu. Rev. Biomed. Eng. **4**, 375–405 (2002)
17. Mumford, D.: Pattern theory and vision. In: Questions Mathématiques En Traitement Du Signal et de L'Image, Chapter 3, pp. 7–13. Institut Henri Poincaré, Paris (1998)
18. Petersen, P.: Riemannian Geometry. Springer, New York (1998)
19. Wolpert, S.A.: The hyperbolic metric and the geometry of the universal curve. J. Differ. Geom. **31**, 417–472 (1990)

Chapter 13
On the Destruction of Resonant Lagrangean Tori in Hamiltonian Systems

Henk W. Broer, Heinz Hanßmann, and Jiangong You

Abstract Starting from Poincaré's fundamental problem of dynamics, we consider perturbations of integrable Hamiltonian systems in the neighbourhood of resonant Lagrangean (i.e. maximal) invariant tori with a single (internal) resonance. Applying KAM Theory and Singularity Theory we investigate how such a torus disintegrates when the action variables vary in the resonant surface. For open subsets of this surface the resulting lower dimensional tori are either hyperbolic or elliptic. For a better understanding of the dynamics, both qualitatively and quantitatively, we also investigate the singular tori and the way in which they are being unfolded by the action variables. In fact, if N is the number of degrees of freedom, singularities up to co-dimension $N - 1$ cannot be avoided. In the case of Kolmogorov non-degeneracy the singular tori are parabolic, while under the weaker non-degeneracy condition of Rüssmann the lower dimensional tori may also undergo e.g. umbilical bifurcations. We emphasize that this application of Singularity Theory only uses internal (or distinguished) parameters and no external ones.

H.W. Broer (✉)
Johann Bernoulli Instituut, Rijksuniversiteit Groningen, 9747 AG Groningen, The Netherlands
e-mail: H.W.Broer@rug.nl

H. Hanßmann (✉)
Mathematisch Instituut, Universiteit Utrecht, 3508 TA Utrecht, The Netherlands
e-mail: hansmann@math.uu.nl

J. You (✉)
Department of Mathematics, Nanjing University, Nanjing 210093, P.R. China
e-mail: jyou@nju.edu.cn

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_13,

13.1 Introduction

Classical perturbation theory largely concerns the continuation of quasi-periodic motions as these occur in integrable Hamiltonian systems for small non-integrable perturbations. The classical perturbation series here diverge on a resonance subset that densely fills the phase space, leading to the notorious small denominators even when avoiding this dense set. This paper deals with the dynamics within such resonance gaps.

Background. We briefly summarize that Kolmogorov–Arnol'd–Moser (KAM) Theory [2] establishes the persistence of quasi-periodic motions that densely fill Lagrangean tori, meaning that the dimension equals the number N of degrees of freedom. If $\omega \in \mathbb{R}^N$ denotes the frequency vector, the resonances alluded to above are given by

$$\langle k, \omega \rangle \ = \ 0 \ , \quad 0 \neq k \in \mathbb{Z}^N \ . \tag{13.1}$$

KAM Theory excludes such resonances by imposing strongly non-resonant, Diophantine conditions

$$| \langle k, \omega \rangle | \ \geq \ \frac{\gamma}{|k|^\tau} \quad \text{for all } k \in \mathbb{Z}^N \backslash \{0\} \tag{13.2}$$

on the frequency vectors, where $\gamma > 0$, $\tau > N-1$, which guarantee the persistence of many Lagrangean tori in the sense of measure theory. We recall [25] that the persistent N-tori are smoothly parametrized over the nowhere dense union of closed half lines defined by (13.2), in the sense of Whitney; colloquially we speak of a *Cantor family of half lines*.

Aim of the paper. We investigate what happens to the Lagrangean tori of the unperturbed system for which the frequency vector is resonant, i.e. satisfying (13.1) for a single fixed nonzero $k \in \mathbb{Z}^N$ (and its integer multiples), so which is contained in a gap of the "Cantor set" defined by (13.2).

Starting point is the real analytic perturbed N-degree-of-freedom Hamiltonian

$$H_\varepsilon(\varphi, I) \ = \ H_0(I) \ + \ \varepsilon H_1(\varphi, I; \varepsilon) \tag{13.3}$$

defined on $\mathbb{T}^N \times \mathbb{R}^N$ with perturbation parameter $0 < \varepsilon \ll 1$. We concentrate on a single resonance, whence a normal form approximation can be reduced to one degree of freedom. See below for details. The reduced system is defined on the cylinder $\mathbb{T} \times \mathbb{R}$ with Hamiltonian of the form

$$\mathscr{H}(p, q; \mu) \ , \quad (q, p) \in \mathbb{T} \times \mathbb{R}, \ \mu \in \mathbb{R}^{N-1} \ ; \tag{13.4}$$

here $\mu = \mu(I)$ is a (distinguished) parameter varying along the resonance surface

$$\langle k, \omega(I) \rangle \ = \ 0 \ , \quad \text{where} \quad \omega(I) \ := \ DH_0(I) \ .$$

We will show that (13.4) is a general family of Hamiltonian functions and our interest is the study of the dynamics they generate. As in all one-degree-of-freedom systems, this dynamics is largely determined by the configuration of level sets. We recall [2] that the closed curves correspond to librational Lagrangean N-tori, while the critical points correspond to invariant $(N-1)$-tori. The latter govern the global geometry of the dynamics and in the $(N-1)$-parameter family (13.4) of functions their behaviour is determined by Singularity Theory [3, 27].

Within such an $(N-1)$-parameter family of functions one may encounter singularities up to co-dimension $N-1$ in a persistent way. Singularity Theory provides us with versal unfoldings of these. These unfoldings also will appear persistently in our general family (13.4).

The compact-open topology. We like to note that for families of planar functions persistence in the sense of structural stability is a generic property. To formulate this property in a precise way one needs a topology on the space of Hamiltonian functions in one degree of freedom. In the present real analytic setting we use the compact-open topology on holomorphic extensions detailed in [11].

The real-analytic compact-open topology on holomorphic extensions fits with local uniform convergence of the corresponding Hamiltonian functions. We recall from [11] that this compact-open topology has the Baire property, which means that countable intersections of dense-open sets are still dense. Moreover, the compact-open topology is stronger than the Whitney C^k topologies used in [22,23]. From the latter it immediately follows that any C^k open property also is open in the compact-open sense. The same holds for denseness, as long as we restrict to properties defined in terms of transversality. Therefore a real analytic unfolding that is versal in the C^k sense also is versal with respect to the compact-open topology. In particular the differentiable Singularity Theory also applies for the real analytic case. In fact, Weierstraß originally developed the theory for the analytic case.

Co-dimensions. We recall that for equilibria to be either hyperbolic or elliptic, the corresponding singularity A_1 has co-dimension 0, which means that in one degree of freedom it has open occurrence in the space of families of Hamiltonian functions. However, the singularities of higher co-dimension determine the geometric organization of those of lower co-dimension, therefore in particular the ones corresponding to hyperbolic and elliptic equilibria. This organization reflects both qualitative and quantitative aspects.

Example 13.1 (A Quasi-Periodic Center-Saddle Bifurcation). As an example[1] in the space of $(N-1)$-parameter families in one degree of freedom we consider

$$\mathscr{H}(p,q;\mu) \;=\; \frac{1}{2}p^2 \;+\; \frac{1}{6}q^3 \;+\; \lambda(\mu)q \tag{13.5}$$

[1] Any non-degenerate example of a co-dimension 1 bifurcation can be reduced to this case. Here we exclude a setting with symmetry or other structural restrictions [20].

with $\lambda(0) = 0$ and $D_\mu\lambda(0) \neq 0$. We remark that this family usually is called *fold*, a versal unfolding of the singularity A_2, see [3, 27]. Clearly (13.5) is a family of one-degree-of-freedom systems for which the equilibrium $(p, q) = (0, 0)$ at $\mu = 0$ is parabolic (and hence neither elliptic nor hyperbolic), and for any (real analytic) small perturbation there is a parameter value $\mu = \mu_0 \approx 0$ for which there is a parabolic equilibrium at a certain point $(p, q) = (p_0, q_0) \approx (0, 0)$. Thus, there is a full neighbourhood of the $(N - 1)$-parameter family defined by (13.5) in the compact-open topology such that a parabolic equilibrium occurs nearby. Hence it is impossible for the set of all families of one-degree-of-freedom systems that only have elliptic and/or hyperbolic equilibria to contain a countable intersection of open and dense sets, i.e. to be generic.

In the reconstruction to the setting of N degrees of freedom, the equilibria correspond to invariant $(N - 1)$-tori, where the normal behaviour is inherited, and the closed level curves correspond to librational Lagrangean N-tori. Addition of the non-symmetric higher order terms presents us with a new perturbation problem. This can be solved by KAM Theory [18]. In this way the entire bifurcation scenario becomes *Cantorized* as explained below, for a detailed example also see Sect. 13.2.

An interesting aspect is that in a fixed energy level we have the same theory, with one parameter less. For $N \geq 3$ we thus have in each energy level elliptic, hyperbolic, and parabolic tori in a center-saddle bifurcation. In both cases a quantitative consequence of this is that for small perturbations the asymptotic distance of the two families of tori, as they approach the parabolic equilibrium, is of order $\sqrt{-\lambda(\mu)}$ as $\mu \to 0$.

Observe that for $N = 2$ the lower dimensional tori really are closed orbits. In this case the energy is the only parameter and no bifurcation takes place inside an energy level. For $N \geq 3$ the bifurcating tori have dimension larger than or equal to 2 and the bifurcation can take place within an energy level.

Unfoldings. The fold example (13.5) is a special case of the family of cuspoids that unfold the co-rank 1 singularities A_{k+1}, $k \in \mathbb{N}$, see [3, 27] for more details. The corresponding KAM perturbation problem has been solved in [5]. In that paper (trans)-versality conditions as dictated by Singularity Theory [3, 27] are used to develop a normal form of Hamiltonians in the neighbourhood of a parabolic torus. Solving the ensuing small divisor problems it is proven that the bifurcation scenario persists in a Cantorized way. See Example 13.2 below for a more detailed description. In this way all possible quasi-periodic bifurcations of normally parabolic tori can be retrieved in resonance gaps, taking N sufficiently high.

An aspect that one has to keep in mind in the present case of $(N - 1)$-parameter families (13.4) defined by the normal form of (13.3) near a single resonance is that part of the terms in $\mathscr{H}$, say the p-terms, come from H_0 and the remaining terms come from the perturbation H_1. The p^2-term in (13.5) occurs in the case where the integrable part H_0 satisfies the Kolmogorov condition

$$\det D^2 H_0(I) \quad \neq \quad 0 \ , \tag{13.6}$$

which is valid on an open and dense set in parameter space, see [2]. For these values of I the above theory of parabolic tori applies, see Sect. 13.2 for more details. In particular, one can read off from the degeneracy of the minimum and the maximum of a certain potential V_μ on $\mathbb{T}$ a lower bound for the number of families of invariant $(N-1)$-tori.

Note that the Kolmogorov non-degeneracy condition (13.6) fails on a co-dimension 1 subset. Here one may still expect a weaker form of non-degeneracy to hold true, expressed by the Rüssmann condition, cf. [8, 26, 29]. In this way also more general singularities can be incorporated, giving rise to all possible applications of Singularity Theory as in [20]. More degenerate singularities may e.g. lead to umbilical torus bifurcations, for an example see Sect. 13.3; in Sect. 13.4 we discuss further possibilities.

We emphasize that we are not to impose genericity conditions on occurring equilibria, but only on the initial Hamiltonian H, see (13.3). Our genericity assumption on H_0 amounts to Rüssmann non-degeneracy. The remaining genericity assumptions on H are obtained via H_1. These genericity conditions amount to versality of the occurring unfolding, thereby excluding too pathological examples. For more details see below.

On Cantorization. The Diophantine condition (13.2) defines a Cantor family of closed half lines, parametrized over a Cantor set of positive measure [9]. This Cantor family in turn parametrizes a Cantor bundle of integrable quasi-periodic invariant tori, in a Whitney smooth way. KAM Theory implies that in case of Kolmogorov non-degeneracy, under small perturbations this Cantor bundle is distorted by a near-identity Whitney smooth conjugation. Where the integrable invariant tori foliate a submanifold or a semi-algebraic set organized by Singularity Theory, we colloquially say that the Diophantine condition (13.2) "Cantorizes" this geometry, with the same terminology for nearly integrable systems. We note that the property of having positive (Hausdorff) measure is preserved by diffeomorphisms. In the sequel we shall also meet Cantor bundles of tori parametrized over (real) Cantor sets. In all cases these bundles can be distinguished by their Hausdorff dimension.

Related work. To our knowledge, invariant tori reconstructed from possibly degenerate equilibria have only been addressed for single resonances, not for multiple resonances.

Cheng [13] considers convex H_0, such that in the reduced system (13.4) the maximum and minimum of the q-dependent part differ. The invariant $(N-1)$-tori corresponding to non-degenerate maxima are hyperbolic tori, but in [13] a degenerate maximum is *not* excluded and the resulting tori are called of "hyperbolic type". It is established that the system (13.3) has a Cantor family of "hyperbolic type" invariant $(N-1)$-tori. In [14] the minima of the q-dependent part of (13.4) are treated, restricting to non-degenerate minima (i.e. elliptic tori). This yields a Cantor family of elliptic invariant $(N-1)$-tori.

The approach by Gallavotti, Gentile, and Giuliani [17] considers the perturbation parameter ε also as a bifurcation parameter. Although they do consider degenerate

singularities, the results only concern Cantor families of elliptic and hyperbolic invariant $(N-1)$-tori. Careful considerations yield expansions of the ε-family of $(N-1)$-tori in e.g. $\sqrt{\varepsilon}$, and quantitative asymptotic information. We note that these results can be retrieved as a direct consequence of our approach. However, the fate of degenerate tori is not explained in [17] and *a fortiori* a complete bifurcation scenario is not discussed.

As soon as one makes the assumption that all equilibria of the reduced system are non-degenerate one obtains elliptic and hyperbolic lower dimensional tori also for multiple resonances, see [10, 12, 15, 28] and references therein. We like to emphasize that this assumption is not generic for reduced systems parametrized by the conjugate actions; one then also has to deal with degenerate equilibria and the corresponding bifurcation theory. A starting point for the development of such a persistence result is formed by [4, 7], where multiple resonances are taken into account. For m-fold resonances corank-m-singularities may occur already under the Kolmogorov condition, and under the Rüssmann condition this may further rise to corank $2m$.

13.2 Kolmogorov Hamiltonians

To explain the kind of results that can be obtained for the perturbation (13.3) of an integrable Hamiltonian $H_0 = H_0(I)$ on $\mathbb{T}^N \times \mathbb{R}^N$ we first restrict to an open subset $U \subset \mathbb{R}^N$ where the Kolmogorov non-degeneracy condition (13.6) is valid for every $I \in U$. By classical KAM Theory [2, 8, 25] it then follows that most Lagrangean tori $\mathbb{T}^N \times \{I\}$, $I \in U$ satisfy the Diophantine conditions (13.2) and persist. For Lagrangean tori with resonant frequency vector that have a single resonance we have the following result.

Theorem 13.1 (Resonant Dynamics). *Consider the perturbed Hamiltonian* (13.3) *on* $\mathbb{T}^N \times U$ *satisfying the Kolmogorov non-degeneracy condition* (13.6) *for all* $I \in U$. *Then for sufficiently small perturbations* H_1, *satisfying suitable genericity conditions, a Lagrangean torus of the unperturbed system with a single resonance* $\langle k, DH_0(I)\rangle = 0$, $k \in \mathbb{Z}^N \backslash \{0\}$ *leads in the perturbed system* (13.3) *to Cantor families of hyperbolic, elliptic, and possibly also parabolic tori. The distribution of these tori is determined by the way in which the genericity conditions on* H_1 *are fulfilled.*

The proof in particular reveals the nature of the genericity conditions, made precise in Lemma 13.1 and the paragraph preceding it.

Proof. The equation $\langle k, DH_0(I)\rangle = 0$ determines a local hypersurface $\mathbb{Y} \subset U$. For $I \in \mathbb{Y}$ the unperturbed Lagrangean torus $\mathbb{T}^N \times \{I\}$ is foliated into invariant tori of dimension $N-1$. Let us put

$$n \;=\; N-1$$

and on U choose local coordinates $\varphi = (x, q)$ with values in $\mathbb{T}^n \times \mathbb{T}$ and $I = (y, p)$ with values in $\mathbb{R}^n \times \mathbb{R}$, where y parametrizes the surface $\mathbb{Y}$. In these local coordinates the single resonance reads

$$\frac{\partial}{\partial p} H_0(y, 0) \;=\; 0 \quad \text{for all } y \in \mathbb{Y} \;, \tag{13.7}$$

cf. [10, 13, 14, 28]. The remaining frequencies form the vector $\omega = D_y H_0(y, 0)$ which is non-resonant at a single resonance $(y, p) = (y^*, 0)$. Treating ω as an external parameter, we expect persistence results only for Diophantine ω, where gaps in the resulting "Cantor set" correspond to multiple resonances. Following [9, 25], we localize to $\hat{y}$ (ε-close to 0) writing $y = y^* + \hat{y}$, thereby shrinking the neighbourhood U if necessary. Moreover we restrict to the lowest order terms

$$H_0(\hat{y}, p; y^*, \omega) \;=\; \langle \omega, \hat{y} \rangle \;+\; \frac{a(y^*)}{2} p^2 \;.$$

From (13.6) together with (13.7) we infer

$$a(y^*) \;\neq\; 0 \quad \text{for all } y^* \in \mathbb{Y} \;,$$

whence we find a lower bound of the function $|a|$ by shrinking the coordinate domain $U \supset \mathbb{Y}$ a bit if necessary.

We now apply a familiar method to replace the system (13.3) by a family of one-degree-of-freedom systems, cf. [2, 10, 13, 28]. Starting point is a normalizing transformation that turns the perturbed Hamiltonian into

$$H_\varepsilon(x, \hat{y}, p, q; y^*, \omega) \;=\; H_0(\hat{y}, p; y^*, \omega) \;+\; \varepsilon \bar{H}_1(\hat{y}, p, q; y^*, \omega) \;+\; \mathscr{O}(\varepsilon^2)$$

where $\bar{H}_1$ is the $\mathbb{T}^n$-average along x of H_1 at $\varepsilon = 0$. In the expansion

$$\begin{aligned} \bar{H}_1 \;=\; & \eta(\hat{y}; y^*, \omega) \;+\; \alpha(\hat{y}; y^*, \omega) p \;+\; \beta(\hat{y}; y^*, \omega) q \\ & +\; \frac{A(y^*)}{2} p^2 \;+\; \frac{B(y^*)}{2} q^2 \;+\; C(y^*) pq \;+\; \ldots \end{aligned}$$

we may have $A(y^*) \equiv 0$, but more importantly $|a(y^*) + \varepsilon A(y^*)|$ is still bounded from below on $\mathbb{Y}$. Re-parametrizing $\omega \mapsto \omega + \varepsilon D_{\hat{y}} \eta(\hat{y}; y^*, \omega)$, maintaining the same symbol ω for the frequency vector, the expansion of H_ε still starts with $\langle \omega, \hat{y} \rangle$. By means of an ε-small shear transformation in p we get rid of terms that are linear in p, and scaling p by $\sqrt{\varepsilon}$ we arrive at

$$H_\varepsilon(x, \hat{y}, p, q; y^*, \omega) \;=\; \langle \omega, \hat{y} \rangle \;+\; \varepsilon \mathscr{H}(p, q; y^*, \omega) \;+\; \mathscr{O}(\varepsilon^2) \tag{13.8}$$

with

$$\mathscr{H}(p,q;y^*,\omega) \;=\; \frac{a(y^*)}{2}p^2 \;+\; V_{y^*}(q)\;,$$

compare with (13.4).

Here V_{y^*} can be interpreted as an n-parameter family of one-dimensional potentials, and critical points q^* of V_{y^*} correspond to invariant n-tori $\mathbb{T}^n \times \{(0,0,q^*;y^*,\omega)\}$ of the "intermediate" integrable system with Hamiltonian $\mathscr{H}_\varepsilon = \langle\omega,\hat{y}\rangle + \varepsilon\mathscr{H}$. The $y^* \in \mathbb{R}^n \cong \mathbb{Y}$ around which $\hat{y}$ is localized has the character of a (distinguished) parameter; let us make this explicit by writing

$$\mu \;:=\; y^*\;.$$

While critical points of a single potential are generically non-degenerate, it is a generic property for the n-parameter family V_μ of potentials to encounter critical points up to co-dimension n. Note that this amounts to a genericity condition on the perturbation H_1 of H_0. More precisely, the μ-values parametrizing a potential V_μ with a degenerate critical point of co-dimension k form an $(n-k)$-dimensional submanifold Λ_k in μ-space.

Lemma 13.1 (Versality). *In the above circumstances, let $\mu^* \in \Lambda_k$ and put $d = k+2$. Then all derivatives at q^* of order $j < d$ vanish, so*

$$V_{\mu^*}(q) \;=\; \frac{b(\mu^*)}{d!}(q-q^*)^d \;+\; \mathscr{O}\left((q-q^*)^{d+1}\right) \tag{13.9}$$

with $b(\mu^*) \neq 0$ and $2 \leq d \leq n+2$ near the critical point q^* of V_{μ^*}. When $d=2$ the critical point q^* is non-degenerate and $(p,q) = (0,q^*)$ is a non-degenerate equilibrium of the one-degree-of-freedom system—a saddle if $ab<0$ and a center if $ab>0$. In case $d \geq 3$ the equilibrium is parabolic and it is furthermore generic for the family V_μ to provide a versal unfolding of the degenerate critical point. Also this genericity condition is a condition on the perturbation H_1.

Let us translate coordinates to $q^* = 0$ and concentrate on a degenerate critical point with $d \leq n+1$. Then

$$\mathscr{H}(p,q;\mu,\omega) \;=\; \frac{a(\mu)}{2}p^2 \;+\; \frac{b(\mu)}{d!}q^d \;+\; \sum_{j=1}^{d-2}\frac{c_j(\mu)}{j!}q^j \;+\; \mathscr{O}(q^{d+1})$$

and $\mathscr{H}_\varepsilon = \langle\omega,\hat{y}\rangle + \varepsilon\mathscr{H}$ has the form (1.4) of [5] with $\lambda_j = \varepsilon c_j(\mu)$. Thus under small perturbation, satisfying the above genericity conditions, the resonant Lagrangean torus leads to the entire bifurcation scenario detailed in [5]. This amounts to the classical hierarchy of the cuspoids [3, 27] unfolding the singularities A_{k+1}, $k \in \mathbb{N}$, that are Cantorized in the now familiar way by taking out a small

neighbourhood of the dense set of resonances mentioned before and using KAM Theory. Granted the proof of Lemma 13.1 (given below), this proves Theorem 13.1. □

Remark 13.1. Since $\mu = y - \hat{y}$ was introduced by localization the rôle of the unfolding parameter λ is ultimately played by the action y conjugate to the toral angles, provided that $c : \mu \mapsto c(\mu)$ is at $\mu = 0$ a submersion from $\mathbb{R}^n$ to $\mathbb{R}^{d-2}$. The tori $\mathbb{T}^n \times \{(0, 0, q^*)\}$ with $d = n + 2$ are generically isolated and may therefore disappear in a resonance gap. In case $n - d + 2 \geq 1$ Diophantine approximation of dependent quantities does yield persistence on Cantor sets, see [5,8,20] for more details.

Proof of Lemma 13.1. For a description of the compact-open topology on holomorphic extensions of real analytic systems we refer to the introduction. We recall that Singularity Theory in the real analytic setting coincides with that in the C^k setting for k large. Therefore the families V_μ are also versal in the real analytic setting.

The versality of the family V_μ of potentials amounts to a genericity condition on the same family. This in turn is implied by a genericity condition on H_1 since the mapping $H_1 \mapsto \bar{H}_1 \mapsto V_\mu$ is a composition of submersions. Indeed, the latter part $\bar{H}_1 \mapsto V_\mu$ is merely a scaling in p after which the quadratic part $\frac{a}{2}p^2$ is split off by means of the Morse Lemma [22]. The normalization $H_1 \mapsto \bar{H}_1$ consists of a coordinate transformation followed by truncation of higher order terms which for sufficiently small ε has the character of a (linear) projection. □

Example 13.2 (Unfolding a Degenerate Minimum). We consider the case $d = 4$ for which the one-degree-of freedom family has the form

$$\mathscr{H}(p, q; \mu, \omega) \;=\; \frac{1}{2}p^2 \;+\; \frac{1}{24}q^4 \;+\; \lambda_1(\mu)q \;+\; \frac{\lambda_2(\mu)}{2}q^2 \;, \tag{13.10}$$

versally unfolding the singularity A_3. Note that this example occurs persistently for $N \geq 4$ degrees of freedom. For definiteness we fix $N = 4$. So we started with four action variables $\mathbb{R}^4 = \{I_1, \ldots, I_4\}$ in which the three-dimensional resonance hypersurface $\mathbb{Y}$ is defined by $\langle k, DH_0(I)\rangle = 0$. Locally we have the variable p transverse to $\mathbb{Y}$ and $\mu = (\mu_1, \mu_2, \mu_3)$ parametrizes $\mathbb{Y}$.

In Fig. 13.1 we show the organization of the local dynamics in dependence of the parameters. We now describe the meaning of the phase portraits for the 4-degree-of-freedom system. All periodic orbits correspond to Lagrangean tori and equilibria to three-dimensional tori. These are elliptic in the case of a center, hyperbolic in the case of a saddle and parabolic in the (two) remaining cases.

We recall that $\mu \mapsto \lambda(\mu)$ is a (local) submersion. Therefore Cantorization in the (μ_1, μ_2, μ_3)-direction amounts to the following.

1. We begin with the cases corresponding to parabolic tori.
   (a) The central point $\lambda = 0$ corresponds to a line that Cantorizes to a (real) Cantor set (i.e. of topological dimension 0) of Hausdorff dimension 1.

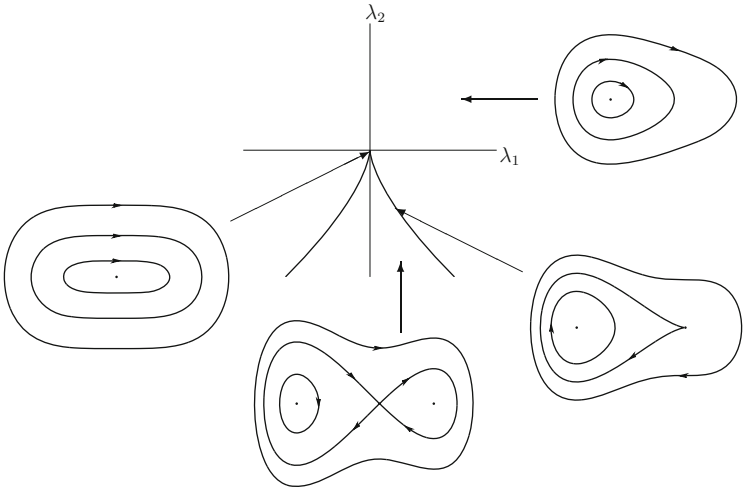

Fig. 13.1 Organization of the local dynamics near a degenerate minimum of the reduced Hamiltonian $\mathscr{H}$. Phase portraits show the reduced one-degree-of-freedom dynamics of (13.10). The interpretation for the N-degree-of-freedom system is given in the text

(b) The fold lines emanating from $\lambda = 0$ correspond to planes that Cantorize to a (real) Cantor set as well, now of Hausdorff dimension 2.

2. The hyperbolic tori are parametrized over the open region in between the fold lines, which corresponds to an open three-dimensional set. Cantorization leads to a union of lines, smoothly parametrized over the two-dimensional Cantor set mentioned above and also ending there. Recall that colloquially we call this a Cantor family of closed half lines; the Hausdorff dimension is 3.
3. Both open regions in Fig. 13.1 parametrize elliptic tori, observing that in between the fold lines each point corresponds to two elliptic tori, and in the other region only to one.

   (a) In the latter case Cantorization leads to a (real) Cantor set.
   (b) In the former case Cantorization leads to two layers of (real) Cantor sets.

   In both cases the Cantor set has Hausdorff dimension 3.

As usual, Cantorization of the librational Lagrangean tori takes place along a Cantor family of lines, of Hausdorff dimension 4. Note that this also uses the p variable to obtain the four-dimensional (μ, p)-set. Here, and in the earlier cases, the corresponding Hausdorff measure is positive and, in fact, even close to full measure when the perturbation is small.

As in the example of the quasi-periodic center-saddle bifurcation, we can give asymptotic estimates based on the geometry sketched in Fig. 13.1. For instance, along the symmetry axis $\lambda_1 = 0$ the distance between the two elliptic tori is of order $\sqrt{-\lambda_2(\mu)}$ as $\mu \to 0$.

It is instructive to compare the above results with the corresponding statements in [17]. When d is odd the bifurcation diagram contains empty regions and regions with both elliptic and hyperbolic tori. When d is even all open regions of the bifurcation diagram Cantor-parametrize at least one elliptic torus if $b > 0$, see Fig. 13.1, while $b < 0$ yields deformations with hyperbolic tori. For a similar approach in the latter case (not using Lindstedt series) see [30].

For small perturbations εH_1 of H_0 Theorem 13.1 allows to recover what [13,14] state in this situation. Indeed, the angular variable q takes values in $\mathbb{T}$ and on this compact set each potential V_μ, μ fixed, assumes minimum and maximum. It is a genericity condition on H_1 for these to be different from each other, i.e. $V_\mu \neq const$ for all μ, and furthermore for V_μ to assume the form (13.9) with d even and $b(\mu) < 0$ at a maximum q^* while $b(\mu) > 0$ at a minimum; let us denote the latter by q_*. For definiteness we concentrate on $a(\mu) > 0$ (which may always be achieved by reversing time if necessary). Then the maximum q^* corresponds to a "hyperbolic type" torus $\mathbb{T}^n \times \{(0,0,q^*)\}$, truly hyperbolic if the maximum is non-degenerate and parabolic otherwise. In both cases we have persistence for ω satisfying the Diophantine conditions (13.2) with N replaced by n. This leads to Cantorization.

The persistence of at least one n-torus from the H_0-resonant $\mathbb{T}^n \times \{(0,0)\} \times \mathbb{T}$ for Diophantine ω had already been established in [13], and without any genericity condition on the perturbation. What our approach adds to this is a precise description how occurring degenerate maxima of the potential, called "of weaker persistency" in [13], lead to a Cantorized bifurcation scenario of the corresponding n-tori. This latter result cannot be obtained without genericity conditions. The generality of the result in [13] does not exclude perturbations that are rather pathological. Correspondingly, that approach does not allow to obtain information on the fine structure where the n-tori fail to be truly hyperbolic.

The minima q_* of V_μ are treated in [14]. In the non-degenerate case these correspond to elliptic n-tori, whence normal-internal resonances have to be avoided as well. Therefore, persistence of a second n-torus is obtained in [14] only on a smaller (though still measure-theoretically large) subset S. For generic perturbations we can now explain the fine structure. In particular the tori coming from degenerate minima q_* do not have to be excluded. In fact, these give the opportunity to enlarge S a bit. For instance, when $d = 4$ in (13.9) we recover the bifurcation diagram given in Fig. 13.1 and next to at least one center for all nonzero $\lambda \in \mathbb{R}^2$ we have an additional saddle in between the two fold lines; hence, here only Diophanticity of internal frequencies is needed to obtain persistence of a second family of invariant n-tori in the resonant zone.

Remark 13.2. Without the genericity conditions on the perturbation there may be tori $\mathbb{T}^n \times \{(0,0,q^*)\}$ with $d > n+2$. In an attempt to still apply the results of [5] we may introduce extra (external) parameters ν to provide a versal unfolding. The perturbed system then displays again the Cantorized bifurcation scenario and contains the original perturbed system as a subsystem. The torus $\mathbb{T}^n \times \{(0,0,q^*)\}$ is an invariant torus of the intermediate system. However, we cannot expect this torus to be present in the perturbed system since this torus is moved by the perturbation.

We still have the weaker conclusion, though, that the Cantorized family of n-tori contains parabolic tori of at most degeneracy d (as found in the intermediate system).

Note that this approach does not allow us to drop the genericity conditions that we had to impose on the perturbation when recovering the results of [13,14]. Indeed, the small constant ε in Theorem 2.1 of [5] depends on d and may tend to 0 as $d \to \infty$. Since the results in [13, 14] are valid for non-generic perturbations as well one may speculate that the latter does not occur.

Example 13.3 (A Non-versal Perturbation). The class of perturbed Hamiltonians

$$H_\varepsilon(\varphi, I) \;=\; H_0(I) \;+\; \varepsilon H_1(\varphi) \tag{13.11}$$

figures a perturbation that is independent of the action variable I. A single resonance (13.7) leads again to a reduced system (13.4) of the form

$$\mathscr{H}(p, q; \mu) \;=\; \frac{a(\mu)}{2} p^2 \;+\; V(q)$$

where the potential V is now equal to the $\mathbb{T}^n$-average $\bar{H}_1$ of H_1, with no further transformations. As a μ-dependent family this potential is trivial, being the same for all parameter values. Invariant n-tori correspond to $(p, q) = (0, q^*)$ with $V'(q^*) = 0$ and if $b := V''(q^*) \neq 0$ the triviality of the family is not problematic. Indeed, the torus is elliptic for $ab > 0$ and hyperbolic for $ab < 0$ with no need for an unfolding. On the other hand, if $b = 0$ then (13.11) is a very degenerate system: the torus is parabolic and the potential cannot provide the necessary unfolding. Remark 13.2 still applies, though.

The degeneracy in Example 13.3 that occurs for $b = 0$ should then be seen as a warning sign that the model (13.11) is problematic and might need to be changed. This kind of warning sign is given whenever Theorem 13.1 does not apply, cf. [1, 27]. The necessary genericity conditions can be explicitly checked in examples and provide one with clues of what exactly is happening. For instance, where the n-parameter family V_μ of potentials encounters critical points of co-dimension exceeding n the perturbed system (13.4) deserves further examination.

This might result in an adjusted model for what one is trying to describe. Another possible outcome is that a symmetry is found in (13.4), and that within the "symmetric universe" the co-dimension no longer exceeds n. The unfolding provided by V_μ then is expected to be versal within the "symmetric universe", see [24], and also the $\mathscr{O}(\varepsilon^2)$-terms in (13.8) do not break the symmetry. Also other reasons for seemingly non-generic behaviour are known to exist, see e.g. [21] for the persistent occurrence of a degenerate bifurcation.

A different approach (that avoids to impose genericity conditions) is pursued in [16], where periodic orbits foliating invariant 2-tori are searched for using the zeroes of the subharmonic Mel'nikov function. In case the latter vanishes

identically, a second order Mel'nikov function is defined with similar properties, and so on. If all higher order Melnikov functions vanish identically, then the whole torus consisting of periodic orbits is shown to survive the perturbation.

13.3 An Umbilic Example

To understand how results similar to those of the previous section can be obtained if the Kolmogorov condition (13.6) is replaced by Rüssmann's non-degeneracy condition, we now consider the following example in $N = 5$ degrees of freedom. Starting point remains the perturbed Hamiltonian (13.3) with $I = (y, p)$, and for the unperturbed part we work with

$$H_0(y,p) \;=\; \sum_{i=1}^{4} \mathrm{e}^{i-1} y_i + \frac{1}{2} y_i^2 + \frac{1}{6} y_i^3 \;+\; \frac{1}{6} p^3 \;, \tag{13.12}$$

where we use that the vector $(1, \mathrm{e}, \mathrm{e}^2, \mathrm{e}^3)$ is Diophantine. Then Rüssmann's non-degeneracy condition

$$\mathbb{R}^5 \;=\; < \frac{\partial^{|\ell|}\omega}{\partial I^\ell} \;\Big|\; 0 \neq \ell \in \mathbb{N}_0^5 > \;\supseteq\; < \frac{\partial \omega}{\partial y_1}, \frac{\partial \omega}{\partial y_2}, \frac{\partial \omega}{\partial y_3}, \frac{\partial \omega}{\partial y_4}, \frac{\partial^2 \omega}{\partial p^2} >$$

that the partial derivatives span the frequency space is satisfied everywhere. We are interested in the fate of the resonant tori $p = 0$. Again we apply a normalizing transformation that turns the perturbed Hamiltonian $H_\varepsilon = H_0 + \varepsilon H_1$ into

$$H_\varepsilon(x, \hat{y}, p, q; \mu, \omega) \;=\; H_0(\hat{y}, p; \mu, \omega) \;+\; \varepsilon \bar{H}_1(\hat{y}, p, q; \mu, \omega) \;+\; \mathscr{O}(\varepsilon^2)$$

where $\bar{H}_1$ is the $\mathbb{T}^4$-average along x of H_1 at $\varepsilon = 0$.

A vanishing 1-jet (in the (p, q)-variables) of a Hamiltonian function merely amounts to $(p, q) = (0, 0)$ being a relative equilibrium. This is already true for H_0, and to achieve this for H_ε we (again) translate coordinates to $q^* = 0$. Using the p^3-term in H_0 a translation in p allows to remove the p^2-term in the expansion of $\bar{H}_1$ in p and q. In case the coefficient $a_{02}(0; \mu, \omega)$ of q^2 does not vanish at $\mu = 0$ we scale p by $\sqrt{\varepsilon}$ and q by $\sqrt[4]{\varepsilon}$ to recover the quasi-periodic center-saddle bifurcation encountered in the previous section. Finally, for nonzero $a_{11}(0; 0, \omega) pq$ we scale p and q both by ε, revealing the relative equilibrium to be hyperbolic. In the expansion

$$\bar{H}_1 \;=\; \sum_{k+l=3} \frac{a_{kl}(\hat{y}; \mu, \omega)}{k!\, l!} p^k q^l \;+\; \textit{h.o.t.}$$

we therefore start with third order terms. We emphasize that for generic perturbations H_1 this cannot be avoided to occur at 1-parameter subfamilies.

Shrinking $\mathbb{Y}$ a bit, if necessary, the coefficient[2] a_{03} is bounded away from zero. Scaling p by $\varepsilon^{\frac{2}{3}}$ and q by $\varepsilon^{\frac{1}{3}}$ we obtain

$$H_\varepsilon(x,\hat{y},p,q;\mu,\omega) \;=\; \langle\omega,\hat{y}\rangle \;+\; \varepsilon^2\mathscr{H}(p,q;\mu,\omega) \;+\; \mathscr{O}(\varepsilon^{\frac{7}{3}})$$

with

$$\mathscr{H}(p,q;\mu,\omega) \;=\; \frac{a}{6}p^3 \;+\; \frac{b}{6}q^3 \;+\; c_1(\mu)q \;+\; c_2(\mu)p \;+\; c_3(\mu)pq$$

where $a \approx 1$, $b = 6a_{03}(0;\mu,\omega)$. The coefficient functions c_1, c_2 and c_3 vanish at $\mu = 0$, this allows to rescale a_{01} by $\varepsilon^{\frac{2}{3}}$ to yield $c_1(\mu)$ and a_{10} by $\varepsilon^{\frac{1}{3}}$ to yield $c_2(\mu)$; a_{11} is not rescaled, we simply put $c_3 = a_{11}(0;\mu,\omega)$. Note that we do not obtain the unfolding of D_4 used in [6, 20], but a form adapted to the critical singularity $\frac{a}{6}p^3 + \frac{b}{6}q^3$, and that this form leads to the hyperbolic umbilic. Still, we conclude that the family of resonant Lagrangean tori $p = 0$ of (13.12) may lead to umbilical torus bifurcations of invariant 4-tori.

13.4 Rüssmann Hamiltonians

The example of the previous section looks in one aspect quite degenerate—the resonance $p = 0$ coincides with the hypersurface $p = 0$ where the Kolmogorov condition (13.6) fails. In general the (transverse) intersection of these should be a co-dimension 2 submanifold, and singularities of the equation $\det D^2H_0 = 0$ may lead to further complications. For instance, one could consider the example in the previous section with one more degree of freedom and add[3] the terms

$$\mathrm{e}^4y_5 \;+\; \frac{1}{2}y_5^2 \;+\; \frac{1}{6}y_5^3 \;+\; \frac{1}{2}y_5p^2$$

to the unperturbed Hamiltonian (13.12). Then the analysis of the previous section concerns $y_5 = 0$ and the obvious question is how the perturbed system behaves when unfolded by y_5—scaled by $\varepsilon^{\frac{2}{3}}$. Note that it is generic for H_0 to satisfy some form of Rüssmann non-degeneracy at every point, cf. [29].

Leaving such complications aside for the moment, a "naive" generalization of the example in the previous section leads to

[2] Here we depart from the general theory of planar singularities that allows to transparently treat the relative equilibria. Indeed, the coordinates p and q already have a "meaning", so we had to sharpen the usual assumption that the (homogeneous) 3-jet does not have multiple roots.

[3] Here we use that the vector $(1, \mathrm{e}, \ldots, \mathrm{e}^4)$ is Diophantine as well.

$$H_0(\hat{y}, p; \mu, \omega) \;=\; \langle \omega, \hat{y} \rangle \;+\; \frac{a(\mu)}{\ell!} p^\ell$$

with $a(\mu) \neq 0$ and $2 \leq \ell \leq n+2$. For instance, if $\ell = 3$ the potential (13.9) leads to the quasi-periodic center-saddle bifurcation (unfolding the singularity A_2) when $d = 2$. For $d = 3$ this similarly leads to umbilic tori (unfolding the singularity D_4) and to the simple singularities E_6 and E_8 (see [19, 20]) when $d = 4$ and 5, respectively.

The umbilic example of the previous section gives some confidence that it should still be possible to find adapted scalings that turn $\frac{a}{\ell!}p^\ell + \varepsilon \bar{H}_1$ for generic perturbation H_1 into versal unfoldings of occurring singularities. Note, however, that the special "starting point" $\frac{a}{\ell!}p^\ell$ leads to a classification that may slightly differ from the classification of one-degree-of-freedom equilibria by means of planar singularities.

13.5 Conclusions

Summarizing we may state that for a *single* resonance the Kolmogorov condition (13.6) restricts the occurring singularities to the family A_{k+1} of corank-1-singularities (13.9), while under the weaker Rüssmann condition also singularities of corank 2 may occur. Theorem 13.1 treats the former situation, while the example in Sect. 13.3 concerns the latter case. All these singularities describe the normal behaviour of Cantor bundles of (degenerate) tori, with Cantorized unfoldings.

The corank-2-singularities D_4, D_5, D_6 and E_6 of low co-dimension are still simple and E_8 and the complete family D_{k+1}, $k \geq 3$, unfolded by the umbilics, are simple as well. Next to these also singularities with modal parameters become possible, leading for high N to all quasi-periodic bifurcations of [20]. We note that, for these bifurcations to take their standard form in a resonance gap, a scaling will be needed.

In the case of an m-fold resonance the above normalization procedure applied to (13.3) leads to an $(N-m)$-parameter family of Hamiltonian systems defined on $\mathbb{T}^m \times \mathbb{R}^m$. Here, non-degenerate minima correspond to elliptic $(N-m)$-dimensional tori [10, 12, 15, 28]. For these cases there exist many results on quasi-periodic persistence, employing KAM Theory.

In the spirit of this paper, one should consider degenerate minima as well. Under the Kolmogorov condition (13.6) this leads to corank-m-singularities. The corresponding quasi-periodic bifurcation theory still has to be developed. Under the Rüssmann condition similarly this gives rise to singularities of corank $2m$. We like to stress that none of these complications can be avoided.

Acknowledgements We thank an anonymous referee for challenging us with the example $H(\varphi, I) = H_0(I) + \epsilon H_1(\varphi)$ of a non-versal perturbation.

References

1. Arnol'd, V.I.: Geometrical Methods in the Theory of Ordinary Differential Equations. Springer, Berlin (1983)
2. Arnol'd, V.I., Kozlov, V.V., Neishtadt, A.I.: Mathematical Aspects of Classical and Celestial Mechanics. In: Arnol'd, V.I. (ed.) Dynamical Systems, vol. III. Springer, Berlin (1988)
3. Arnol'd, V.I., Vasil'ev, V.A., Goryunov, V.V., Lyashko, O.V.: Singularity theory. I: Singularities, local and global theory. In: Arnol'd, V.I. (ed.) Dynamical Systems, vol. VI. Springer, Berlin (1993)
4. Broer, H.W., Ciocci, M.C., Hanßmann, H., Vanderbauwhede, A.: Quasi-periodic stability of normally resonant tori. Physica D **238**(3), 309–318 (2009)
5. Broer, H.W., Hanßmann, H., You, J.: Bifurcations of normally parabolic tori in Hamiltonian systems. Nonlinearity **18**, 1735–1769 (2005)
6. Broer, H.W., Hanßmann, H., You, J.: Umbilical torus bifurcations in Hamiltonian systems. J. Differ. Equ. **222**, 233–262 (2006)
7. Broer, H.W., Hoo, J., Naudot, V.: Normal linear stability of quasi-periodic tori. J. Differ. Equ. **232**(2), 355–418 (2007)
8. Broer, H.W., Huitema, G.B., Sevryuk, M.B.: Quasi-Periodic Motions in Families of Dynamical Systems: Order amidst Chaos. LNM, vol. 1645. Springer, Berlin (1996)
9. Broer, H.W., Huitema, G.B., Takens, F.: Unfoldings of quasi-periodic tori. Mem. AMS **83** 421, 1–82 (1990)
10. Broer, H.W., Sevryuk, M.B.: KAM theory: quasi-periodicity in dynamical systems. In: Broer, H.W., Hasselblatt, B., Takens, F. (eds.) Handbook of Dynamical Systems, vol. 3, pp. 249–344. North-Holland, Amsterdam (2010)
11. Broer, H.W., Tangerman, F.M.: From a differentiable to a real analytic perturbation theory, applications to the Kupka Smale theorems. Ergodic Theory Dynam. Syst. **6**, 345–362 (1986)
12. Cong, F., Küpper, T., Li, Y., You, J.: KAM-type theorem on resonant surfaces for nearly integrable Hamiltonian systems. J. Nonlinear Sci. **10**, 49–68 (2000)
13. Cheng, C.-Q.: Birkhoff–Kolmogorov–Arnold–Moser tori in convex Hamiltonian systems. Commun. Math. Phys. **177**, 529–559 (1996)
14. Cheng, C.-Q.: Lower Dimensional Invariant Tori in the Regions of Instability for Nearly Integrable Hamiltonian Systems. Commun. Math. Phys. **203**, 385–419 (1999)
15. Cheng, C.-Q., Wang, S.: The surviving of lower dimensional tori from a resonant torus of Hamiltonian systems. J. Differ. Equ. **155**(2), 311–326 (1999)
16. Corsi, L., Gentile, G.: Melnikov theory to all orders and Puiseux series for subharmonic solutions. J. Math. Phys. **49**(112701), 1–29 (2008)
17. Gallavoti, G., Gentile, G., Giuliani, A.: Fractional Lindstedt series. J. Math. Phys. **47**(012702), 1–33 (2006)
18. Hanßmann, H.: The Quasi-Periodic Centre-Saddle Bifurcation. J. Differ. Equ. **142**(2), 305–370 (1998)
19. Hanßmann, H.: Hamiltonian torus bifurcations related to simple singularities. In: Ladde, G.S., Medhin, N.G., Sambandham, M. (eds.) Dynamic Systems and Applications, Atlanta 2003, pp. 679–685. Dynamic Publishers, Atlanta (2004)
20. Hanßmann, H.: Local and Semi-Local Bifurcations in Hamiltonian Dynamical Systems—Results and Examples. LNM, vol. 1893. Springer, Berlin (2007)
21. Hanßmann, H., Sommer, B.: A degenerate bifurcation in the Hénon–Heiles family. Celestial Mech. Dynam. Astronom. **81**(3), 249–261 (2001)
22. Hirsch, M.W.: Differential Topology. Springer, Berlin (1976)
23. Markus, L., Meyer, K.R.: Generic Hamiltonian dynamical systems are neither integrable nor ergodic. Mem. AMS **144**, 1–52 (1974)
24. Poénaru, V.: Singularités C^∞ en Présence de Symétrie. LNM, vol. 510. Springer, Berlin (1976)
25. Pöschel, J.: Integrability of Hamiltonian Systems on Cantor Sets. Commun. Pure Appl. Math. **35**, 653–696 (1982)

26. Rüssmann, H.: Invariant tori in non-degenerate nearly integrable Hamiltonian systems. Regul. Chaotic Dyn. **6**, 119–204 (2001)
27. Thom, R.: Structural Stability and Morphogenesis. An Outline of a General Theory of Models, 2nd edn. Addison-Wesley, Reading (1989)
28. Sevryuk, M.B.: The classical KAM theory at the dawn of the twenty-first century. Moscow Math. J. **3**(3), 1113–1144 (2003)
29. Xu, J., You, J.: A note on the paper "Invariant tori for nearly integrable Hamiltonian systems with degeneracy". Nanjing University. Preprint (2006)
30. You, J.: A KAM theorem for hyperbolic-type degenerate lower dimensional tori in Hamiltonian systems. Commun. Math. Phys. **192**(1), 145–168 (1998)

Chapter 14
Deformation of Geometry and Bifurcations of Vortex Rings

James Montaldi and Tadashi Tokieda

Abstract We construct a smooth family of Hamiltonian systems, together with a family of group symmetries and momentum maps, for the dynamics of point vortices on surfaces parametrized by the curvature of the surface. Equivariant bifurcations in this family are characterized, whence the stability of the Thomson heptagon is deduced without recourse to the Birkhoff normal form, which has hitherto been a necessary tool.

Introduction

This paper introduces one geometric idea and implements it in one dynamical problem.

Here is the problem. On the Euclidean plane, a ring of identical point vortices shaped as a regular n-gon is a relative equilibrium, in that it spins while keeping the same shape. Is this solution stable if we perturb the initial shape away from the regular n-gon? When $n < 7$, linear stability analysis (first carried out by Thomson [22]) concludes that the solution is stable; when $n > 7$, it is likewise unstable. But when $n = 7$, degeneracy makes linear analysis inapplicable and

J. Montaldi (✉)
School of Mathematics, University of Manchester, Oxford Road, Manchester M13 9PL, UK
e-mail: j.montaldi@manchester.ac.uk

T. Tokieda
Trinity Hall, Cambridge CB2 1TJ, UK
e-mail: t.tokieda@dpmms.cam.ac.uk

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_14,

prevents us from concluding. Is the regular 7-gon stable or unstable? This is known as the *Thomson heptagon* problem. It has been answered in the affirmative, see [11, 16, 20], although as pointed out in [11] the argument in [16] is incomplete. Indeed part of our approach could be viewed as completing the argument of [16].

The spirit of the approach goes back to Poincaré. In the theory of dynamical systems, the simplest solution methods to problems require some nondegeneracy condition (nonzero determinant, nonresonant frequencies, ...). When the problem, call it $\mathscr{P}$, is degenerate, we have to mobilize heavy machinery (cf. [20] for a proof with recourse to the Birkhoff normal form that the Thomson heptagon is nonlinearly stable, as well as for historical details). But there is an alternative approach. Embed $\mathscr{P} = \mathscr{P}_0$ in a parametric family $\mathscr{P}_\lambda$ of problems and deform it away from degeneracy. The problem $\mathscr{P}$ can become tractable when regarded as $\lim_{\lambda \to 0} \mathscr{P}_\lambda$, if we happen to understand well enough the bifurcations that occur in such a deformation. Thus, this approach trades one machinery for another, of bifurcation theory. The point is that the latter sometimes sheds an unusual light on the problem compared to the former.

In the classic applications of this idea, people deform the dynamical system by adding a perturbation term in λ. This, however, we cannot do in our special instance of the heptagon problem if we wish to preserve the hydrodynamic motivation: as long as we are studying the dynamics on the plane, it makes little physical sense to tamper with the Hamiltonian.

We therefore deform not so much the dynamical system *but rather the phase space* on which the system evolves. A deformed choice of the phase space fixes canonically, by the hydrodynamic motivation, a deformed Hamiltonian formalism. Explicitly, we take λ to be the Gaussian curvature[1] and deform the original plane to $\lambda > 0$ (family of spheres) and to $\lambda < 0$ (family of hyperbolic planes). Corresponding to this family of surfaces parametrized by λ, we must write a whole parametric family of Hamiltonian systems for point vortices: a family of symplectic (Kähler) forms depending on λ, a family of symmetry groups and momentum maps depending on λ, a family of invariant Hamiltonians depending on λ—all dependences arranged to be smooth. We do this in Sect. 14.1. In Sect. 14.2 we carry out the stability analysis for the parametric family. Bifurcations are characterized, and the nonlinear stability of the heptagon is deduced, in Sect. 14.3.

The idea of *deforming the geometry* underlying the dynamics of point vortices, in particular as a route to a better understanding of the Thomson heptagon problem, arose during an evening conversation between the two authors in Peyresq, in the summer of 2003. We have since discussed it in seminars and conferences, and part of it has leaked into the literature [1]. We set down the full story in this paper. The stability of the Thomson heptagon is stated below as Corollary 14.7.

[1] Actually formulaic convenience leads us to take 4λ to be the Gaussian curvature.

14.1 Smooth Family of Geometries

In this section we describe the family of surfaces of constant curvature, containing the hyperbolic planes ($\lambda < 0$), the Euclidean plane ($\lambda = 0$), and the spheres ($\lambda > 0$). Each is a homogenous space, i.e. an orbit of its group of symmetries, and for the different signs of λ we describe the different groups. We begin with the 1-parameter family of Lie algebras $\mathfrak{g}_\lambda$ (rather than groups) in Sect. 14.1.1, and in order to obtain the corresponding family of surfaces $\mathscr{M}_\lambda$, we shift the usual linear action of these Lie algebras to obtain affine linear actions. In Sect. 14.1.2 the geometry of $\mathscr{M}_\lambda$ is described; among other things it has curvature 4λ. The Hamiltonian for point vortices on $\mathscr{M}_\lambda$ is based on Green's function for the Laplacian on $\mathscr{M}_\lambda$, and we meet three possible choices of Green's function according to choices of the "boundary condition". In Sect. 14.1.3 we comment on the implications of the different choices of Green's function.

14.1.1 Lie Algebras

Let $\lambda \in \mathbb{R}$. On $\mathbb{R}^3$ with coordinates (x, y, u), consider the family of metrics

$$\mathrm{d}s^2 = \mathrm{d}x^2 + \mathrm{d}y^2 + \lambda\,\mathrm{d}u^2. \tag{14.1}$$

The Lie group of linear transformations preserving this metric will be denoted $\mathsf{SO}(q_\lambda)$, where $q_\lambda = \mathrm{diag}[1, 1, \lambda]$ is the metric tensor. For the Lie algebra, we have $X \in \mathfrak{so}(q_\lambda)$ if and only if $X^T q_\lambda + q_\lambda X = 0$. A basis for $\mathfrak{so}(q_\lambda)$ is

$$X_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -\lambda \\ 0 & 1 & 0 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 0 & \lambda \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{14.2}$$

(Strictly speaking, the third column of X_i is arbitrary when $\lambda = 0$, meaning the family of all automorphisms of (14.1) is not *flat*. We are picking a component which *is* a flat family over λ.) This basis satisfies the commutation relations

$$[X_1, X_2] = \lambda X_3, \quad [X_2, X_3] = X_1, \quad [X_3, X_1] = X_2.$$

From now on, we shall abbreviate $G_\lambda = \mathsf{SO}(q_\lambda)$ and $\mathfrak{g}_\lambda = \mathfrak{so}(q_\lambda)$.

The Lie algebra $\mathfrak{g}_\lambda$ is isomorphic to $\mathfrak{so}(3)$ for $\lambda > 0$, to $\mathfrak{se}(2)$ for $\lambda = 0$, and to $\mathfrak{sl}(2)$ for $\lambda < 0$. Indeed, for $\lambda \neq 0$, the standard commutation relations are recovered by rescaling the basis to $\{|\lambda|^{-1/2} X_1, |\lambda|^{-1/2} X_2, X_3\}$. So

$$G_\lambda \simeq \begin{cases} \mathsf{SO}(3) & \text{if} \lambda > 0 \\ \mathsf{SE}(2) & \text{if} \lambda = 0 \\ \mathsf{SL}(2,\mathbb{R}) & \text{if} \lambda < 0. \end{cases}$$

It is seen from the commutation relations that in the adjoint representation of $\mathfrak{g}_\lambda$ the basis elements are represented by $\mathrm{ad}_{X_j} = -X_j^T$; in other words, if $\sum_j a_j X_j \in \mathfrak{g}_\lambda$ is written as a vector $\mathbf{u} = (a_1\ a_2\ a_3)^T$, then $\mathrm{ad}_{X_j}(\mathbf{u}) = -X_j^T \mathbf{u}$. In the *coadjoint* representation the basis elements $X_j \in \mathfrak{g}_\lambda$ are represented by the matrices X_j themselves. Thus the original $\mathbb{R}^3$ from which we started may be naturally identified with $\mathfrak{g}_\lambda^*$.

Affine Action

While the coadjoint action defined by the matrices X_j depends continuously on λ, it is not possible to track a single orbit continuously as λ crosses 0. Yet, in what follows, we wish to do just that. To this end we must shift the linear coadjoint action by a translation, making it an affine action.

The affine action of the Lie algebra is given by

$$X \cdot \mu = X\mu + \tau(X), \tag{14.3}$$

where $X\mu$ is the linear action part (matrix times vector) and

$$\tau(aX_1 + bX_2 + cX_3) = \begin{pmatrix} -b/2 \\ a/2 \\ 0 \end{pmatrix}$$

is the translation. The orbit we track is the one through the origin, cf. Sect. 14.1.2.

Remark 14.1. Our translation τ, a function of the element of $\mathfrak{g}_\lambda$, is a 1-cocycle taking values in $\mathfrak{g}_\lambda^*$, a *symplectic cocycle* in Souriau's terminology [21] because the matrix of τ is skew-symmetric. It is known that every cocycle is exact when the group is semi-simple, and our G_λ is semi-simple for $\lambda \neq 0$. Here $\tau = \delta((0,\ 0,\ 1/2\lambda)^T)$, since by definition $\delta(\mu)(X) = -\mathrm{coad}_X \mu = -X\mu$. The natural invariant Poisson structure on $\mathbb{R}^3 = \mathfrak{g}_\lambda^*$ with the cocycle τ is given by (cf. [15, 21])

$$\{f,\ g\}(\mu) = \langle \mu,\ [\mathrm{d}f(\mu), \mathrm{d}g(\mu)] \rangle - \langle \tau(\mathrm{d}f(\mu)), \mathrm{d}g(\mu) \rangle \tag{14.4}$$

under the identification $\mathrm{d}f(\mu), \mathrm{d}g(\mu) \in (\mathfrak{g}_\lambda^*)^* \simeq \mathfrak{g}_\lambda$. The Casimir for this Poisson structure is $x^2 + y^2 + \lambda u^2 - u$, so level sets of this function are the orbits of the

shifted coadjoint action (14.3), of which we shall take advantage below. For the record, the Kostant–Kirillov–Souriau symplectic form on the affine coadjoint orbits is given by the same formula,

$$\Omega_\mu(\mathbf{u}, \mathbf{v}) = \langle \mu, [\xi, \eta] \rangle - \langle \tau(\xi), \eta \rangle \,,$$

where $\mathbf{u} = \mathrm{coad}_\xi \mu$ and $\mathbf{v} = \mathrm{coad}_\eta \mu$.

14.1.2 Surfaces

Now consider the family of quadratic surfaces through the origin in $\mathbb{R}^3$,

$$x^2 + y^2 + \lambda u^2 - u = 0. \tag{14.5}$$

When $\lambda > 0$, this looks like an ellipsoid with centre at $(x, y, u) = (0, 0, 1/2\lambda)$. With the metric (14.1), however, this ellipsoid-looking surface is in fact a sphere of radius $1/2\sqrt{\lambda}$. Its Gaussian curvature is 4λ.

When $\lambda = 0$, (14.5) defines the paraboloid $u = x^2 + y^2$, and the metric is the usual metric on the xy-plane lifted to the paraboloid by orthogonal projection, so is of curvature 0.

When $\lambda < 0$, the metric (14.1) becomes Lorentzian, but restricted to either sheet of the 2-sheeted hyperboloid defined by (14.5) it induces the hyperbolic metric of constant negative curvature 4λ; we consider just the "upper sheet" that passes through the origin, see Fig. 14.1.

We refer to the surface (14.5) with metric induced from (14.1) as $\mathscr{M}_\lambda$. It is easy to check that $\mathscr{M}_\lambda$ is invariant under the infinitesimal action (14.3), and is therefore an orbit of the affine coadjoint G_λ-action on $\mathfrak{g}_\lambda^*$.

To create a *uniform coordinate system* on $\mathscr{M}_\lambda$, we use stereographic projection on the xy-plane, centred at the point $(0, 0, 1/\lambda)$ (where $\mathscr{M}_\lambda$ intersects the u-axis, besides the origin); for $\lambda = 0$ this is the orthogonal projection. We also identify the xy-plane with $\mathbb{C}$, via $z = x + \mathrm{i}y$. The map inverse to the projection has the formula

$$z \mapsto \begin{pmatrix} x + \mathrm{i}y \\ u \end{pmatrix} = \frac{1}{1 + \lambda |z|^2} \begin{pmatrix} z \\ |z|^2 \end{pmatrix}. \tag{14.6}$$

The domain of this map is $\{z \in \mathbb{C} \mid 1 + \lambda |z|^2 > 0\}$, which is the entire plane if $\lambda \geqslant 0$ and a bounded disc (Poincaré disc) if $\lambda < 0$. For the sphere, the equator corresponds to $|z|^2 = 1/\lambda$, while the point antipodal to z is $-1/\lambda \bar{z}$.

The metric on the surface $\mathscr{M}_\lambda$ induced from that in (14.1), in terms of the complex variable z, is

$$\mathrm{d}s^2 = \frac{1}{\sigma^2} |\mathrm{d}z|^2,$$

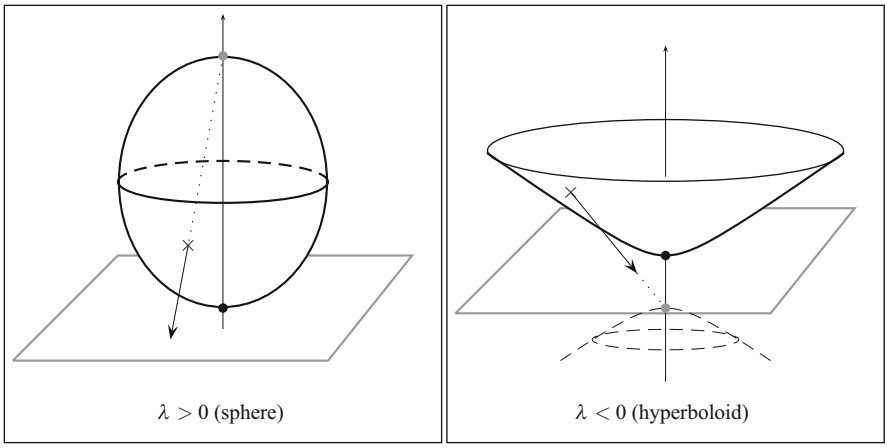

Fig. 14.1 Geometries in the 1-parameter family

where $\sigma = 1 + \lambda|z|^2$, a notation we shall use throughout. The circle $|z| = r$ in $\mathbb{C}$ maps to a circle of radius a on $\mathscr{M}_\lambda$, where[2]

$$a = \begin{cases} \frac{1}{\sqrt{\lambda}} \tan^{-1}\left(r\sqrt{\lambda}\right) & \text{if} \lambda > 0 \\ r & \text{if} \lambda = 0 \\ \frac{1}{\sqrt{-\lambda}} \tanh^{-1}\left(r\sqrt{-\lambda}\right) & \text{if} \lambda < 0. \end{cases} \tag{14.7}$$

Pulling back the vector fields X_j of the Lie algebra (or rather their affine variants shifted by τ) via the stereographic projection yields

$$\xi_1(x,y) = \tfrac{1}{2}\left(2\lambda xy,\ 1-\lambda(x^2-y^2)\right), \quad \xi_2(x,y) = -\tfrac{1}{2}\left(1+\lambda(x^2-y^2),\ 2\lambda xy\right),$$

$$\xi_3(x,y) = (-y,x),$$

or in complex variables

$$\xi_1 + \mathrm{i}\xi_2 = -\mathrm{i}\partial_{\bar{z}} - \mathrm{i}\lambda\, z^2 \partial_z, \quad \xi_3 = \mathrm{i}\left(z\,\partial_z - \bar{z}\,\partial_{\bar{z}}\right),$$

or in polar coordinates

$$\xi_1 + \mathrm{i}\xi_2 = \tfrac{1}{2}\mathrm{e}^{\mathrm{i}\theta}\,\frac{1-\lambda r^2}{r}\,\partial_\theta + \tfrac{1}{2}\,\mathrm{i}\,\mathrm{e}^{\mathrm{i}\theta}\sigma\,\partial_r, \quad \xi_3 = \partial_\theta.$$

[2]Despite three formulae, a is a single analytic function of r, λ, with series expansion $a = r - \frac{1}{3}r^3\lambda + \frac{1}{5}r^5\lambda^2 - \frac{1}{7}r^7\lambda^3 + \cdots$ convergent for $|r^2\lambda| < 1$.

Symplectic Structures

Up to a scalar multiple, there exists a unique $\mathsf{SO}(q_\lambda)$-invariant symplectic form on $\mathscr{M}_\lambda$. We choose the scalar so that

$$\Omega_\lambda = \frac{2}{\sigma^2}\mathrm{d}x \wedge \mathrm{d}y = \frac{\mathrm{i}}{\sigma^2}\mathrm{d}z \wedge \mathrm{d}\bar{z}. \tag{14.8}$$

The choice of scaling is such that the sphere $\mathscr{M}_\lambda$ of radius $1/2\sqrt{\lambda}$ acquires symplectic area $2\pi\lambda^{-1}$. With respect to the basis $\{X_1, X_2, X_3\}$ for the Lie algebra, the momentum map takes the form

$$\mathbf{J}_\lambda(z) = \frac{1}{\sigma}(z, |z|^2). \tag{14.9}$$

This coincides with the inclusion $\mathscr{M}_\lambda \hookrightarrow \mathbb{R}^3$ given in (14.6), which shows that Ω_λ coincides with the KKS symplectic form on the affine coadjoint orbit.

Green's Functions

The metric (14.1) on $\mathbb{R}^3$ induces the metric on $\mathscr{M}_\lambda$. In terms of the uniform coordinate system (14.6), the metric tensor is $\sigma^{-2}\mathrm{diag}[1, 1]$. The Laplace–Beltrami operator on $\mathscr{M}_\lambda$ is

$$\Delta f = \sigma^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) f = \tfrac{1}{4}\sigma^2 \frac{\partial^2}{\partial z \partial \bar{z}}\, f.$$

The 2-point Green's function for this operator is

$$G(z; w) = \log |z - w|^2. \tag{14.10}$$

This satisfies $\Delta_z G = 0$ for $z \neq w$ and has a logarithmic singularity at $z = w$. When we regard the plane as a model for (most of) the sphere, G has another singularity at $z = \infty$.

An alternative Green's function is

$$G_1(z; w) = \log \frac{|z - w|^2}{|1 + \lambda z \bar{w}|^2}. \tag{14.11}$$

This satisfies $\Delta_z G_1(z; w) = 0$ for $z \notin \{w, -1/\lambda\bar{w}\}$ and has a logarithmic singularity at those excluded points (which are antipodal to each other), but is regular at $z = \infty$.

The Green's function that is usually used on the sphere is the "log of Euclidean distance", whose expression after stereographic projection is

$$G(z;w) = \log \frac{|z-w|^2}{(1+\lambda|z|^2)(1+\lambda|w|^2)}. \tag{14.12}$$

Away from the pole at $z = w$, this satisfies $\Delta_z G(z;w) = -4\lambda$, so it is not, if we go by the book, a Green's function. This function is regular at $z = \infty$, and also has a well-defined limit as $w \to \infty$, namely $G(z;\infty) = -\log\sigma$ (up to an additive constant of $-\log\lambda$), which checks against $\Delta_z(-\log\sigma) = -4\lambda$.

In the next section we comment on the differences among these Green's functions in the context of the dynamics of point vortices.

14.1.3 Hamiltonians for Point Vortices

We recall how the Hamiltonian formalism for the dynamics of point vortices works.

Let $(u,v)^T$ be the velocity field of an inviscid, incompressible flow on a domain $D \subseteq \mathbb{R}^2 \simeq \mathbb{C}$. The incompressibility $\frac{\partial}{\partial x}u + \frac{\partial}{\partial y}v = 0$ implies the existence of a *stream function* $\psi : D \to \mathbb{R}$ such that

$$u = \frac{\partial}{\partial y}\psi, \quad v = -\frac{\partial}{\partial x}\psi. \tag{14.13}$$

The curl of the velocity[3] is $\frac{\partial}{\partial x}v - \frac{\partial}{\partial y}u = -\Delta\psi$. The boundary condition for an inviscid flow is that $(u,v)^T$ be tangent to ∂D everywhere, equivalently that every connected component of ∂D be a level set of ψ. The total circulation along all the boundary components is, by Stokes's theorem,

$$\int_{\partial D} u\,\mathrm{d}x + v\,\mathrm{d}y = \int_D -\Delta\psi\,\mathrm{d}x\mathrm{d}y, \tag{14.14}$$

the total curl present on D.

Now consider a model situation where the flow is generated by a curl concentrated at a singularity $z_0 = x_0 + \mathrm{i}y_0$ and the circulation around that singularity is $2\pi\kappa$:

$$-\Delta\psi(z) = 2\pi\kappa\delta(z - z_0).$$

[3]The minus sign makes $-\Delta$ a positive operator. But we shall be casual about the sign and use $+\Delta$ as well as $-\Delta$. All that the casualness causes is to reverse the direction of the flow.

We recognize that, up to the sign and a scalar coefficient, ψ is Green's function. We say that we have a *point vortex* of *vorticity* κ at z_0. In the situation where we have point vortices of vorticities $\kappa_1, \ldots, \kappa_n$ at $z_1, \ldots, z_n$, each vortex moves carried by the sum of the flows generated by all the other vortices. Equation (14.13) shows that the dynamics then is Hamiltonian.

The theory is written analogously on any domain of any Riemann surface. In particular, on the surfaces $\mathscr{M}_\lambda$ discussed above, the Hamiltonian is

$$H_\lambda(z_1, \ldots, z_N) = -\frac{1}{4\pi} \sum_{i<j} \kappa_i \kappa_j G_\lambda(z_i; z_j).$$

Note an unusual feature of this Hamiltonian system: unlike in classical mechanics, the phase space here is not the cotangent bundle of anything, but rather the n-fold product $\mathscr{M}_\lambda \times \cdots \times \mathscr{M}_\lambda$ (minus the diagonals if we want *a priori* to avoid collisions) with the weighted-sum symplectic form $\kappa_1 \Omega_\lambda \oplus \cdots \oplus \kappa_n \Omega_\lambda$.

For $\lambda > 0$, $\mathscr{M}_\lambda$ is compact without boundary. In this case, the fact of nature (14.14) forces the total vorticity to be zero: $\sum_j \kappa_j = 0$. This, in principle, bans placing a lone point vortex on $\mathscr{M}_\lambda$ for $\lambda > 0$ or on any closed Riemann surface. The dodge around this ban, favoured in the literature, is to impose a constant *background vorticity*

$$-\text{sum of circulations/area} = -\lambda \sum_j \kappa_j,$$

which results in (14.12). For a (geo)physical example, a rigidly rotating sphere entails such a background vorticity. But even if background vorticity dodges around $\lambda > 0$, continuing it to $\lambda \leqslant 0$ gets us into trouble, for on these noncompact surfaces, background vorticity imparts an infinite amount of energy to the flow. So for $\lambda \leqslant 0$ the family (14.10) seems preferable.

However, welding together (14.12) for $\lambda \geqslant 0$, (14.10) for $\lambda < 0$ has a decisive defect: the resulting family is *not smooth* in λ. There are three options for having a smooth family.

1. Use (14.10). This costs postulating an immobile vortex of vorticity $-\sum_j \kappa_j$ at the North Pole for $\lambda > 0$.
2. Use (14.11). This costs introducing "counter-vortices" at antipodes to the z_js for $\lambda > 0$ (recall that the point antipodal to z is $-1/\lambda\bar{z}$).
3. Use (14.12). This costs infinite energy for $\lambda \leqslant 0$.

In the planar case $\lambda = 0$ all three Green's functions (14.10)–(14.12) agree.

In this paper, we opt for the family defined in (14.12) with the constant background vorticity, because the infinite energy of a tame flow would not shock any fluid dynamicist—flows on the plane uniform at infinity and such are handled routinely—whereas the smoothness of the family, without postulating extraneous objects, is essential for us. In Sect. 14.4 we sketch how the analysis can be adapted to the other two options.

14.2 Nondegenerate Analysis of Vortex Rings

In this section we study a ring of n vortices with identical vorticities κ, shaped as an n-gon on $\mathscr{M}_\lambda$. We evaluate the Hamiltonian, the momentum map, and the augmented Hamiltonian at the regular ring in Sect. 14.2.1. We obtain the Hessian of the augmented Hamiltonian and its spectral data in Sect. 14.2.2. In Sect. 14.2.3 we construct the symplectic slice and say what we can, as far as this linear analysis goes, about the stability of the ring. The answers will depend not only on n but also on λ.

Recall that the augmented Hamiltonian is given by $H - \omega J$, where $J = J_\lambda$ is the conserved quantity coming from the rotational symmetry; it represents the Hamiltonian in a frame rotating with angular velocity ω. Its critical points are therefore equilibria in the rotating frame, i.e. relative equilibria, and the Hessian restricted to the symplectic slice determines the stability.

14.2.1 Regular Ring

Let γ be a primitive n th root of unity. When the vortices are placed at the vertices $z_j = r\mathrm{e}^{2\pi \mathrm{i} j/n}$ $(j = 1, \ldots, n)$ of a regular n-gon of radius r, the Hamiltonian H_λ takes the value

$$h_\lambda(r^2) = -\frac{n\kappa^2}{8\pi} \sum_{j=1}^{n-1} G_\lambda(r, r\gamma^j),$$

where $G_\lambda = \log \circ \rho_\lambda$ with $\rho_\lambda(r, r\gamma^j) = |1 - \gamma^j|^2 r^2/(1 + \lambda r^2)^2$. In view of the identity $\prod_{j=1}^{n-1}(1 - \gamma^j) = n$,

$$\prod_{j=1}^{n-1} \rho_\lambda(r, r\gamma^j) = n^2 \left(\frac{r}{1 + \lambda r^2}\right)^{2(n-1)},$$

hence

$$h_\lambda(r^2) = -\frac{n(n-1)\kappa^2}{8\pi} \log \frac{r^2}{(1 + \lambda r^2)^2} + \text{const.}$$

Recall the momentum map given in (14.9). The first component vanishes for these regular rings, while the value of the second component at the regular ring is

$$J_\lambda(r^2) = \frac{n\kappa r^2}{1 + \lambda r^2}.$$

(Notice that the full momentum map is typeset in bold, while this component is not.) Ignoring the constant, the augmented Hamiltonian then takes the value

$$\begin{aligned}\hat{h}_\lambda(r^2) &= h_\lambda(r^2) - \omega J_\lambda(r^2)\\ &= -\frac{n(n-1)\kappa^2}{8\pi}\log\frac{r^2}{(1+\lambda r^2)^2} - \omega\frac{n\kappa r^2}{1+\lambda r^2},\end{aligned}$$

which admits a critical point at $r = r_0 \neq 0$ if and only if

$$\omega = \omega_0 = -\frac{(n-1)\kappa}{8\pi}\,\frac{1-\lambda^2 r_0^4}{r_0^2}. \tag{14.15}$$

This is the angular velocity of the regular ring. (It is a little surprising that ω_0 is even in λ.)

14.2.2 Hessians

We continue with the system of n identical point vortices, all with vorticity κ. In the uniform coordinate system on $\mathscr{M}_\lambda$, the Hamiltonian with Green's function (14.12) is

$$H_\lambda(z_1,\dots,z_n) = -\frac{\kappa^2}{4\pi}\sum_{i<j}\log\frac{|z_i - z_j|^2}{\sigma_i\sigma_j},$$

where $\sigma_j = 1 + \lambda|z_j|^2$ $(j = 1,\dots,n)$. The augmented Hamiltonian then is

$$\hat{H}_\lambda(z_1,\dots,z_n) = H_\lambda(z_1,\dots,z_n) - \omega\sum_{j=1}^{n}\kappa\frac{|z_j|^2}{\sigma_j}.$$

We saw above that this is critical at $z_j = r_0 e^{2i\pi j/n}$ and $\omega = \omega_0$ as in (14.15). The entries in the Hessian are

$$\begin{aligned}
\frac{\partial^2}{\partial r_j^2}\hat{H}_\lambda &= A & \frac{\partial^2}{\partial r_j \partial r_k}\hat{H}_\lambda &= \frac{\kappa^2}{4\pi r_0^2\left(1-\cos\frac{2\pi(j-k)}{n}\right)}\\
\frac{\partial^2}{\partial \theta_j^2}\hat{H}_\lambda &= \frac{\kappa^2}{24\pi}(n^2-1) & \frac{\partial^2}{\partial \theta_j \partial \theta_k}\hat{H}_\lambda &= -\frac{\kappa^2}{4\pi\left(1-\cos\frac{2\pi(j-k)}{n}\right)}\\
\frac{\partial^2}{\partial r_j \partial \theta_k}\hat{H}_\lambda &= 0 \quad (\forall j,k).
\end{aligned} \tag{14.16}$$

At the critical point, and with $\omega = \omega_0$, we have

$$A = \frac{(n-1)\kappa^2}{24\pi r_0^2 \sigma^2}\left((5-n)\sigma^2 + 6\tilde{\sigma}^2\right), \tag{14.17}$$

where

$$\sigma = 1 + \lambda r_0^2, \quad \tilde{\sigma} = 1 - \lambda r_0^2$$

(so that $\sigma + \tilde{\sigma} = 2$). We used the identity, valid for $0 \leqslant \ell \leqslant n$, cf. [8]:

$$\sum_{j=1}^{n-1} \frac{\cos(2\pi\ell j/n)}{1-\cos(2\pi j/n)} = \frac{1}{6}(n^2-1) - \ell(n-\ell). \tag{14.18}$$

Eigenvalues of the Hessian

The Hessian matrix $\mathrm{d}^2\hat{H}_\lambda$ is block-diagonal, with two $n \times n$ blocks both of which are symmetric circulant, so the eigenvalues can be written down at once.

Following the notation in [13], for $\ell = 0, 1, \ldots, \lfloor n/2 \rfloor$ define the *Fourier tangent vectors*

$$\begin{aligned} \zeta_r^{(\ell)} &= \alpha_r^{(\ell)} + \mathrm{i}\beta_r^{(\ell)} = \sum_{j=1}^{n} \exp(-2\pi\mathrm{i}\,\ell j/n)\,\delta r_j \\ \zeta_\theta^{(\ell)} &= \alpha_\theta^{(\ell)} + \mathrm{i}\beta_\theta^{(\ell)} = \tfrac{1}{r_0}\sum_{j=1}^{n} \exp(-2\pi\mathrm{i}\,\ell j/n)\,\delta\theta_j. \end{aligned} \tag{14.19}$$

Here δr_j denotes the unit tangent vector in the r_j-direction, and similarly for $\delta\theta_j/r_0$. For $\ell = 0,\ n/2$, we have $\beta_r^{(\ell)} = \beta_\theta^{(\ell)} = 0$ and the $\zeta^{(\ell)}$s are real. The αs and βs form a set of $2n$ linearly independent vectors, forming a basis for the tangent space. For the record,

$$\delta r_j = \frac{1}{n}\sum_{\ell=1}^{n} \exp(2\pi\mathrm{i}\,\ell j/n)\zeta_r^{(\ell)},$$

and similarly for $\delta\theta_j$. The span $V_\ell = \left\langle \alpha_r^{(\ell)}, \alpha_\theta^{(\ell)}, \beta_r^{(\ell)}, \beta_\theta^{(\ell)} \right\rangle$ is a subspace of *Fourier modes*; for $\ell = 0,\ n/2$ they are 2-dimensional, while for all other indices ℓ they are 4-dimensional.

As each block of $\mathrm{d}^2\hat{H}_\lambda$ is circulant as well as symmetric, $\zeta_r^{(\ell)}$ and $\zeta_\theta^{(\ell)}$ (or rather their real and imaginary parts) are the eigenvectors. The eigenvalues $\epsilon_r^{(\ell)}$ and $\epsilon_\theta^{(\ell)}$ (of course real) are found to be

$$\begin{aligned}\epsilon_r^{(\ell)} &= A + \tfrac{\kappa^2}{24\pi r_0^2}\left((n^2-1) - 6\ell(n-\ell)\right)\\ &= \tfrac{\kappa^2}{4\pi r_0^2}\left(2(n-1)\tfrac{1+\lambda^2 r_0^4}{\sigma^2} - \ell(n-\ell)\right)\\ \epsilon_\theta^{(\ell)} &= \tfrac{\kappa^2}{4\pi}\ell(n-\ell),\end{aligned} \tag{14.20}$$

where A is given in (14.17). In case $\ell = 1$, these simplify to

$$\epsilon_r^{(1)} = (n-1)\frac{\kappa^2\tilde{\sigma}^2}{4\pi r_0^2\sigma^2}, \quad \epsilon_\theta^{(1)} = (n-1)\frac{\kappa^2}{4\pi}, \tag{14.21}$$

which are both strictly positive.

14.2.3 Symplectic Slice

Not all the eigenvalues are relevant to stability. First, those that are zero because they correspond to directions along the group orbit should be discarded. Second, those corresponding to directions transverse to the level set of the conserved quantities should be discarded, too. This is the process of reduction: restrict to $\operatorname{Ker} \mathrm{d}\mathbf{J}$ (where $\mathbf{J} = \mathbf{J}_\lambda$ is given in (14.9)), then take the complement to the group orbit in that kernel. The resulting space is called the *symplectic slice*, which we denote by $\mathscr{N}$.

We find $\operatorname{Ker} \mathrm{d}\mathbf{J}$ using the Fourier basis above. Identify $\mathfrak{g}_\lambda^*$ with $\mathbb{C} \times \mathbb{R}$. Then

$$\mathrm{d}\mathbf{J}\zeta_r^{(0)} = \frac{nr_0}{\sigma^2}\begin{pmatrix}0\\1\end{pmatrix}, \quad \mathrm{d}\mathbf{J}\zeta_r^{(1)} = \frac{n\tilde{\sigma}}{\sigma^2}\begin{pmatrix}1\\0\end{pmatrix}, \quad \mathrm{d}\mathbf{J}\zeta_\theta^{(1)} = \frac{n}{\sigma}\begin{pmatrix}\mathrm{i}\\0\end{pmatrix},$$

while $\mathrm{d}\mathbf{J}$ vanishes on all other Fourier tangent vectors. Write $V_1' = V_1 \cap \operatorname{Ker} \mathrm{d}\mathbf{J}$. Then

$$V_1' = \left\langle \sigma\alpha_r^{(1)} - \tilde{\sigma}\beta_\theta^{(1)},\ \sigma\beta_r^{(1)} + \tilde{\sigma}\alpha_\theta^{(1)} \right\rangle. \tag{14.22}$$

$\operatorname{Ker} \mathrm{d}\mathbf{J}$ is spanned by the V_ℓ s with $\ell > 1$, V_1', and $\zeta_\theta^{(0)}$. The subspace generated by $\zeta_\theta^{(0)}$ is tangent to the group orbit (an infinitesimal rotation about the origin), so must be discarded. Finally the symplectic slice is

$$\mathscr{N} = V_1' \oplus \bigoplus_{\ell=2}^{\lfloor n/2 \rfloor} V_\ell. \tag{14.23}$$

The relevant eigenvalues are therefore $\epsilon_r^{(\ell)}$, $\epsilon_\theta^{(\ell)}$ for $\ell \geqslant 1$. Now, with respect to the basis for V_1' given in (14.22), the restriction of the Hessian to V_1' is a scalar multiple of the identity. So it has a double eigenvalue

$$\epsilon_1' = \frac{n}{2}\,\sigma^2\epsilon_r^{(1)} + \frac{n}{2r_0^2}\tilde{\sigma}^{\,2}\epsilon_\theta^{(1)} = \frac{n(n-1)\kappa^2}{4\pi r_0^2}\tilde{\sigma}^{\,2}, \tag{14.24}$$

which is strictly positive unless $\lambda r_0^2 = 1$. However, $\lambda r_0^2 = 1$ corresponds to the equator on the sphere, at which the momentum value is left fixed by all of $\mathsf{SO}(3)$. In this case $\mathscr{N}$ drops in dimension, V_1' no longer lies in $\mathscr{N}$, and the corresponding eigenvalue ϵ_1' becomes irrelevant to the stability of the relative equilibrium.

Stability

The relative equilibria in question are rotating rings; thus they are periodic trajectories. The precise sense of stability we are adopting is like the ordinary one of Lyapunov stability, but in terms of G_μ-invariant open sets, where μ is the momentum value at the trajectory and G_μ is the stabilizer of μ: in detail, for every G_μ-invariant neighbourhood V of the trajectory, there exists a G_μ-invariant neighbourhood $U \subseteq V$ such that every trajectory starting in U remains in V for all time.

We can make do with a coarse criterion: if the restriction of the Hessian of the augmented Hamiltonian to the symplectic slice is positive-definite, then the relative equilibrium is stable in our sense. A finer criterion is: if this Hessian is merely non-negative but the augmented Hamiltonian admits a local extremum at the relative equilibrium, then the relative equilibrium is still stable [17, Theorem 1.2].

Among the relevant eigenvalues, $\epsilon_\theta^{(\ell)} > 0$ for $\ell \geqslant 1$ and $\epsilon_1' > 0$. It remains to check the sign of $\epsilon_r^{(\ell)}$ for $\ell \geqslant 2$. From the expression in (14.20), it is clear that the least eigenvalue occurs for $\ell = \lfloor n/2 \rfloor$. This is the criterion we have been after: the relevant eigenvalues are all strictly positive if and only if

$$\frac{1+\lambda^2 r_0^4}{(1+\lambda r_0^2)^2} > \frac{1}{2(n-1)}\left\lfloor \frac{n^2}{4} \right\rfloor. \tag{14.25}$$

The left-hand side is unbounded as a function of r_0 if $\lambda < 0$. The right-hand side is a strictly increasing, unbounded function of n for $n \geqslant 3$. Hence, on one hand, for fixed λ, r_0 a value of n exists beyond which the inequality fails; on the other hand, if $\lambda < 0$, for fixed n, by enlarging r_0^2 sufficiently close to $-1/\lambda$ (its supremum on the hyperbolic plane), we can ensure the inequality holds. We conclude with a result which for spheres is due to [2] (see also [13]):

Theorem 14.2. *On $\mathscr{M}_\lambda$, the relative equilibrium of n identical point vortices in a regular ring of radius a is Lyapunov-stable if (14.25) is satisfied, where a and r_0 are related by (14.7). In particular, in the hyperbolic scenario $\lambda < 0$ this ring is stable if λr_0^2 is sufficiently close to -1.*

Remark 14.3. Physical intuition confirms that a spherical surface destabilizes a ring whereas a hyperbolic surface stabilizes it. Think of a vortex of the ring. As λ gets positive, the adjacent vortices remain relatively near while the diametrically opposite vortices become relatively far, so the former, which tend to knock our vortex perpendicularly to the ring, exert more influence than the latter, which tend to slide our vortex tangentially to the ring. As λ gets negative, the effects are felt the other way round.

Remark 14.4. If the inequality in (14.25) is reversed, then the ring is *linearly unstable*. This is because the $V_{\lfloor n/2 \rfloor}$ mode will have real eigenvalues. Indeed, for n even $V_{n/2}$ is of dimension 2 and the Hessian is indefinite, which suffice to conclude that the eigenvalues are real. For n odd $V_{(n-1)/2}$ is 4-dimensional, so having an indefinite Hamiltonian does not imply the eigenvalues are real. However, the negative eigenspace is spanned by $\alpha_r^{(\ell)}, \beta_r^{(\ell)}$ which is Lagrangian, and then the realness of the eigenvalues follows.

If the parameters are such that (14.25) is an equality, then the stability is not determined by linear analysis. This determination is what our deformation plus bifurcation approach achieves, in Sects. 14.3.5.1 and 14.3.5.2.

We now spell out the conclusions concretely for each of the three geometries. See Fig. 14.2 and the table of bifurcation points below. On the spheres $\lambda > 0$ there are two bifurcation points: the value of λr_0^2 listed in the table below and its reciprocal.

n	4	5	6	7	8	9	10	11	12	13
λr_0^2	0.268	0.172	0.0557	0	−0.0627	−0.101	−0.143	−0.172	−0.202	−0.225

Plane

There is an obvious scale-invariance, and the stability/instability of the ring is independent of the radius a. We recover J.J. Thomson's original result [22] that the ring is stable when $n < 7$ and unstable when $n > 7$. When $n = 7$, we have the so-called Thomson heptagon, whose stability is not determined by linear analysis (but cf. Corollary 14.7).

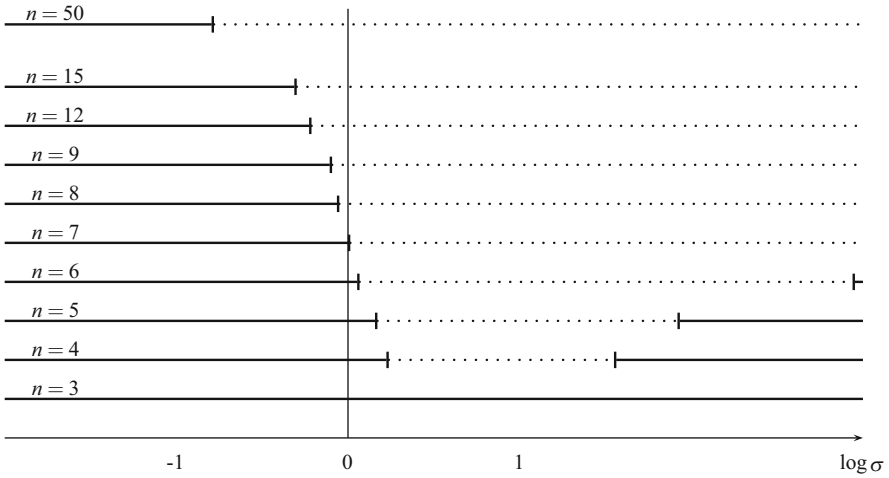

Fig. 14.2 Ranges of stability (*solid lines*) and instability (*dotted lines*) for the dimensionless quantity $\log\sigma = \log(1 + \lambda r_0^2)$

Spheres

When $n \geqslant 7$, the ring is always unstable. When $n < 7$, there is a range of a over which it is stable. Example: when $n = 6$, it is stable for $(\lambda r_0^2 - 2)^2 > 3$ which translates, bearing in mind $\lambda > 0$, into

$$r_0^2 < \frac{2-\sqrt{3}}{\lambda} \quad \text{or} \quad r_0^2 > \frac{2+\sqrt{3}}{\lambda}.$$

The two inequalities correspond to neighbourhoods of the South and North Poles, respectively. For $n = 3$ the ring is always stable, at any radius a. The linear stability of rings of vortices on the sphere was first studied by Polvani and Dritschel [18], and the full nonlinear stability in [2]—see also [13] for more details.

Hyperbolic Planes

When $n \leqslant 7$, the ring is always stable. When $n > 7$, it is stable (for a given λ) for a sufficiently large. Example: when $n = 15$, it is stable for

$$r_0^2 > \frac{2-\sqrt{3}}{|\lambda|}.$$

14.3 Bifurcations Across the Degeneracy

To understand the bifurcations in detail, it is imperative to exploit the symmetries of the system, which for the ring of point vortices is the dihedral group. We begin Sect. 14.3.1 by setting up the dihedral symmetry of the system, and then in Sect. 14.3.2 list the bifurcations we expect in Hamiltonian systems with dihedral symmetry. Following that, in Sect. 14.3.3, we state the main theorem (Theorem 14.6) on which of these bifurcations occur in the dynamics of point vortices; it leads to the nonlinear stability of the Thomson heptagon (Corollary 14.7). In Sect. 14.3.4 we visit the geometry of the bifurcating rings, which enjoy less symmetry (spontaneous symmetry breaking) and are illustrated in Fig. 14.6. Finally, Sect. 14.3.5 proves the main theorem: we perform the calculations needed to justify these "expectations" and decide which of the expectations are actually realized. For the sake of completeness, Sect. 14.3.6 summarizes results from [14] on bifurcations from the equator; they are not covered by the other results as the momentum value there is degenerate (in the sense that it is fixed by the entire group rather than just a 1-parameter subgroup).

14.3.1 Dihedral Group Action

For the simplicity of language, we shall confuse $\mathscr{M}_\lambda$ and the uniform chart $\mathbb{C}$ of Sect. 14.1.2. Points and group actions on $\mathbb{C}$ should be interpreted as their lifts on $\mathscr{M}_\lambda$.

The full system is invariant under the symmetry group G_λ (depending on λ) as in Sect. 14.1.1. For every λ, G_λ contains rotations about the origin, and reflections in lines through the origin together generating a subgroup of G_λ isomorphic to $\mathsf{O}(2)$. Consider now a system of n identical point vortices. A dihedral subgroup[4] $\mathbf{D}_n \subseteq \mathsf{O}(2) \times S_n$ acts on the phase space $\mathbb{C}^n$ by

$$\begin{aligned} c \cdot (z_1, \ldots, z_n) &= (cz_n, cz_1, \ldots, cz_{n-1}) \\ m \cdot (z_1, \ldots, z_n) &= (\bar{z}_{n-1}, \bar{z}_{n-2}, \ldots, \bar{z}_1, \bar{z}_n). \end{aligned} \tag{14.26}$$

where $c = \exp(2\pi \mathrm{i}/n)$ is a cyclic rotation and m is a mirror reflection. If the points $z_j = r_0 \exp(2\pi \mathrm{i} j/n)$ are placed as a regular ring, then that configuration is fixed by this $\mathbf{D}_n$, and m acts as a reflection in the line passing through z_n. In case n is odd, all reflections in $\mathbf{D}_n$ are conjugate, whereas in case n is even, there are two distinct conjugacy classes of reflections, one consisting of those through opposite vertices of a regular n-gon (all conjugate to m), the other consisting of those through

[4] S_n is the group of permutations of the n vortices and $\mathbf{D}_n$ has order $2n$.

mid-points of opposite edges (all conjugate to $m' = c\,m$). This will come into play in deciding what bifurcating solutions appear.

Since the $\mathbf{D}_n$-action fixes the ring, $\mathbf{D}_n$ acts on the tangent space to the phase space at this ring. The Fourier basis (14.19) for n point vortices is adapted to this action:

$$\begin{aligned} c\cdot\zeta_r^{(\ell)} &= \mathrm{e}^{-2\pi\mathrm{i}\ell/n}\,\zeta_r^{(\ell)} & m\cdot\zeta_r^{(\ell)} &= \bar{\zeta}_r^{(\ell)} \\ c\cdot\zeta_\theta^{(\ell)} &= \mathrm{e}^{-2\pi\mathrm{i}\ell/n}\,\zeta_\theta^{(\ell)} & m\cdot\zeta_\theta^{(\ell)} &= -\bar{\zeta}_\theta^{(\ell)}, \end{aligned} \tag{14.27}$$

and $m'\cdot\zeta_r^{(\ell)} = \mathrm{e}^{-2\pi\mathrm{i}\ell/n}\,\bar{\zeta}_r^{(\ell)}$, $m'\cdot\zeta_\theta^{(\ell)} = -\mathrm{e}^{-2\pi\mathrm{i}\ell/n}\,\bar{\zeta}_\theta^{(\ell)}$. For the 2-dimensional subspace V_1' of the symplectic slice $\mathscr{N}$ (14.22), write

$$\zeta' = (\sigma\alpha_r^{(1)} - \tilde{\sigma}\beta_\theta^{(1)}) + \mathrm{i}(\sigma\beta_r^{(1)} + \tilde{\sigma}\alpha_\theta^{(1)}) = \sigma\zeta_r^{(1)} + \mathrm{i}\tilde{\sigma}\zeta_\theta^{(1)}. \tag{14.28}$$

Then (14.27) implies that $c\cdot\zeta' = \mathrm{e}^{-2\pi\mathrm{i}/n}\,\zeta'$, $m\cdot\zeta' = \overline{\zeta'}$, and $m'\cdot\zeta' = \mathrm{e}^{-2\pi\mathrm{i}/n}\,\overline{\zeta'}$.

As the parameter λ varies, the eigenvalue $\epsilon_r^{(\ell)}$, $\ell \geqslant 2$ in (14.20) may cross 0 and may involve a bifurcation in the mode V_ℓ. In contrast $\epsilon_r^{(1)}$ and $\epsilon_\theta^{(\ell)}$, being strictly positive, never involve bifurcations and in particular the mode V_1' never bifurcates.

14.3.2 Dihedral Bifurcations

The type of bifurcation expected in a symmetric Hamiltonian system is controlled by the group action on the generalized kernel of the linear system at the bifurcation point. This was first investigated by Golubitsky and Stewart [6], cf. also [4]. It follows from [6] that in a *generic* family of linear Hamiltonian systems, a pair of eigenvalues come together along the imaginary axis, collide at the origin, and split along the real axis. This splitting transition is indeed what happens in our problem, as seen from the expressions (14.20) for the eigenvalues. Part of the *genericity* hypothesis of [6] is that the generalized kernel be an irreducible symplectic representation, which is satisfied here.

The greatest common divisor of n and ℓ will be denoted by (n,ℓ). We see from (14.27) that the cyclic subgroup $\mathbb{Z}_{(n,\ell)} \subset \mathbf{D}_n$ acts trivially on V_ℓ. Consequently on V_ℓ there is an effective action of $\mathbf{D}_{n/(n,\ell)}$. It turns out that V_ℓ is an irreducible symplectic representation of this group. Two cases are to be distinguished: $\ell = n/2$ (n even) when $V_{n/2}$ is 2-dimensional with an action of $\mathbf{D}_2 = \mathbb{Z}_2 \times \mathbb{Z}_2$, and $\ell \neq n/2$ when the V_ℓ s for $0 < \ell < n/2$ are all 4-dimensional. For bifurcations the modes $1 < \ell \leqslant n/2$ alone are of interest to us. Much of the analysis of generic bifurcations with dihedral symmetry is found in [7], and although there they deal with general vector fields rather than with Hamiltonian ones, the conclusions turn out to be the same. Analysis of the gradient case is also in [3].

In the dynamics of point vortices, if the bifurcating mode is $\ell \neq \lfloor n/2 \rfloor$, then the linear system[5] at the relative equilibrium has real eigenvalues in the $\lfloor n/2 \rfloor$ mode, hence is unstable. If $\ell = \lfloor n/2 \rfloor$, then the bifurcation involves a loss of stability in the mode ℓ. This means that in our analysis below, a local minimum corresponds to stable relative equilibria only if we are looking at the $\lfloor n/2 \rfloor$ mode.

Bifurcations on V_ℓ for $\ell = n/2$

This is the 2-dimensional symplectic span

$$\left\langle \zeta_r^{(n/2)}, \zeta_\theta^{(n/2)} \right\rangle,$$

with an action of $\mathbf{D}_2 \simeq \mathbb{Z}_2 \times \mathbb{Z}_2$. The kernel of the Hessian at the bifurcation point is the one-dimensional subspace spanned by $\zeta_r^{(n/2)}$, with a $\mathbb{Z}_2$-action. Then the generic bifurcation is a $\mathbb{Z}_2$-pitchfork, which can be either sub- or super-critical: if subcritical, the bifurcating solutions are unstable, and if supercritical, they are stable (provided the original "central" solution was stable). A normal form is given by the family

$$f_u(x, y) = -ux^2 \pm x^4 + y^2 + \text{h.o.t.} \tag{14.29}$$

"h.o.t." stands for higher-order terms in x^2, y^2, u. The $+$ sign in front of y^2 is justified by the eigenvalue in the $\zeta_\theta^{(n/2)}$-direction, which is always positive. The $-$ sign in front of u is a choice, dictated by the fact that increasing λr_0^2 makes a critical point pass from local minimum to saddle, as shown in Fig. 14.2. The sign $+$ or $-$ in front of x^4 corresponds to supercritical or subcritical, respectively. See Fig. 14.3a, b and the lecture notes [4] for a fuller discussion. In Remark 14.11 we explain why the bifurcations occurring here are in fact all supercritical, for all even n.

Bifurcations on V_ℓ for $\ell \neq n/2$

Of these only $\ell = \frac{1}{2}(n-1)$ involves stable relative equilibria, other modes bifurcate only if the linear system already has real eigenvalues; that said, the other values of ℓ do involve the appearance of new unstable relative equilibria, so are of interest. Write $k = n/(n, \ell)$. Then $k \geqslant 4$ and $\mathbf{D}_k$ acts effectively on V_ℓ. The type of generic bifurcation we get depends on k. The 4-dimensional V_ℓ is a direct

[5] The vector field, not the Hessian.

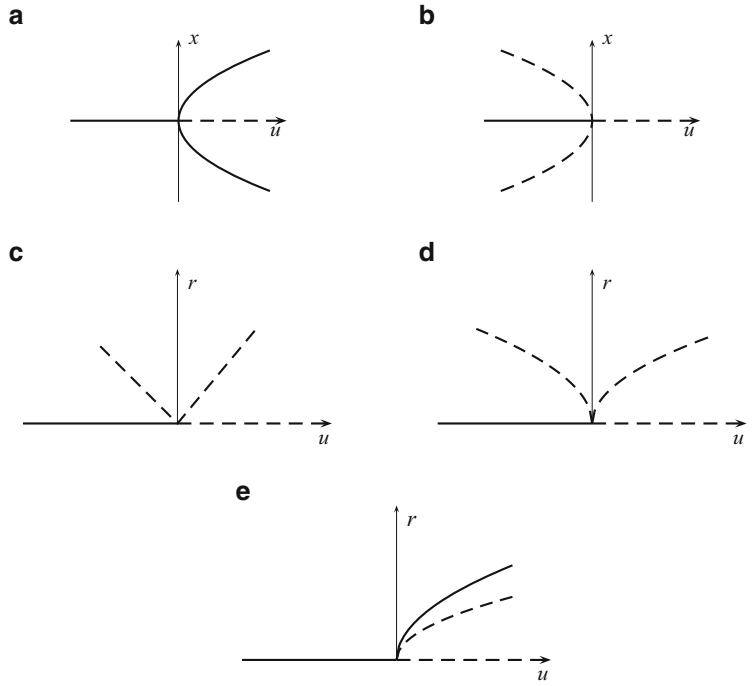

Fig. 14.3 Bifurcation diagrams for generic $\mathbf{D}_k$-bifurcations, $k \geqslant 2$. $r = \sqrt{x^2 + y^2}$, and u is the parameter as in the text. *Solid lines* refer to local minima, *dashed lines* to saddles and local maxima. For $k \geqslant 3$ each nontrivial branch corresponds to k solutions after applying the rotations from $\mathbf{D}_k$. Figure (**e**) is drawn for $\alpha > 1$, where α is the coefficient in (14.30); if $\alpha < -1$, reflect the diagram in the r-axis. (**a**) $\mathbf{D}_2$ supercritical pitchfork. (**b**) $\mathbf{D}_2$ subcritical pitchfork. (**c**) $\mathbf{D}_3$ (transcritical). (**d**) $\mathbf{D}_4$ transcritical ($|\alpha| < |\beta|$). (**e**) $\mathbf{D}_k$ pitchfork, $k \geqslant 4$ (with $|\alpha| > |\beta|$ for $k = 4$)

sum of two $\mathbf{D}_k$-invariant 2-dimensional Lagrangian subspaces, on one of which the Hessian vanishes at the bifurcation point, on the other it is always positive-definite. A $\mathbf{D}_k$-invariant function on such a space is a function of the fundamental invariants

$$N(x, y) = x^2 + y^2, \quad P(x, y) = \mathrm{Re}(x + \mathrm{i}y)^k.$$

A generic 1-parameter family of such functions is given by

$$f_u = -uN + \alpha N^2 + \beta P + \text{h.o.t.}, \tag{14.30}$$

where "h.o.t." stands for higher-order terms in N, P, u. Figure 14.4 shows the level sets of f_u for $k = 3, \ldots, 6$, as u varies through 0.

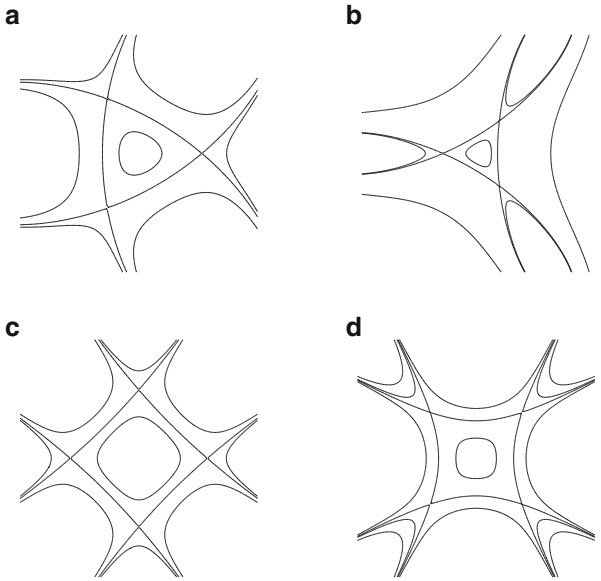

Fig. 14.4 Contours of the generic 1-parameter family of $\mathbf{D}_k$-invariant functions (14.30), for $k = 3, 4$ [produced with `Maple`, with a judicious choice of level sets]. The figures (**a**)–(**d**) are all transcritical bifurcations. (**a**) $\mathbf{D}_3$ with $u < 0$. (**b**) $\mathbf{D}_3$ with $u > 0$. (**c**) $\mathbf{D}_4$ with $|\alpha| < |\beta|$ and $u < 0$. (**d**) $\mathbf{D}_4$ with $|\alpha| < |\beta|$ and $u > 0$.

- If $k = 3$ or if $k = 4$ with $\beta > \alpha$, then the bifurcation can be said to be transcritical, in that the bifurcating branches exist on both sides of the bifurcation point $u = 0$, and the k bifurcating points are all saddles (and hence unstable equilibria), each of which is fixed by a mirror reflection conjugate to m. See the bifurcation diagrams in Fig. 14.3c, d.
- If $k = 4$ and $\alpha > \beta$ or if $k > 4$, then the bifurcation is like a pitchfork, in that all bifurcating equilibria coexist on the same side of the bifurcation point. But unlike the pitchfork, two types of bifurcating solutions appear, possibly with different stability properties; if k is even, then one has symmetry type $\langle m \rangle$ and the other $\langle m' \rangle$. See the bifurcation diagram in Fig. 14.3e.

Remark 14.5. The finite determinacy and unfolding theorems of singularity theory guarantee that f_0 in (14.29) and (14.30) is finitely determined and h.o.t. may be ignored. If $n \leqslant 3$, then f_0 has codimension 1 provided $\beta \neq 0$, and f_u is a versal unfolding of f_0, so that any deformation is equivalent to it. If $n \geqslant 4$, then f_0 has codimension 2, and a versal unfolding is

$$f_{u,v} = -uN + (\alpha + v)N^2 + \beta P$$

provided $\beta \neq 0$ (and $\alpha \neq \pm\beta$ when $n = 4$). The parameter v defines a *topologically* trivial deformation, i.e. v can be eliminated via a continuous change of coordinates rather than a smooth one; nevertheless this homeomorphism will be a

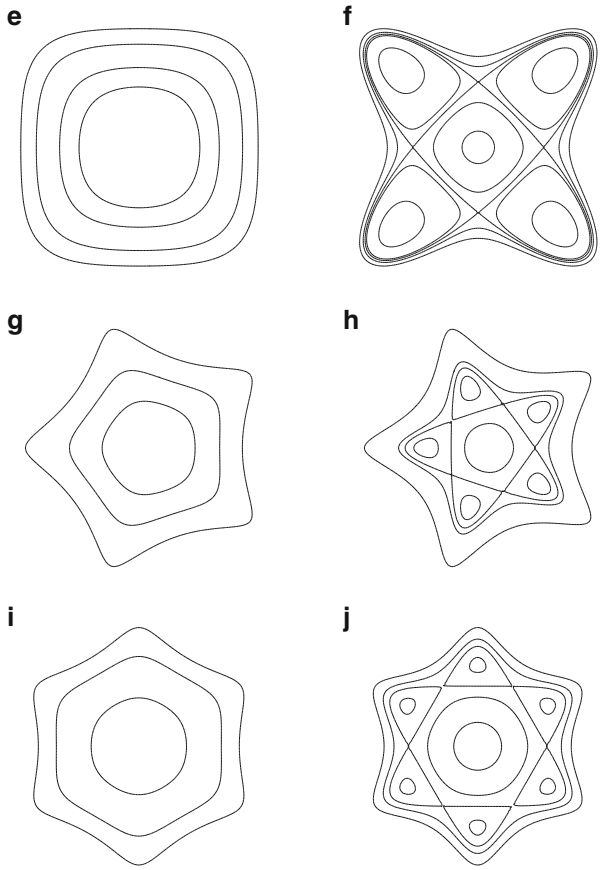

Fig. 14.4 (continued) Contours of the generic 1-parameter family of $\mathbf{D}_k$-invariant functions (14.30), for $k \geqslant 4$. These figures are all dihedral pitchfork bifurcations. (**e**) $\mathbf{D}_4$ with $|\alpha| > |\beta|$, $u < 0$. (**f**) $\mathbf{D}_4$ with $|\alpha| > |\beta|$, $u > 0$. (**g**) $\mathbf{D}_5$ with $u < 0$. (**h**) $\mathbf{D}_5$ with $u > 0$. (**i**) $\mathbf{D}_6$ with $u < 0$. (**j**) $\mathbf{D}_6$ with $u > 0$

diffeomorphism away from the origin, so critical points are preserved. The modulus v that arises when $n = 4$ is related to the cross-ratio of the four lines making up $f^{-1}(0)$.

14.3.3 Bifurcations of Vortex Rings

We work with the parameter $\lambda r_0^2 > -1$. Recall that r_0 is the radius of the vortex ring measured on $\mathbb{C}$ after the stereographic projection, and that it is related by (14.7) to the radius a measured on $\mathscr{M}_\lambda$. The 0 curvature case is $\lambda = 0$, and in the spherical case $\lambda > 0$ the values λr_0^2 and $1/\lambda r_0^2$ are equivalent as they represent antipodal rings on the sphere. We therefore let λr_0^2 vary in the range $(-1, 1]$.

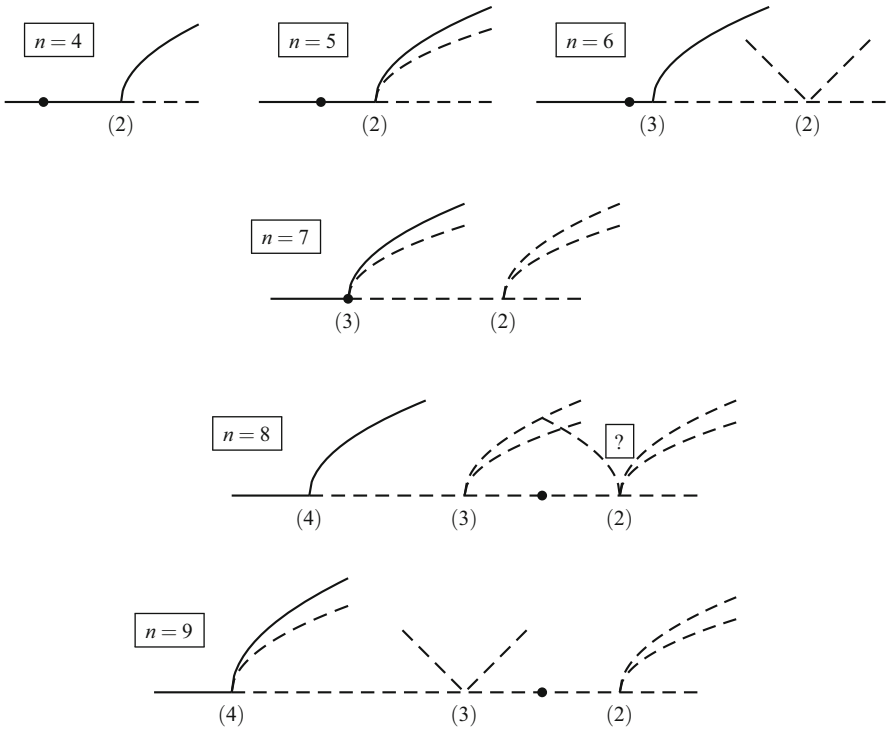

Fig. 14.5 Bifurcation diagrams for the ring of n identical vortices for low values of n. The number in parentheses is the mode number bifurcating at that point. The *black dot* on the axis represents schematically the point where $\lambda = 0$ (the plane). λ increases toward right. We do not know whether the bifurcating branches from the lower modes branch to the right or the left, though we believe they are as shown. The case $n = 8$, $\ell = 2$ has an effective action of $\mathbf{D}_4$, so could be transcritical or pitchfork—we do not know which occurs

Now for the main theorem. Stability means Lyapunov stability modulo rotations (same as orbital stability in our situation). Instability means full spectral instability, i.e. at least one of the eigenvalues is real and positive. The ring of $n = 3$ vortices is always nonlinearly stable.

Theorem 14.6. *Let $n \geqslant 4$. With $\lambda r_0^2 \in (-1, 1]$ as a parameter, the regular ring of n identical vortices undergoes the following bifurcations, illustrated in Fig. 14.5:*

all n *The ring is stable for $\lambda r_0^2 < b_n$, where b_n is the unique root in $(-1, 1]$ of*[6]

$$\frac{1 + b_n^2}{(1 + b_n)^2} = \frac{1}{2(n-1)} \left\lfloor \frac{n^2}{4} \right\rfloor. \tag{14.31}$$

[6]Cf. (14.25).

n even As λr_0^2 crosses b_n, the ring loses stability via a supercritical pitchfork bifurcation, and the (stable) bifurcating solution consists of a pair of $n/2$-gons with different values for the radius.

n odd As λr_0^2 crosses b_n, the ring loses stability via a supercritical bifurcation as depicted in Fig. 14.3e, to 2 types of relative equilibria, each with a line of symmetry.

all n As λr_0^2 increases further, the ring undergoes a sequence of bifurcations, one in each of the modes $\lfloor n/2 \rfloor > \ell > 1$; all the relative equilibria involved are unstable, and the bifurcating solutions have $\mathbf{D}_{(n,\ell)}$-symmetry.[7]

As a special case, we recover the following result of Kurakin and Yudovich [11] and Schmidt [20]. The calculation justifying it is the subject of Sect. 14.3.5.2.

Corollary 14.7. *The Thomson heptagon is nonlinearly stable.*

Bifurcation Values of λr_0^2

The tables below spell out the values of λr_0^2 where the bifurcations occur, for $n = 6, 7, 8, 9$. They are found by solving $\epsilon_r^{(\ell)} = 0$ (14.20) for λr_0^2; the first values are those of b_n mentioned in Theorem 14.6.

$n = 6$

mode	$\ell = 3$	$\ell = 2$
λr_0^2	0.056	0.127

$n = 7$

mode	$\ell = 3$	$\ell = 2$
λr_0^2	0	0.101

$n = 8$

mode	$\ell = 4$	$\ell = 3$	$\ell = 2$
λr_0^2	−0.063	−0.033	0.084

$n = 9$

mode	$\ell = 4$	$\ell = 3$	$\ell = 2$
λr_0^2	−0.101	−0.056	0.072

In all the tables, the bifurcation of the $\ell = 2$ mode occurs for $\lambda > 0$ (on the sphere); this is easily checked to be true for all n.

14.3.4 Geometry of Bifurcating Rings

At a bifurcation, the bifurcating mode controls the geometry/symmetry of the bifurcating solution. Points in V_ℓ all correspond to configurations with cyclic

[7] We put $\mathbf{D}_1 = \mathbb{Z}_2$ acting by reflection.

symmetry $\mathbb{Z}_{(n,\ell)} \subset \mathbf{D}_n$, which allows the $\mathbf{D}_n$-action on V_ℓ to factor through a $\mathbf{D}_n/\mathbb{Z}_{(n,\ell)} \simeq \mathbf{D}_k$-action, where

$$k = n/(n,\ell).$$

Moreover, the bifurcating solutions are all fixed by a reflection in $\mathbf{D}_k$ (conjugate to m or to m'), which implies that they are symmetric under a $\mathbb{Z}_{(n,\ell)}$-action and under a reflection, together giving a symmetry of $\mathbf{D}_{(n,\ell)}$. This means that if $(n,\ell) > 1$, then the configuration consists of k rings of (n,ℓ) vortices in each. Typical deformed configurations with the correct symmetry in each mode for $n = 3,\ldots,8$ are shown in Fig. 14.6.

In particular, when n is even and $\ell = n/2$, the solutions have symmetry isomorphic to $\mathbf{D}_{n/2}$. Now in $\mathbf{D}_n$ there sit 2 non-conjugate copies of $\mathbf{D}_{n/2}$, one containing m, the other containing m', and since $\zeta_r^{(n/2)}$ is fixed by m, the bifurcating solutions must have the symmetry $\mathbf{D}_{n/2}$ containing m. Consequently the bifurcating solution consists of a pair of regular $n/2$-gons, in general of different radii, staggered by $2\pi/n$ as shown in Fig. 14.6a, d, k.

When $\ell \neq n/2$, the $\mathbf{D}_n$-action factors through a $\mathbf{D}_k$-action, and all bifurcating solutions have reflexive symmetry of m or m' (as seen from Fig. 14.4). Now if k is odd, the resulting reflections are conjugate, so a configuration fixed by m will also be fixed by some conjugate of m'. This fact is illustrated in Fig. 14.6c where $n = 6$, $\ell = 2$, $k = 3$. Indeed, as $\mathbf{D}_n$ has n reflections while $\mathbf{D}_k$ has k, in the representation $\mathbf{D}_n \to \mathbf{D}_k$ we must have (n,ℓ) reflections in $\mathbf{D}_n$ that get identified, thereby fixing the same configurations. On the other hand, if k is even (as in $n = 8$, $\ell = 2$, Fig. 14.6g, h, the nonconjugate m and m' in $\mathbf{D}_n$ have as images 2 nonconjugate reflections in $\mathbf{D}_k$, so the latter's fixed-point sets correspond to different configurations.

Finally, whenever ℓ divides n, a perturbation in the $\zeta_r^{(\ell)}$-direction produces n/ℓ rings of ℓ-gons, in general of slightly different radii, and in the configurations with reflexive symmetry m the vortices in the different ℓ-gons line up with the original n-gon.

14.3.5 Degenerate Critical Points

This section is dedicated to proving Theorem 14.6. We begin by presenting a criterion for a degenerate critical point to be a local minimum, then in Sects. 14.3.5.1 and 14.3.5.2 respectively apply the criterion to the cases where n is even and n is odd.

Lemma 14.8. *Let f be an analytic function defined on a neighbourhood of 0 in $\mathbb{R}^n$ with a degenerate critical point at 0, such that $f(0) = 0$. Write $B = \mathrm{d}^2 f(0)$, $C = \mathrm{d}^3 f(0)$, $D = \mathrm{d}^4 f(0)$. If B is positive-semidefinite and if for all $\mathbf{a} \in \mathrm{Ker}\, B \setminus \{0\}$, $\mathbf{b} \in \mathbb{R}^n$ we have*

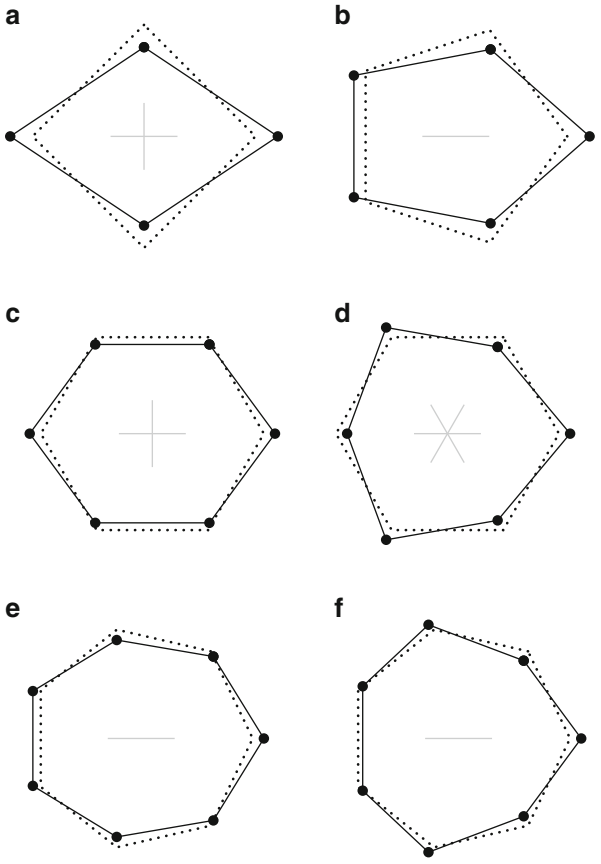

Fig. 14.6 Perturbations of the n-ring in mode ℓ. These configurations are invariant under a subgroup isomorphic to $\mathbf{D}_{(n,\ell)}$, and the *grey lines* in the centre of each represent the lines of reflection. The *dotted figures* are the regular n-gons. See Sect. 14.3.4 for explanations. (**a**) $n = 4,\ \ell = 2$. (**b**) $n = 5,\ \ell = 2$. (**c**) $n = 6,\ \ell = 2$. (**d**) $n = 6,\ \ell = 3$. (**e**) $n = 7,\ \ell = 2$. (**f**) $n = 7,\ \ell = 3$. (**g**) $n = 8,\ \ell = 2$ (Fix m). (**h**) $n = 8,\ \ell = 2$ (Fix m'). (**i**) $n = 8,\ \ell = 3$ (Fix m). (**j**) $n = 8,\ \ell = 3$ (Fix m'). (**k**) $n = 8,\ \ell = 4$

$$C\mathbf{a}^3 = 0, \quad D\mathbf{a}^4 + 6C\mathbf{a}^2\mathbf{b} + 3B\,\mathbf{b}^2 > 0,$$

then f is strictly positive on a punctured neighbourhood of 0.

The symbol like $C\mathbf{a}^2\mathbf{b}$ means "evaluate the trilinear form C at $\mathbf{a}$ in 2 of its 3 arguments and at $\mathbf{b}$ in the 1 remaining argument". In our application of this lemma, f will be the augmented Hamiltonian $\tilde{H}_\lambda$ (Sect. 14.2.2), and for the point-vortex problem this is analytic.

Proof. Suppose for a contradiction that $f^{-1}(0)$ intersects every punctured neighbourhood of 0.

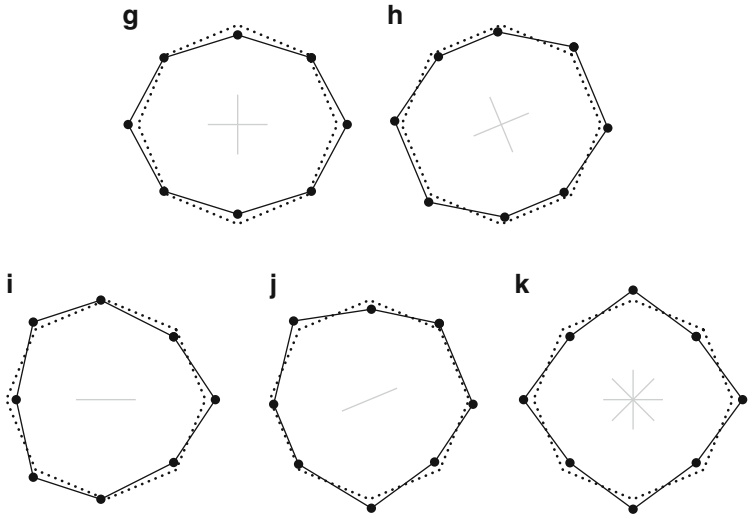

Fig. 14.6 (continued)

Use the splitting lemma (e.g. [19]) to write

$$f(x, y) = Q(x) + h(y),$$

where Q is a homogeneous quadratic form and h is a function with vanishing 2-jet. Explicitly, $Q(x) = \frac{1}{2} x^T B\, x$ and $y \in \operatorname{Ker} B$. Since $y \in \operatorname{Ker} B$, the hypothesis $C\mathbf{a}^3 = 0$ implies that h in fact has vanishing 3-jet. Since the hypotheses of the lemma on the derivatives are ***intrinsic*** (invariant under change of coordinates), in the new coordinates x, y they become $\mathrm{d}^3 h = 0$ and $\mathrm{d}^4 h\, \mathbf{a}^4 > 0$ for all $\mathbf{a} \in \operatorname{Ker} B \setminus \{0\}$.

Use the curve selection lemma [5, 9] to deduce the existence of an analytic curve $\gamma(t)$ passing through 0 along which f vanishes. In the new coordinates, write $\gamma(t) = (x, y) = (\xi(t), \eta(t))$; then $Q(\xi(t)) + h(\eta(t)) = 0$. Expand ξ, η in Taylor series:

$$\xi(t) = \xi_1 t^r + \xi_2 t^{r+1} + \cdots, \quad \eta(t) = \eta_1 t^r + \eta_2 t^{r+1} + \cdots,$$

where r is the order of the curve (at least one of ξ_1, η_1 is nonzero). The leading terms of $f \circ (\xi, \eta)$, all of which must vanish, are

$$Q(\xi(t)) = \tfrac{1}{2} B \xi_1^2\, t^{2r} + B \xi_1 \xi_2\, t^{2r+1} + \cdots$$

By inspection this forces $\xi_1 = \cdots = \xi_r = 0$. The coefficient of t^{4r} is then $B\xi_{r+1}^2 + \frac{1}{4!} \mathrm{d}^4 h\, \eta_1^4$, which must vanish. But the order of the curve being r and $\xi_1 = 0$, we must have $\eta_1 \neq 0$, hence from the hypothesis $B\xi_{r+1}^2 + \frac{1}{4!} \mathrm{d}^4 h\, \eta_1^2 > 0$, a contradiction. □

Consider now any Hamiltonian system on any symplectic manifold M, with Hamiltonian H, symmetry group G, momentum map $\mathbf{J} : M \to \mathfrak{g}^*$, and suppose $x \in M$ lies on a relative equilibrium with finite stabilizer (possibly trivial). Define $T_0 := \mathfrak{g}\cdot x \cap \operatorname{Ker} \mathrm{d}\mathbf{J}$. The symplectic slice $\mathscr{N}$ is then any G_x-invariant complement to T_0 in $\operatorname{Ker} \mathrm{d}\mathbf{J}$.

Proposition 14.9. *If the hypotheses of Lemma 14.8 are satisfied with $\mathbb{R}^n$ replaced by $\mathscr{N}$ and f by $\hat{H}$, then the relative equilibrium is Lyapunov stable.*

Proof. It is enough to show that in the reduced space M_μ, $\mu = \mathbf{J}(x)$, the reduced Hamiltonian admits a local extremum at x, cf. [17]. Since the action is locally free, $\mathbf{J}$ is a submersion near x. The tangent space of the submanifold $\mathbf{J}^{-1}(\mu)$ is $T_0 \oplus \mathscr{N}$, and $\mathrm{d}\mathbf{J}_\lambda(x)$ maps $\mathscr{N}$ isomorphically to $T_{\bar{x}}M_\mu$, where $\bar{x}$ is the image of x in M_μ. Let $\bar{\gamma}$ be a curve in M_μ through $\bar{x}$. Then $\bar{\gamma}$ lifts to a smooth curve in M tangent to $\mathscr{N}$ at x. The claim now follows because the hypotheses of Lemma 14.8 applied on M_μ are equivalent to the same hypotheses applied on $\mathscr{N}$. □

As the argument and calculations for the stability of the bifurcating points take distinct turns depending on the parity of n, we treat the even and odd cases separately. The even case is fairly easy, the odd case is much harder.

14.3.5.1 n Even

The critical mode is $\ell = n/2$, we have $\zeta^{(n/2)} = \alpha^{(n/2)}$, and both c and m act by multiplication by -1. This means that on this Fourier mode $V_{n/2} = \left\langle \alpha_r^{(n/2)}, \alpha_\theta^{(n/2)} \right\rangle$, the augmented Hamiltonian $\hat{H}_\lambda$ is an even function. We therefore expect, for generic families of functions, $\hat{H}_\lambda$ restricted to $V_{n/2}$ to be equivalent to a family of the form

$$f_u(x, y) = \pm x^4 + ux^2 + y^2,$$

where u is a parameter depending on r_0, λ. The $+$ sign in front of the y^2 term is justified because the y-direction here corresponds to $\alpha_\theta^{(\ell)}$, whose eigenvalue from (14.20) is $\kappa^2 n^2/16\pi > 0$.

The $\mathbf{D}_n$-invariance of $\hat{H}_\lambda$ helps us to figure out which terms arise in its Taylor series. For example, if $f \in V_\ell^*$, $g \in V_m^*$, and fg is invariant, then $m = \ell$.

Lemma 14.10. *For all $\mathbf{a} \in V_{n/2}$ and $\mathbf{b} \in \mathscr{N}$, we have $C\mathbf{a}^2\mathbf{b} = 0$.*

Proof. On the symplectic slice expand $\hat{H}_\lambda$ in Taylor series. Each term is invariant, in particular the 3rd-order term $C\mathbf{x}^3$ for $\mathbf{x} \in \mathscr{N}$. Given $\mathbf{a}_i \in V_{\ell_i}$ $(i = 1, 2, 3)$, the quantity $C\mathbf{a}_1\mathbf{a}_2\mathbf{a}_3$ lies in the tensor product $V_{\ell_1} \otimes V_{\ell_2} \otimes V_{\ell_3}$, which contains invariant functions if and only if $\ell_1 \pm \ell_2 \pm \ell_3 \equiv 0 \bmod n$. For $\mathbf{a}_1 = \mathbf{a}_2 = \mathbf{a} \in V_{n/2}$,

$C\mathbf{a}^2\mathbf{b}$ (necessarily invariant) can be nonzero only if $\mathbf{b} \in V_0$. But $V_0 \cap \mathscr{N} = \{0\}$, implying that if $\mathbf{b} \in \mathscr{N}$ and $C\mathbf{a}^2\mathbf{b}$ is invariant, then $\mathbf{b} = 0$, so that $C\mathbf{a}^2\mathbf{b} = 0$. □

It remains to calculate $D\mathbf{a}^4$ for $\mathbf{a} \in V_{n/2}$ in order to apply Proposition 14.9. The criterion for a bifurcation (14.31) reads, for even n,

$$\frac{1+\lambda^2 r_0^4}{(1+\lambda r_0^2)^2} = \frac{n^2}{8(n-1)}. \tag{14.32}$$

Put $z_j = (r_0 + (-1)^j\, t)\mathrm{e}^{2\pi \mathrm{i} j/n}$ and $f(t) = \hat{H}_\lambda(z_1, \ldots, z_n)$. We shall expand

$$f(t) = -\frac{1}{4\pi}\sum_{i<j} \log|z_i - z_j|^2 + \frac{n-1}{4\pi}\sum_j \log(1+\lambda|z_j|^2) - \omega \sum_j \frac{|z_j|^2}{1+\lambda|z_j|^2}.$$

to the 4th order in t.

Calculation

At the bifurcation point, r_0 and λ are related by (14.32) and $\omega = \omega_0$ is given by (14.15). We are expanding

$$\begin{aligned} f(t) = &-\frac{n}{8\pi} \sum_{1\leqslant k\leqslant n-1,\ k \text{ odd}} \log\left(r_0^2 + t^2 - (r_0^2 - t^2)\cos(2\pi k/n)\right) \\ &-\frac{n(n-2)}{32\pi}\left(\log(r_0+t)^2 + \log(r_0-t)^2\right) \\ &+\frac{n(n-1)}{8\pi}\left(\log(1+\lambda(r_0+t)^2) + \log(1+\lambda(r_0-t)^2)\right) \\ &-\omega_0 \frac{n}{2}\left(\frac{(r_0+t)^2}{1+\lambda(r_0+t)^2} + \frac{(r_0-t)^2}{1+\lambda(r_0-t)^2}\right) \quad + \quad \cdots \end{aligned}$$

where $\cdots$ is a constant independent of r_0, t, n, which will henceforth be ignored. Taking Taylor series in t of all of these terms to order 4 is simple, except for the first line, which comes out as

$$-\frac{n^2}{8\pi}\log r_0 - \frac{n^2(n-2)}{32\pi r_0^2}t^2 + \frac{n^2(n-2)(n^2+2n-12)}{768\pi r_0^4}t^4 + O(t^6)$$

(up to an additive constant), thanks to identities akin to (14.18). The coefficient of t^2 in the Taylor series is then

$$\frac{n}{32\pi r_0^2\sigma^2}\left(-(n-2)^2\sigma^2+4(n-1)(1-\lambda r_0^2)^2\right)$$

which can be shown to vanish subject to the bifurcation relation (14.32). The coefficient of t^4 is

$$\frac{n}{768\pi r_0^4\sigma^4}\ \big[(n-2)(n^3+2n^2-12n+24)\sigma^4 \\ +24(n-1)\lambda r_0^2(19\lambda^3 r_0^6-54\lambda^2 r_0^4+43\lambda r_0^2-4)\big].$$

Denote by T the term in square brackets. We wish to show that $T>0$ at the bifurcation point. Solving (14.32) for λr_0^2 yields 2 roots, substituting which into T in turn yields 2 values, say T_1, T_2 (functions of n). Write $m=n-2$. A calculation (using `Maple`!) reveals that T_1+T_2 is equal to

$$\frac{128(m+1)}{(m^2-4m-4)^4}\ \big(384+2560m+4m^9+4832m^2+5024m^3+10616m^4 \\ +15888m^5+10778m^6+3266m^7+177m^8\big),$$

while T_1T_2 is equal to

$$\frac{4096(m+1)^2}{(m^2-4m-4)^4}\ \big(16m^{10}+648m^9+7409m^8+1044m^7+39960m^6+85512m^5 \\ +57332m^4+25824m^3+15136m^2+5376m+576\big).$$

In view of $m\geqslant 0$ and $n\geqslant 2$, both are manifestly strictly positive, so that each of T_1 and T_2 is indeed strictly positive. □

Remark 14.11. Since the relative equilibrium is stable at the point of bifurcation, the resulting pitchfork bifurcations are ***supercritical***: the bifurcating relative equilibria are stable and coexist with the unstable central one. Thus these stable bifurcating relative equilibria exist in a neighbourhood of the bifurcation point, λr_0^2 satisfying (14.25) with the inequality reversed. See also Fig. 14.2.

To persuade ourselves that this is a genuine pitchfork, we need to check the nondegeneracy condition which is that the eigenvalues of the Hessian move through 0 at nonzero speed with respect to the parameter λr_0^2 (or just λ or r_0 separately). The expression (14.20) for the eigenvalues permits an easy check.

As the mode that bifurcates is $\ell=n/2$, the bifurcation occurs in the fixed-point space for the subgroup $\mathbf{D}_{n/2}$ as explained in Sect. 14.3.4. The bifurcating solutions have $\mathbf{D}_{n/2}$-symmetry, i.e. consist of 2 regular $n/2$-gons at slightly different radii from the common centre, and these bifurcating solutions are ***stable***, at least close to the bifurcation point.

14.3.5.2 n Odd

There are two reasons why n odd is much harder than n even. First, $\operatorname{Ker} B$ (degeneracy space) is 2-dimensional and its basis elements are less simple (namely $\zeta_r^{(n-1)/2}$ rather than $\zeta_r^{(n/2)}$). Second, the third derivative contributions are nonzero and no analogue of Lemma 14.10 holds. We proceed as far as we can with general odd n, and then specialize to numerical calculations for a few low values of n.

The criterion for a bifurcation (14.31) reads, for odd n,

$$\frac{1+\lambda^2 r_0^4}{(1+\lambda r_0^2)^2} = \frac{n+1}{8}. \tag{14.33}$$

The critical mode is $\ell = \frac{1}{2}(n-1)$, and

$$V_c := \operatorname{Ker} B = \left\langle \alpha_r^{((n-1)/2)},\ \beta_r^{((n-1)/2)} \right\rangle \subset V_{(n-1)/2}.$$

We wish to apply Proposition 14.9, based on Lemma 14.8 with $\mathbf{a} \in V_c$ and $\mathbf{b} \in \mathscr{N}$. The calculations are simplified by the following observations. Recall the definition of V_1' from (14.22).

Proposition 14.12. *1. No cubic invariant exists on V_c, consequently $C\mathbf{a}^3 = 0$.*
2. Up to scalar multiple, there exists a unique quartic invariant on V_c, consequently $D\mathbf{a}^4$ is a multiple of $|\mathbf{a}|^4$.
3. Up to scalar multiple, there exists a unique cubic invariant of the form $C\mathbf{a}^2\mathbf{b}$ with $\mathbf{a} \in V_c$ and $\mathbf{b} \in \mathscr{N}$, and invariance forces $\mathbf{b} \in V_1'$.

Proof. Because $\frac{1}{2}(n-1)$ is coprime to n, the action of $\mathbf{D}_n$ on V_c is equivalent to the usual representation of $\mathbf{D}_n$ in the plane, though with an unusual choice of generator, cf. (14.27). $\mathbf{D}_n$-invariant functions on V_c are functions of $N = x^2 + y^2$ and $P = \operatorname{Re}(x + \mathrm{i}y)^n$, cf. comment just before (14.30). Write $\mathbf{a} = (x, y)$.

(i) As $n \geqslant 5$, this representation accommodates no cubic invariants, and as $C\mathbf{a}^3$ must be invariant, it is 0.
(ii) Likewise, the unique quartic invariant on V_c is N^2, so $D\mathbf{a}^4$ is a scalar multiple of $N^2 = |\mathbf{a}|^4$.
(iii) If $\mathbf{a} \in V_c$ and $\mathbf{b} \in V_m$, then $C\mathbf{a}^2\mathbf{b} \in V_{2\cdot(n-1)/2+m} \oplus V_{2\cdot(n-1)/2-m} \oplus V_m$. For this to be invariant, we need $m = 0$ or $m = 1$. The former is ruled out by the assumption $\mathbf{b} \in \mathscr{N}$, so $\mathbf{b} \in \mathscr{N} \cap V_1 = V_1'$. □

To understand better the invariant $C\mathbf{a}^2\mathbf{b}$, let x, y be as before on V_c and u, v be coordinates on V_1', chosen so that $m \cdot (x, y) = (x, -y)$ and $m \cdot (u, v) = (u, -v)$; in a nutshell $m \cdot (z, w) = (\bar{z}, \bar{w})$ in terms of complex variables $z = x + \mathrm{i}y$ and $w = u + \mathrm{i}v$. We then have $c \cdot (z, w) = (c^{(n-1)/2}z, cw)$. The cubics of the form $C\mathbf{a}^2\mathbf{b}$ are the real and imaginary parts of z^2w, $z^2\bar{w}$, $|z|^2w$. However, only the first of these is invariant under c, and only its real part is invariant under m. Thus,

$$C\mathbf{a}^2\mathbf{b} = \gamma\,(z^2 w + \bar{z}^2 \bar{w}) = 2\gamma\,\big(u(x^2 - y^2) - 2vxy\big)$$

for some value of $\gamma \in \mathbb{R}$; explicitly $\gamma = \frac{1}{4}\frac{\partial^3}{\partial x^2 \partial u}(C\mathbf{a}^2\mathbf{b}) = \frac{3}{2}\frac{\partial^3}{\partial x^2 \partial u}\hat{H}_\lambda$.

The key quantity $D\mathbf{a}^4 + C\mathbf{a}^2\mathbf{b} + B\mathbf{b}^2$ becomes, on completing the square,

$$\delta|z|^4 + 2\gamma\,\mathrm{Re}(z^2 w) + \beta|w|^2 = \delta\left|z^2 + \frac{\gamma}{\delta}\bar{w}\right|^2 + \frac{\beta\delta - \gamma^2}{\delta}|w|^2. \tag{14.34}$$

Manifestly this is positive for all $z, w \neq 0$ if and only if $\delta > 0$ and $\beta\delta > \gamma^2$. The value of β is

$$\beta = \epsilon_1' = \frac{n(n-1)\kappa^2}{4\pi r_0^2}\tilde{\sigma}^2$$

where as before $\tilde{\sigma} = 1 - \lambda r_0^2$, cf. (14.24). There remains the task of calculating γ and δ.

Calculation

The awkward trigonometric expressions prevented us (and `Maple`) from reaching closed forms for γ and δ. We therefore proceed to evaluate them numerically. In all these evaluations, r_0 is related to the parameter λ by (14.33). For $n = 7$, of course $\lambda = 0$ and r_0 is arbitrary.

$n = 5$:

$$\beta = \frac{37.1}{r_0^2}, \quad \gamma = -\frac{9.04}{r_0^3}, \quad \delta = \frac{15.8}{r_0^4}.$$

$\beta\delta - \gamma^2 = 504.7/r_0^6 > 0$ hence the pentagon is stable.

$n = 7$:

It transpires (`Maple`) that for $n = 7$ we have

$$\beta = \frac{21}{2\pi r_0^2}, \quad \gamma = \frac{63}{4\pi r_0^3}, \quad \delta = \frac{1071}{8\pi r_0^4}.$$

Computationally these numbers are correct to a high degree of precision—but we have no proof that they are rational multiples of $1/\pi r_0^k$. At any rate it is certain that $\beta\delta - \gamma^2 > 0$, hence the heptagon is stable. This establishes Corollary 14.7.

$n = 9$:

$$\beta = \frac{6.9}{r_0^2}, \quad \gamma = \frac{13.8}{r_0^3}, \quad \delta = \frac{182.4}{r_0^4}.$$

$\beta\delta - \gamma^2 = 1075.60/r_0^4 > 0$ hence the enneagon[8] is stable.

$n = 11$:

$$\beta = \frac{12.0}{r_0^2}, \quad \gamma = \frac{29.7}{r_0^3}, \quad \delta = \frac{555.1}{r_0^4}.$$

$\beta\delta - \gamma^2 = 5787.6/r_0^6 > 0$ hence the hendecagon[9] is stable.

Conjecture 14.13. For all values of n, the n-gon is stable at the bifurcation point.

Above, we have proved this for all even n and for $n = 3, 5, 7, 9, 11$.

14.3.6 Bifurcations from the Equator

For all values of $n > 3$, the $\ell = 1$ mode "bifurcates" at $\lambda r_0^2 = 1$, i.e. when the ring of vortices lies on the equator of the sphere. The momentum value at such a ring is fixed by all of $\mathsf{SO}(3)$: indeed, for the usual coadjoint equivariant momentum map for this problem, the momentum value is 0. Let us summarize from [14, Proposition 3.8] what bifurcations occur in this situation. On each near-zero momentum sphere we get, besides the regular ring of n vortices, the following configurations.

n odd For each of the n planes through the poles of the sphere and containing one of the vortices, 2 configurations consisting of $\frac{1}{2}(n-1)$ pairs and that 1 vortex on the plane; the vortices in each pair are each other's reflection in that plane. The notation in [14] is $C_h(\frac{1}{2}(n-1)R,\, E)$, E referring to the single vortex on the plane, the R to the reflection pairs.

n even In this case there are two distinct types of bifurcating solution, arising from the two distinct types of reflection in D_n:

1. For each of the $\frac{1}{2}n$ planes through the poles and containing a pair of diametrically opposite vortices, 1 configuration consisting of $\frac{1}{2}n - 1$ reflection pairs and those 2 vortices on the plane. The notation in [14] is $C_h((\frac{1}{2}n - 1)R,\, 2E)$.

[8] "Nonagon" mixes Latin and Greek.

[9] "Undecagon" is another Greco-Latin hybrid.

2. For each of the $\frac{1}{2}n$ planes through the poles and passing midway between adjacent vortices, 1 configuration consisting of $\frac{1}{2}n$ reflection pairs. The notation in [14] is $C_h(\frac{1}{2}n\,R)$.

14.4 What Happens with Other Hamiltonians

Toward the end of Sects. 14.1.2 and 14.1.3, we met three options for families of Green's functions, which all agree for the plane $\lambda = 0$. We have been opting for (14.12). Here we sketch the conditions that guarantee the stability of the ring of identical vortices for the other two options (14.10), (14.11); the methods are the same as those of Sects. 14.2 and 14.3.

14.4.1 Green's Function $G = \log|z - w|^2$

Angular Velocity

The regular ring of n vortices with identical vorticities κ rotates at angular velocity

$$\omega = -\frac{(n-1)\kappa}{8\pi}\frac{\sigma^2}{r_0^2}.$$

Unlike the expression (14.15) this is not even in λ.

Stability

The Hessian of the augmented Hamiltonian is the same as (14.16), except that A is changed to

$$A = -\frac{(n-1)\kappa^2}{24\pi r_0^2\sigma}\left((n-11) + (n+13)\lambda r_0^2\right)$$

and the eigenvalues of the Hessian become

$$\epsilon_r^{(\ell)} = \frac{\kappa^2}{4\pi r_0^2}\left((n-1)\frac{6+5\lambda^2 r_0^4}{3\sigma^2} - \ell(n-\ell)\right)$$

while the expressions for $\epsilon_\theta^{(\ell)}$ are as before. Also as before, the $\ell = \lfloor n/2 \rfloor$ mode has the least eigenvalue, so the ring is stable provided $\epsilon_r^{\lfloor n/2 \rfloor} > 0$. The criterion is

$$\frac{1+\frac{5}{6}\lambda^2 r_0^4}{(1+\lambda r_0^2)^2} > \frac{1}{2(n-1)}\left\lfloor \frac{n^2}{4} \right\rfloor.$$

Compared with (14.25), for each n this new inequality is satisfied by a slightly narrower range of the effective parameter λr_0^2. The transition from stable to unstable still occurs at λr_0^2 of the same sign as for the previous Hamiltonian, and for $\lambda = 0$ the two Hamiltonians agree. Hence the Thomson heptagon is still stable.

Bifurcations

It seems likely that the bifurcations are of the same types as those explained in Sect. 14.3; we have checked this for $n = 5, 7, 9$.

14.4.2 Green's Function $G = \log \frac{|z-w|^2}{|1+\lambda z\bar{w}|^2}$

On the hyperbolic plane $\lambda < 0$ this reduces to the Hamiltonian adopted by Kimura [10]. For $\lambda = 0$ it is the standard Green's function on the plane, while for $\lambda > 0$ it corresponds to Green's function for the Laplacian on the sphere with "counter-vortices". The Hamiltonian will model $2n$ vortices placed pairwise at antipodal points, each pair having opposite vorticities. Thus the "ring" becomes 2 rings, one of n vortices of vorticity κ near the North Pole, the other of n vortices of vorticity $-\kappa$ near the South Pole. In Laurent-Polz [12] these configurations are referred to as $\mathbf{D}_{nh}(2R)$ when n is even and $\mathbf{D}_{nd}(R, R')$ when n is odd (in the former the rings are aligned, while in the latter they are staggered). The stability results of [12] are not directly applicable here, as he considers stability with respect to perturbations of all $2n$ vortices, whereas we are considering a restricted class of perturbations: those preserving the antipodal pairing of the configurations. If a configuration is stable for Laurent-Polz, then *a fortiori* it will be stable for our setting.

The calculations based on this option of Green's function get so cumbersome that the stability problem seems no longer tractable analytically. It seems likely that the results are similar to those in Sect. 14.3, though the details of where the bifurcations occur will differ. We did calculate that the angular velocity ω analogous to (14.15) of the ring is

$$-\frac{\kappa}{8\pi r_0^2}\frac{1+\lambda r^2}{1-(-\lambda r^2)^n}\left((n-1)(1+\lambda r^2)\left(1+\left(-\lambda r^2\right)^{n-1}\right)+2\lambda r^2\left(1-\left(-\lambda r^2\right)^{n-1}\right)\right).$$

For $\lambda > 0$ (spheres) this can be deduced from [12, Proposition 3.11].

Acknowledgements JM thanks Tudor Ratiu and the staff of the Bernoulli Centre in Lausanne for their hospitality, as much of this paper was written during an extended visit there. TT thanks L. Mahadevan and the staff of SEAS at Harvard for their hospitality, as much of this paper was finished during an extended visit there.

References

1. Boatto, S.: Curvature perturbations and stability of a ring of vortices. Discrete Contin. Dyn. Syst. **10**, 349–375 (2008)
2. Boatto, S., Cabral, H.: Nonlinear stability of a latitudinal ring of point-vortices on a nonrotating sphere. SIAM J. Appl. Math. **64**, 216–230 (2003)
3. Bridges, T.J., Furter, J.E.: Singularity Theory and Equivariant Symplectic Maps. Lecture Notes in Mathematics, vol. 1558. Springer, Berlin (1993)
4. Buono, P.L., Laurent-Polz, F., Montaldi, J.: Symmetric Hamiltonian bifurcations. In: Montaldi, J., Ratiu, T.S. (eds.) Geometric Mechanics and Symmetry: The Peyresq Lectures. LMS Lecture Notes Series, vol. 306, pp. 357–402. Cambridge UP, Cambridge (2005)
5. Denkowska, Z., Łojasiewicz, S., Stasica, J.: Certaines propriétés élémentaires des ensembles sous-analytiques. Bull. Acad. Polonaise Sci. Sér. Sci. Math. **27**, 529–535 (1979)
6. Golubitsky, M., Stewart, I.: Generic bifurcation of Hamiltonian systems with symmetry. Physica D **24**, 391–405 (1987)
7. Golubitsky, M., Stewart, I., Schaeffer, D.G.: Singularities and Groups in Bifurcation Theory, vol. II. Springer, Berlin (1988)
8. Hansen, E.: A Table of Series and Products. Prentice-Hall, Englewood Cliffs (1975)
9. Hironaka, H.: Introduction to Real Analytic Sets and Real-Analytic Maps. Instituto 'L. Tonelli', Pisa (1973)
10. Kimura, Y.: Vortex motion on surfaces with constant curvature. Proc. R. Soc. Lond. A **455**, 245–259 (1999)
11. Kurakin, L.G., Yudovich, V.I.: The stability of stationary rotation of a regular vortex polygon. Chaos **12**, 574–595 (2002)
12. Laurent-Polz, F.: Point vortices on the sphere: a case with opposite vorticities. Nonlinearity **15**, 143–171 (2002)
13. Laurent-Polz, F., Montaldi, J., Roberts, R.M.: Point vortices on the sphere: stability of symmetric relative equilibria. J. Geom. Mech. **3**, 439–486 (2011)
14. Lim, C., Montaldi, J., Roberts, R.M.: Relative equilibria of point vortices on the sphere. Physica D **148**, 97–135 (2001)
15. Marsden, J., Ratiu, T.S.: Introduction to Mechanics and Symmetry, 2nd edn. Springer, Berlin (1999)
16. Mertz, G.: Stability of body-centered polygonal configurations of ideal vortices. Phys. Fluids **21**, 1092–1095 (1978)
17. Montaldi, J.: Persistence and stability of relative equilibria. Nonlinearity **11**, 449–466 (1997)
18. Polvani, L., Dritschel, D.: Wave and vortex dynamics on the surface of a sphere. J. Fluid Mech. **255**, 35–64 (1993)
19. Poston, T., Stewart, I.: Catastrophe Theory and Applications. Dover, New York (1996) (Original edition Prentice Hall, 1978)
20. Schmidt, D.: The stability of the Thomson heptagon. Regul. Chaotic Dyn. **9**, 519–528 (2004)
21. Souriau, J.-M.: Structure of Dynamical Systems: A Symplectic View of Physics. Progress in Mathematics, vol. 149. Birkhäuser, Boston (1997)
22. Thomson, J.J.: A Treatise on the Motion of Vortex Rings. An Essay to Which the Adams Prize Was Adjudged in 1882, in the University of Cambridge. Macmillan, London (1883)

Chapter 15
Gradient Flows in the Normal and Kähler Metrics and Triple Bracket Generated Metriplectic Systems

Anthony M. Bloch, Philip J. Morrison, and Tudor S. Ratiu

Dedicated to Jürgen Scheurle on the occasion of his 60th birthday

Abstract The dynamics of gradient and Hamiltonian flows with particular application to flows on adjoint orbits of a Lie group and the extension of this setting to flows on a loop group are discussed. Different types of gradient flows that arise from different metrics including the so-called normal metric on adjoint orbits of a Lie group and the Kähler metric are compared. It is discussed how a Kähler metric can arise from a complex structure induced by the Hilbert transform. Hybrid and metriplectic flows which combine Hamiltonian and gradient components are examined. A class of metriplectic systems that is generated by completely antisymmetric triple brackets (trilinear brackets) is described and for finite-dimensional systems given a Lie algebraic interpretation. A variety of explicit examples of the several types of flows are given. It is shown that this geometry describes a number of classical ordinary and partial differential equations of interest and that the different metrics give rise to different kinds of dissipation that occur in applications.

A.M. Bloch
Department of Mathematics, The University of Michigan, 530 Church Street, Ann Arbor, MI 48109-1043, USA
e-mail: abloch@umich.edu

P.J. Morrison
Department of Physics and Institute for Fusion Studies, University of Texas, 2515 Speedway Stop C1600, Austin, TX 78712-0264, USA
e-mail: morrison@physics.utexas.edu

T.S. Ratiu (✉)
Department of Mathematics and Bernoulli Center, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
e-mail: tudor.ratiu@epfl.ch

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_15,

Keywords Adjoint orbits • Gradient flows • Hamiltonian systems • Integrable systems • Loop groups • Metriplectic systems • Thermodynamics

15.1 Introduction

Dynamical systems describing physical phenomena, finite or infinite dimensional, typically have some terms that are, in some sense, Hamiltonian and others that can be recognized as dissipative, with the Hamiltonian part associated with a Poisson bracket and the dissipative one being some kind of gradient flow. The description of Hamiltonian systems has received much attention over nearly two centuries and, although some forms of dissipation have been intensively studied, the understanding and classification of dissipative dynamics, a much broader topic, is considerably less well developed. Early modern treatments of geometric Hamiltonian mechanics include those of [1, 4, 5, 68]; the literature on this topic is now immense. A special type of gradient flow that preserves the invariants of the Hamiltonian terms, the double bracket formalism due to [21] (see also [11, 12]), occurs in a variety of contexts (see [17, 18]) and is well adapted to practical numerical computations (e.g., [28, 72]). Other examples of infinite dimensional gradient flows include the Cahn–Hilliard equation (e.g., [62]), the celebrated Ricci flow (e.g., [23, 33]) which is a nonlinear diffusion-like equation, and certain astrophysical systems (e.g., [40]). Various types of gradient flows in the infinite dimensional setting, including double bracket flows that are nonlocal, were treated in [35–39]. A general form for combined Hamiltonian and gradient flows was described in [53] and were termed *metriplectic* (see also [46, 49, 55, 60]). Due to the profusion of such types of systems consisting of the sum of a Hamiltonian and a dissipative vector field, it is of interest to investigate the general geometric structure of such evolutionary equations and isolate their key properties. The goal of this paper is precisely such a study, in both the ODE and PDE contexts. Specifically, we investigate the dynamics of gradient and Hamiltonian flows, with particular applications to those on coadjoint orbits of Lie groups and the extension of this setting to loop groups. We compare the different types of gradient flows that arise from various metrics, in particular, the so-called normal and the Kähler metrics on adjoint orbits of compact Lie groups. We discuss how a Kähler metric on a loop group arises from the complex structure induced by the Hilbert transform. In addition, we present the general theory of metriplectic flows in both the finite and infinite dimensional setting. These systems have the remarkable property that they conserve energy, produce entropy, and equilibria are found by a maximum entropy principle (see Sect. 15.4 for the precise statements). Particular attention is given to metriplectic flows that arise from completely antisymmetric triple brackets. For finite dimensional systems, we show how the triple bracket has a natural Lie algebraic formulation and for infinite dimensional systems we present a procedure for constructing general classes of metriplectic PDEs. We also discuss energy dissipating hybrid systems that are the sum of a Hamiltonian and a gradient

term. Several examples of hybrid and metriplectic flows are given, including finite dimensional systems such as the Toda lattice on the real line and metriplectic ones arising from $\mathfrak{so}(3)$ brackets. In the infinite dimensional context, we present a metriplectic 1+1 dissipative system and hybrid systems, such as KdV with dissipation and the Ott-Sudan equation [61] that describes Landau damping. This paper is partly expository, reviewing the necessary background geometry on adjoint orbits and loop groups, as well as gradient and metriplectic flows in various settings, although we make no claim to be comprehensive. New material on the various types of gradient flows that can occur in the loop group setting is included, as well as considerations generalizing metriplectic flows in both the finite and infinite dimensional setting.

The paper is organized as follows. In Sect. 15.2, we review background material, such as metrics on adjoint orbits, Toda flows, and the double bracket equation. Sections 15.3 and 15.4 contain the main new results of the paper. In Sect. 15.3, we discuss metrics on loop groups and related gradient flows, while in Sect. 15.4, we present our results on metriplectic systems, in both finite and infinite dimensions, and give examples.

15.2 Metrics on Adjoint Orbits of Compact Lie Groups and Associated Dynamical Systems

15.2.1 Double Bracket Systems

Let $\mathfrak{g}_u$ be the compact real form of a complex semisimple Lie algebra $\mathfrak{g}$, G_u a compact connected real Lie group with Lie algebra $\mathfrak{g}_u$, and κ the Killing form (on $\mathfrak{g}$ or $\mathfrak{g}_u$, depending on the context).

The "normal" metric on the adjoint orbit $\mathscr{O}$ of G_u through $L_0 \in \mathfrak{g}_u$ (see [6, 9, Chapter 8]) is given as follows. Decompose orthogonally $\mathfrak{g}_u = \mathfrak{g}_u^L \oplus \mathfrak{g}_{uL}$, relative to the invariant inner product $\langle\ ,\ \rangle := -\kappa(\ ,\ )$, where $\mathfrak{g}_{uL} := \ker \mathrm{ad}_L$ is the centralizer of L and $\mathfrak{g}_u^L = \mathrm{range}\ \mathrm{ad}_L$; as usual, $\mathrm{ad}_L := [L, \cdot]$. For $X \in \mathfrak{g}_u$ denote by $X^L \in \mathfrak{g}_u^L$ and $X_L \in \mathfrak{g}_{uL}$ the orthogonal projections of X on $\mathfrak{g}_u^L$ and $\mathfrak{g}_{uL}$, respectively. Recall that a general vector tangent at L to the adjoint orbit $\mathscr{O}$ is necessarily of the form $[L, X]$ for some $X \in \mathfrak{g}_u$. The *normal metric* on $\mathscr{O}$ is the G_u-invariant Riemannian metric whose value at $L \in \mathscr{O}$ is given by

$$\langle [L, X], [L, Y] \rangle_{\mathrm{normal}} := \langle X^L, Y^L \rangle \tag{15.1}$$

for any $X, Y \in \mathfrak{g}_u$.

Fix $N \in \mathfrak{g}_u$ and consider the flow

$$\frac{d}{dt} L(t) = [L(t), [L(t), N]]\,, \qquad L(0) = L_0 \in \mathfrak{g}_u\,, \tag{15.2}$$

on the adjoint orbit $\mathscr{O}$ of G_u. We recall the following well-known result [13–16, 21, 22].

Proposition 15.1. *The vector field given by the ordinary differential equation (15.2) is the gradient of the function $H(L) = \kappa(L, N)$ relative to the normal metric on $\mathscr{O}$.*

Proof. By the definition of the gradient $\operatorname{grad} H(L) \in T_L\mathscr{O} \subset \mathfrak{g}_u$ relative to the normal metric, we have for any $L \in \mathscr{O}$ and $\delta L \in \mathfrak{g}_u$,

$$dH(L) \cdot [L, \delta L] = \langle \operatorname{grad} H(L), [L, \delta L] \rangle_{\text{normal}} \tag{15.3}$$

where $\cdot$ denotes the natural pairing between 1-forms and tangent vectors and $[L, \delta L]$ is an arbitrary tangent vector at L to $\mathscr{O}$. Set $\operatorname{grad} H(L) = [L, X] = [L, X^L]$. Then (15.3) becomes

$$-\langle [L, \delta L], N \rangle = \langle [L, X], [L, \delta L] \rangle_{\text{normal}}$$

or, equivalently,

$$\langle [L, N], \delta L \rangle = \langle X^L, \delta L^L \rangle = \langle X^L, \delta L \rangle .$$

Since $[L, N] \in \mathfrak{g}_u^L$, this implies that $X^L = [L, N]$, and hence $\operatorname{grad} H(L) = [L, [L, N]]$, as stated. □

The same computation, for a general function $H \in C^\infty(\mathfrak{g}_u)$, yields

$$\operatorname{grad} H(L) = -[L, [L, \nabla H(L)]] \tag{15.4}$$

where $\nabla H(L)$ denotes the gradient of the function H relative to the invariant inner product $\langle \ , \ \rangle := -\kappa(\ , \ )$, i.e., $dH(L) \cdot X = \langle \nabla H(L), X \rangle$ for any $X \in \mathfrak{g}_u$.

15.2.2 The Finite Toda System

The double bracket equation (15.2) is intimately related to the finite non-compact Toda lattice system. This is a Hamiltonian system modeling n particles moving freely on the x-axis and interacting under an exponential potential. Denoting the position of the kth particle by x_k, the Hamiltonian is given by

$$H(x, y) = \frac{1}{2} \sum_{k=1}^{n} y_k^2 + \sum_{k=1}^{n-1} e^{x_k - x_{k+1}}$$

and hence the associated Hamiltonian equations are

$$\dot{x}_k = \frac{\partial H}{\partial y_k} = y_k\,, \qquad \dot{y}_k = -\frac{\partial H}{\partial x_k} = e^{x_{k-1}-x_k} - e^{x_k - x_{k+1}}\,, \tag{15.5}$$

where we use the conventions $e^{x_0-x_1} = e^{x_n-x_{n+1}} = 0$, which corresponds to formally setting $x_0 = -\infty$ and $x_{n+1} = +\infty$.

This system of equations has an extraordinarily rich structure. Part of this is revealed by Flaschka's change of variables [27] given by

$$a_k = \frac{1}{2}e^{(x_k - x_{k+1})/2} \quad \text{and} \quad b_k = -\frac{1}{2}y_k\,, \tag{15.6}$$

which transform (15.5) to

$$\begin{cases} \dot{a}_k = a_k(b_{k+1} - b_k)\,, & k = 1, \dots, n-1\,, \\ \dot{b}_k = 2(a_k^2 - a_{k-1}^2)\,, & k = 1, \dots, n\,, \end{cases}$$

with the boundary conditions $a_0 = a_n = 0$. This system is equivalent to the Lax equation

$$\frac{d}{dt}L = [B, L] = BL - LB\,, \tag{15.7}$$

where

$$L = \begin{pmatrix} b_1 & a_1 & 0 & \cdots & 0 \\ a_1 & b_2 & a_2 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & \cdots & b_{n-1} & a_{n-1} \\ 0 & & \cdots & a_{n-1} & b_n \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & a_1 & 0 & \cdots & 0 \\ -a_1 & 0 & a_2 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & \cdots & 0 & a_{n-1} \\ 0 & & \cdots & -a_{n-1} & 0 \end{pmatrix}. \tag{15.8}$$

If $O(t)$ is the orthogonal matrix solving the equation

$$\frac{d}{dt}O = BO\,, \qquad O(0) = \text{ Identity}\,,$$

then from (15.7) we have

$$\frac{d}{dt}(O^{-1}LO) = 0\,.$$

Thus, $O^{-1}LO = L(0)$, i.e., $L(t)$ is related to $L(0)$ by conjugation with an orthogonal matrix and thus the eigenvalues of L, which are real and distinct, are preserved along the flow. This is enough to show that this system is explicitly solvable, or integrable. Equivalently, after fixing the center of mass, i.e., setting $b_1 + \cdots + b_n = 0$, the $n-1$ integrals in involution whose differentials are linearly

independent on an open dense set of phase space $\{(a_1, \ldots, a_{n-1}, b_1, \ldots, b_n) \mid b_1 + \cdots + b_n = 0\}$ are Tr $L^2, \ldots,$ Tr L^n.

15.2.3 Lie Algebra Integrability of the Toda System

Let us quickly recall the well-known Lie algebraic approach to integrability of the Toda lattice. Let $\mathfrak{g}$ be a Lie algebra with an invariant non-degenerate bilinear symmetric form $\langle\,,\rangle$, i.e., $\langle[\xi,\eta],\zeta\rangle = \langle\xi,[\eta,\zeta]\rangle$ for all $\xi,\eta,\zeta \in \mathfrak{g}$ and $\langle\xi,\cdot\rangle = 0$ implies $\xi = 0$. Suppose that $\mathfrak{k}, \mathfrak{s} \subset \mathfrak{g}$ are Lie subalgebras and that, *as vector spaces*, $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{s}$. Let $\pi_\mathfrak{k} : \mathfrak{g} \to \mathfrak{k}$, $\pi_\mathfrak{s} : \mathfrak{g} \to \mathfrak{s}$ be the two projections induced by this vector space direct sum decomposition. Since $\mathfrak{g} \ni \xi \overset{\sim}{\longmapsto} \langle\xi,\cdot\rangle \in \mathfrak{g}^*$ is a vector space isomorphism, it naturally induces the isomorphisms $\mathfrak{k}^\perp \cong \mathfrak{s}^*$, $\mathfrak{s}^\perp \cong \mathfrak{k}^*$. By non-degeneracy of $\langle\,,\rangle$, we have $\mathfrak{g} = \mathfrak{s}^\perp \oplus \mathfrak{k}^\perp$; denote by $\pi_{\mathfrak{k}^\perp} : \mathfrak{g} \to \mathfrak{k}^\perp$, $\pi_{\mathfrak{s}^\perp} : \mathfrak{g} \to \mathfrak{s}^\perp$ the two projections induced by this vector space direct sum decomposition. In particular, $\mathfrak{g}$, $\mathfrak{s}^\perp$, $\mathfrak{k}^\perp$ all carry natural Lie–Poisson structures. The $(-)$Lie–Poisson bracket of $\mathfrak{s}^* \cong \mathfrak{k}^\perp$ is given by

$$\{\varphi,\psi\}(\xi) = -\langle\xi,[\pi_\mathfrak{s}\nabla\varphi(\xi),\pi_\mathfrak{s}\nabla\psi(\xi)]\rangle\,, \qquad \xi \in \mathfrak{k}^\perp, \tag{15.9}$$

where $\varphi,\psi : \mathfrak{k}^\perp \to \mathbb{R}$ are any smooth functions, extended arbitrarily to smooth functions, also denoted by φ and ψ, on $\mathfrak{g}$ and $\nabla\varphi$, $\nabla\psi$ are the gradients of these arbitrary extensions relative to $\langle\,,\rangle$. This formula follows from the fact that the gradient on $\mathfrak{k}^\perp$ of $\varphi|_{\mathfrak{k}^\perp}$, which is an element of $\mathfrak{s}$ due to the isomorphism $\mathfrak{k}^\perp \cong \mathfrak{s}^*$, equals $\pi_\mathfrak{s}\nabla\varphi$. Thus, the Hamiltonian vector field of $\psi \in C^\infty(\mathfrak{k}^\perp)$, given by $\dot\varphi = \{\varphi,\psi\}$ for any $\varphi \in C^\infty(\mathfrak{k}^\perp)$, has the expression

$$X_\psi(\xi) = -\pi_{\mathfrak{k}^\perp}\left[\pi_\mathfrak{s}\nabla\psi(\xi),\xi\right], \qquad \xi \in \mathfrak{k}^\perp \tag{15.10}$$

with the same conventions as above.

If $\psi \in C^\infty(\mathfrak{g})$ is invariant, i.e., $[\nabla\psi(\zeta),\zeta] = 0$ for all $\zeta \in \mathfrak{g}$, then (15.10) simplifies to

$$X_\psi(\xi) = [\pi_\mathfrak{k}\nabla\psi(\xi),\xi] = -\left[\pi_\mathfrak{s}\nabla\psi(\xi),\xi\right], \qquad \xi \in \mathfrak{k}^\perp. \tag{15.11}$$

The Adler–Kostant–Symes Theorem (see [3,44,66,69,70] for many theorems of the same type) states that if φ and ψ are both invariant functions on $\mathfrak{g}$, then $\{\varphi,\psi\} = 0$ on $\mathfrak{k}^\perp$ which is equivalent to the commutation of the flows of the Hamiltonian vector fields (15.11).

Suppose that $G = KS$, where G is a Lie group with Lie algebra $\mathfrak{g}$ and $K, S \subset G$ are closed subgroups with Lie algebras $\mathfrak{k}$ and $\mathfrak{s}$, respectively. The writing $G = KS$ means that each element $g \in G$ can be uniquely decomposed as $g = ks$, where $k \in K$ and $s \in S$ and that this decomposition defines a smooth diffeomorphism $K \times S \approx G$. The coadjoint action of S on $\mathfrak{s}^*$ has the following expression, if $\mathfrak{s}^*$ is

identified with $\mathfrak{k}^\perp$ via $\langle\,,\rangle$: if $s \in S$, $\xi \in \mathfrak{k}^\perp$, then $s \cdot \xi = \pi_{\mathfrak{k}^\perp} \operatorname{Ad}_s \xi$, where $\operatorname{Ad}_s \xi$ is the adjoint action in G of the element $s \in S \subset G$ on $\xi \in \mathfrak{k}^\perp \subset \mathfrak{g}$.

For the Toda lattice (15.7), this general setup applies in the following way. Let $G = \mathrm{GL}(n,\mathbb{R})$, $K = \mathrm{SO}(n)$, $S = \{\text{invertible lower triangular matrices}\}$, $G = KS$ is the Gram–Schmidt orthonormalization process, $\mathfrak{g} = \mathfrak{gl}(n,\mathbb{R})$, $\mathfrak{k} = \mathfrak{so}(n)$, $\mathfrak{s} = \{\text{lower triangular matrices}\}$, $\langle \xi, \eta \rangle := \operatorname{Tr}(\xi\eta)$ for all $\xi, \eta \in \mathfrak{gl}(n,\mathbb{R})$, $\mathfrak{k}^\perp = \mathfrak{sym}(n)$ the vector space of symmetric matrices, and $\mathfrak{s}^\perp = \mathfrak{n}$, the nilpotent Lie algebra of strictly lower triangular matrices. The set of matrices L in (15.8) is a union of S-coadjoint orbits parametrized by the value of the trace; for example, the set of trace zero matrices L of the form (15.8) equals the S-coadjoint orbit through the symmetric matrix that has everywhere zero entries with the exception of the upper and lower first diagonals where all entries are equal to one. Thus, the Toda lattice is a Poisson system whose restriction to a symplectic leaf is a classical Hamiltonian system with $n-1$ degrees of freedom. The Hamiltonian of the Toda lattice is $\frac{1}{2} \operatorname{Tr} L^2$ and the $f_k(L) := \frac{1}{k} \operatorname{Tr} L^k$, $k = 1, \ldots, n-1$ are the $n-1$ integrals in involution (by the Adler–Kostant–Symes Theorem) and are generically independent.

15.2.4 The Toda System as a Double Bracket Equation

If N is the matrix $\operatorname{diag}\{1, 2, \ldots, n\}$, the Toda equations (15.7) may be written in the double bracket form (15.2) for $B := [N, L]$. This was shown in [11]; the consequences of this fact were further analyzed for general compact Lie algebras in [13–15]. As shown in Proposition 15.1, the double bracket equation, with L replaced by $\mathrm{i}L$ and N by $\mathrm{i}N$, restricted to a level set of the integrals described above, i.e., restricted to a generic adjoint orbit of $\mathrm{SU}(n)$, is the gradient flow of the function $\operatorname{Tr} LN$ with respect to the normal metric; see [15] for this approach.

This observation easily implies that the flow tends asymptotically to a diagonal matrix with the eigenvalues of $L(0)$ on the diagonal and ordered according to magnitude, recovering the result of [24, 56, 71].

15.2.5 Riemannian Metrics on $\mathscr{O}$

Now, we recall that, in addition to the normal metric on an adjoint orbit, there are other natural G_u-invariant metrics: the induced and the group invariant Kähler metrics (as discussed in [6, §4], [7], and [9, Chapter 8]).

First, there is the *induced metric* b on $\mathscr{O}$, defined by $b := \iota^* (-\kappa(\ ,\ ))$, where $\iota : \mathscr{O} \hookrightarrow \mathfrak{g}_u$ is the inclusion and $\langle\ ,\ \rangle := -\kappa(\ ,\ )$ is thought of as a constant Riemannian metric on $\mathfrak{g}_u$. Therefore,

$$b(L)([L,X],[L,Y]) := \langle [L,X],[L,Y]\rangle \tag{15.12}$$

for any $L \in \mathscr{O}$, $X, Y \in \mathfrak{g}_u$. The induced metric on $\mathscr{O}$ is also G_u-invariant.

Second, there are the G_u-invariant Kähler metrics on $\mathscr{O}$ compatible with the natural complex structure (of course, induced by the complex structure of G). These are in bijective correspondence (by the transgression homomorphism) with the set of G_u-invariant sections of the trivial vector bundle over $\mathscr{O}$ whose fiber at $L \in \mathscr{O}$ is the center of $\ker(\mathrm{ad}_L)$ and whose scalar product with all positive roots is positive [9, Proposition 8.83]. Among these, there is the G_u-invariant Kähler metric b_2 which is compatible with both the natural complex structure on $\mathscr{O}$ and has as imaginary part the orbit symplectic structure; b_2 is called the *standard Kähler metric* on $\mathscr{O}$.

The G_u-invariant Riemannian metrics on a maximal dimensional orbit $\mathscr{O}$ are completely determined by T-invariant inner products on the direct sum of the two-dimensional root spaces of $\mathfrak{g}_u$, which is the tangent space to $\mathscr{O}$ at the point $L_0 \in \mathfrak{t}$ in the interior of the positive Weyl chamber; recall that $\mathscr{O}$ intersects the positive Weyl chamber in a unique point. The negative of the Killing form induces on each such two-dimensional space an inner product. This inner product, left translated at all points of $\mathscr{O}$ by elements of G_u, yields the normal metric on $\mathscr{O}$. Any other G_u-invariant inner product on $\mathscr{O}$ is obtained by left translating at all points of $\mathscr{O}$ the inner product on this direct sum of two-dimensional root spaces obtained by multiplying in each two-dimensional summand the inner product with a positive real constant.

Since L_0 lies in the interior of the positive Weyl chamber (because $\mathscr{O}$ is maximal dimensional), $\alpha(L_0) > 0$ for all positive roots α of $\mathfrak{g}_u$. Then the constant by which the natural inner product on the two-dimensional root space needs to be multiplied in order to get the standard Kähler metric is $\alpha(L_0)$, whereas to get the induced metric, it is $\alpha(L_0)^2$ [6, Remark 2 in §4]. We can formulate this differently, as in [15]. Since, by (15.12) and (15.1),

$$\begin{aligned} b(L)([L,X],[L,Y]) = \langle [L,X],[L,Y]\rangle &= \left\langle [L,X^L],[L,Y^L]\right\rangle = \left\langle -[L,[L,X^L]],Y^L\right\rangle \\ &= \left\langle -[L,[L,X^L]]^L,Y^L\right\rangle = \langle -\operatorname{ad}_L^2[L,X],[L,Y]\rangle_{\text{normal}} \end{aligned}$$

we have

$$b(L)([L,X],[L,Y]) = b_1(L)(\mathscr{A}(L)^2[L,X],[L,Y]), \tag{15.13}$$

where we denote now by b_1 the normal metric and $\mathscr{A}(L) := \sqrt{(\mathrm{i}\,\mathrm{ad}_L)^2}$ is the positive square root of $(\mathrm{i}\,\mathrm{ad}_L)^2 = -\operatorname{ad}_L^2 = \mathscr{A}(L)^2$. The standard Kähler metric on $\mathscr{O}$ is then given by

$$b_2(L)[L,X],[L,Y]) = b_1(\mathscr{A}(L)[L,X],[L,Y]). \tag{15.14}$$

Note that, as opposed to the normal and induced metrics which have explicit expressions, the standard Kähler metric on $\mathscr{O}$ requires the spectral decomposition of

$\mathscr{A}(L)$ at any point $L \in \mathscr{O}$. Or, as explained above, one expresses it at the point L_0 in the positive Weyl chamber in terms of the positive roots and then left translates the resulting inner product at any point of $\mathscr{O}$. The normal metric does not depend on the operators $\mathscr{A}(L)$, whereas the standard Kähler and induced metrics do.

15.3 Gradient Flows on the Loop Group of the Circle

In this section we introduce three weak Riemannian metrics on the subgroup of average zero functions of the connected component of the loop group $\widetilde{\mathrm{L}}(S^1)$ of the circle, analogous to the normal, standard Kähler, and induced metrics on adjoint orbits of compact semisimple Lie groups. Of course, we shall not work on adjoint orbits of this group because they degenerate to points, $\widetilde{\mathrm{L}}(S^1)$ being a commutative group. Then we shall compute the gradient flows for these three metrics.

15.3.1 The Loop Group of S^1

Recall (e.g., [65]) that the loop group $\widetilde{\mathrm{L}}(S^1)$ of the circle S^1 consists of smooth maps of S^1 to S^1. With pointwise multiplication, $\widetilde{\mathrm{L}}(S^1)$ is a commutative group. Often, elements of $\widetilde{\mathrm{L}}(S^1)$ are written as $e^{\mathrm{i}f}$, where

$$f \in \widetilde{\mathrm{L}}(\mathbb{R}) := \left\{ g : [-\pi, \pi] \to \mathbb{R} \mid g \text{ is } C^\infty,\ g(\pi) = g(-\pi) + 2n\pi, \text{ for some } n \in \mathbb{Z} \right\};$$

n is the *winding number* of the closed curve $[-\pi, \pi] \ni t \mapsto e^{\mathrm{i}g(t)} \in S^1$ about the origin. More precisely, there is an exact sequence of groups

$$\begin{array}{ccccccccc} 0 & \longrightarrow & \mathbb{Z} & \longrightarrow & \widetilde{\mathrm{L}}(\mathbb{R}) & \xrightarrow{\widetilde{\exp}} & \widetilde{\mathrm{L}}(S^1) & \longrightarrow & \mathbb{Z} & \longrightarrow 0 \\ & & n & \longmapsto & 2\pi n; & f & \longmapsto & e^{\mathrm{i}f} & \longmapsto & \frac{f(\pi)-f(-\pi)}{2\pi} \end{array}$$

which shows that $\ker \widetilde{\exp} = \mathbb{Z}$ and $\operatorname{coker} \widetilde{\exp} = \{0\}$. Thus the connected components of $\widetilde{\mathrm{L}}(S^1)$ are indexed by the winding number. The connected component of the identity $\widetilde{\mathrm{L}}(S^1)_0$ consists of loops with winding number zero about the origin.

If one insists on working with smooth loops, then one can consider $\widetilde{\mathrm{L}}(S^1)$ and $\widetilde{\mathrm{L}}(S^1)_0$ as Fréchet Lie groups either in the convenient calculus of [45] or in the tame category of [32].

Alternatively, one can work with loops $e^{\mathrm{i}f}$ for $f : [-\pi, \pi] \to \mathbb{R}$ of Sobolev class H^s, where $s \geq 1$ (or appropriate $W^{s,p}$ or Hölder spaces). By standard theory (see, e.g., [63] or [2]), it is checked that $\widetilde{\mathrm{L}}(S^1)$ is a Hilbert Lie group (see, e.g., [20] or [58]). We shall not add the index s on $\widetilde{\mathrm{L}}(\mathbb{R})$ and $\widetilde{\mathrm{L}}(S^1)$; from now on we work exclusively in this category of H^s Sobolev class maps and loops. A simple

proof of the fact that $\widetilde{\mathrm{L}}(\mathbb{R})$ is a Hilbert Lie group was given to us by K.-H. Neeb. First, note that $\widetilde{\mathrm{L}}(\mathbb{R})$ is a closed additive subgroup of the Hilbert space $H^s(\mathbb{R}) := \{h : \mathbb{R} \to \mathbb{R} \mid h \text{ of class } H^s\}$. Second, $\widetilde{\mathrm{L}}(\mathbb{R}) = \widetilde{\mathrm{L}}(\mathbb{R})_0 \times \mathbb{Z}$ as topological groups, where $\widetilde{\mathrm{L}}(\mathbb{R})_0 := \{g \in \widetilde{\mathrm{L}}(\mathbb{R}) \mid g(\pi) = g(-\pi)\}$ is the closed vector subspace of $H^s(\mathbb{R})$ consisting of periodic functions; hence it is an additive Hilbert Lie group. Therefore, there is a unique Hilbert Lie group structure on $\widetilde{\mathrm{L}}(\mathbb{R})$ for which $\widetilde{\mathrm{L}}(\mathbb{R})_0$ is the connected component of the identity. For general criteria that characterize Lie subgroups in infinite dimensions, see [59, Theorem IV.3.3] (even for certain classes of Lie groups modeled on locally convex spaces). Third, since $\widetilde{\exp} : \widetilde{\mathrm{L}}(\mathbb{R}) \to \widetilde{\mathrm{L}}(S^1)$ maps bijectively each connected component of $\widetilde{\mathrm{L}}(\mathbb{R})$ to a connected component of $\widetilde{\mathrm{L}}(S^1)$, it induces a Hilbert Lie group structure on $\widetilde{\mathrm{L}}(S^1)$.

The commutative Hilbert Lie algebra of $\widetilde{\mathrm{L}}(S^1)$ is clearly $H^s(S^1, \mathbb{R}) := \{u : S^1 \to \mathbb{R} \mid u \text{ of class } H^s\}$, the space of periodic H^s maps, and the exponential map $\exp : H^s(S^1, \mathbb{R}) \to \widetilde{\mathrm{L}}(S^1)$ is given by $\exp(u)(\theta) = e^{\mathrm{i}u(\theta)}$, where $\theta \in \mathbb{R}/2\pi\mathbb{Z} = S^1$.

15.3.2 The Based Loop Group of S^1

The inner product on the Hilbert space $L^2(S^1)$ of L^2 real valued functions on S^1 is defined by

$$\langle f, g\rangle := \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\, f(\theta)g(\theta)\,, \quad f, g \in L^2(S^1).$$

Following [7,64], we introduce the closed Hilbert Lie subgroup $\mathrm{L}(S^1) := \{\varphi \in \widetilde{\mathrm{L}}(S^1) \mid \varphi(1) = 1\}$ of $\widetilde{\mathrm{L}}(S^1)$ whose closed commutative Hilbert Lie algebra is $\mathrm{L}(\mathbb{R}) := \{u \in H^s(S^1, \mathbb{R}) \mid u(1) = 0\}$. The exponential map $\exp : \mathrm{L}(\mathbb{R}) \ni u \mapsto e^{\mathrm{i}u} \in \mathrm{L}(S^1)$ is a Lie group isomorphism (with $\mathrm{L}(\mathbb{R})$ thought of as a commutative group relative to addition), a fact that will play a very important role later on (see also [65, page 151, §8.9]).

There is a natural 2-cocycle ω on $\mathrm{L}(\mathbb{R})$, namely

$$\omega(u, v) := \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\, u'(\theta)v(\theta) = \langle u', v\rangle\ , \tag{15.15}$$

where $u' := du/d\theta$. Therefore, there is a central extension of Lie algebras

$$0 \longrightarrow \mathbb{R} \longrightarrow \widehat{\mathrm{L}(\mathbb{R})} \longrightarrow \mathrm{L}(\mathbb{R}) \longrightarrow 0$$

which, as shown in [67], integrates to a central extension of Lie groups

$$1 \longrightarrow S^1 \longrightarrow \widehat{\mathrm{L}(S^1)} \longrightarrow \mathrm{L}(S^1) \longrightarrow 1.$$

The "geometric duals" of $\mathrm{L}(\mathbb{R})$ and $\widehat{\mathrm{L}(\mathbb{R})} = \mathbb{R} \oplus \mathrm{L}(\mathbb{R})$ are themselves, relative to the weak L^2-pairing. It turns out that the coadjoint action of $\widehat{\mathrm{L}(S^1)}$ on $\widehat{\mathrm{L}(\mathbb{R})}$ preserves $\{1\} \oplus \mathrm{L}(\mathbb{R})$ so that, as usual, the coadjoint action of $\widehat{\mathrm{L}(S^1)}$ on $\mathrm{L}(\mathbb{R})$ is an affine action which, in this case, because the group is commutative, equals

$$\mathrm{Ad}^*_{e^{\mathrm{i}f}} \mu = \frac{f'}{f} = (\log|f|)' \qquad e^{\mathrm{i}f} \in \mathrm{L}(S^1), \quad \mu \in \mathrm{L}(\mathbb{R}).$$

Thus, the orbit of the constant function 0 is $\widehat{\mathrm{L}(S^1)}/S^1$ (where the denominator is thought of as constant loops), i.e., it equals $\mathrm{L}(S^1)$. Therefore, every element $u \in \mathrm{L}(\mathbb{R})$ of its Lie algebra has, in Fourier representation, vanishing zero order Fourier coefficient , i.e., $\hat{u}(0) = 0$.

Thus, the based loop group is a coadjoint orbit of its natural central extension and, according to Sect. 15.2, has three distinguished weak Riemannian metrics. These were computed explicitly in [7, 64, 65]; we recall them below.

15.3.3 $\mathrm{L}(S^1)$ as a Weak Kähler Manifold

Note that on $\mathrm{L}(\mathbb{R})$, the cocycle (15.15) is weakly non-degenerate. Therefore, left (or right) translating it at every point of the group $\mathrm{L}(S^1)$ yields a weakly non-degenerate closed two-form, i.e., a symplectic form. Thus, as expected, since it is a coadjoint orbit, the Hilbert Lie group $\mathrm{L}(S^1)$ carries an invariant symplectic form whose value at the identity element 1 (the constant loop equal to 1) is given by (15.15).

Now we introduce the *Hilbert transform* on the circle

$$\begin{aligned} \mathscr{H}u(\theta) &:= \frac{1}{2\pi} \fint_{-\pi}^{\pi} \mathrm{d}s\, u(s) \cot\left(\frac{\theta - s}{2}\right) = \frac{1}{2\pi} \fint_{-\pi}^{\pi} \mathrm{d}s\, u(\theta - s) \cot\left(\frac{s}{2}\right) \\ &:= \lim_{\varepsilon \to 0+} \frac{1}{\pi} \int_{\varepsilon \le |s| \le \pi} \mathrm{d}s\, u(\theta - s) \cot\left(\frac{s}{2}\right) \end{aligned} \tag{15.16}$$

for any $u \in L^2(S^1)$, where $\fint$ denotes the Cauchy principal value. We adopt here the sign conventions in [43, Formulas (3.202) and (6.38), Vol. 1]. If $u \in L^2(S^1)$, then $\mathscr{H}u \in L^2(S^1)$ and it is defined for almost every $\theta \in [-\pi, \pi]$ (Lusin's Theorem, [43, §6.19, Vol. 1]). The Hilbert transform has the following remarkable properties that will be used later on:

- If $u(\theta) = \sum_{n=-\infty}^{\infty} \hat{u}(n) e^{\mathrm{i}n\theta} \in L^2(S^1)$, where $\hat{u}(n) := \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta\, u(\theta) e^{-\mathrm{i}n\theta}$, so $\overline{\hat{u}(n)} = \hat{u}(-n)$ since u is real valued, then

$$\mathscr{H}u(\theta) = -\mathrm{i}\sum_{n=-\infty}^{\infty} \hat{u}(n)\,\mathrm{sign}(n)\,e^{\mathrm{i}n\theta} \in L^2(S^1) \tag{15.17}$$

which follows from the identity $\widehat{\mathscr{H}f}(n) = -\mathrm{i}\hat{f}(n)\,\mathrm{sign}(n)$ [43, Formulas (6.100) or (6.124), Vol. 1]. Here, $\mathrm{sign}(n) = 1$ if $n \in \mathbb{N}$, $\mathrm{sign}(n) = -1$ if $n \in -\mathbb{N}$, and $\mathrm{sign}(0) = 0$. Note that $\mathscr{H}u$ is also real valued since $\hat{u}(n)\,\mathrm{sign}(n) = -\hat{u}(-n)\,\mathrm{sign}(-n)$. The formula above implies that [43, Formula (6.126), Vol. 1]

$$\int_{-\pi}^{\pi} \mathrm{d}s\,\mathscr{H}u(s) = 0.$$

- For every $u \in L^2(S^1)$, we have the orthogonality property [43, Formula (6.127), Vol. 1]:

$$\langle u, \mathscr{H}u\rangle = 0.$$

- Take the orthonormal Hilbert basis $\{\varphi_n(\theta) := e^{\mathrm{i}n\theta} \mid n \in \mathbb{Z}\}$ of $L^2(S^1)$. Then [43, Formula (6.131), Vol. 1]:

$$\mathscr{H}\varphi_n(\theta) = -\mathrm{i}\,\mathrm{sign}(n)\,\varphi_n(\theta), \quad \text{for all} \quad n \in \mathbb{Z}.$$

  So, the eigenvalues of $\mathscr{H}$ are: $-\mathrm{i}$ for all $n > 0$, i for all $n < 0$, and 0 if $n = 0$.
- If $u, v \in L^2(S^1)$ then [43, Formula (6.99), Vol. 1]

$$\langle u, v\rangle = \frac{1}{4\pi^2}\left(\int_{-\pi}^{\pi} \mathrm{d}s\,u(s)\right)\left(\int_{-\pi}^{\pi} \mathrm{d}s\,v(s)\right) + \langle \mathscr{H}u, \mathscr{H}v\rangle$$

  and hence [43, Formula (6.97), Vol. 1]

$$\|u\|^2_{L^2(S^1)} = \left(\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}s\,u(s)\right)^2 + \|\mathscr{H}u\|^2_{L^2(S^1)}$$

  for any $u \in L^2(S^1)$. This shows that $\|\mathscr{H}u\|^2_{L^2(S^1)} \leq \|u\|^2_{L^2(S^1)}$ and the constant 1 is the best possible [43, Formulas (6.167) and (6.168), Vol. 1]. In particular, if the average of u is zero, then $\mathscr{H}$ is an isometry of $L^2(S^1)$.
- The Hilbert transform is skew-adjoint relative to the $L^2(S^1)$-inner product, i.e., $\mathscr{H}^* = -\mathscr{H}$ [43, Formula (6.98) or (6.106), Vol. 1].
- For any $u \in L^2(S^1)$ we have [43, Formulas (6.34), (6.82), or (6.156), Vol. 1]:

$$\mathscr{H}^2u(\theta) = -u(\theta) + \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}s\,u(s) = -u(\theta) + \hat{u}(0).$$

- For any $u \in H^s(S^1)$ with $s \geq 0$ we have $\mathscr{H}u \in H^s(S^1)$; this is an immediate consequence of (15.17). If $s \geq 1$, then $\mathscr{H}u' = (\mathscr{H}u)'$, i.e., $\mathscr{H} \circ \frac{d}{d\theta} = \frac{d}{d\theta} \circ \mathscr{H}$ on $H^s(S^1)$ with $s \geq 1$.

  Using these properties, if $u(\theta) = \sum_{n=-\infty}^{\infty} \hat{u}(n) e^{\mathrm{i}n\theta} \in H^1(S^1)$, then

$$u'(\theta) = \sum_{n=-\infty}^{\infty} \hat{u}(n) \mathrm{i}n e^{\mathrm{i}n\theta} \in L^2(S^1)$$

and hence

$$\left(\mathscr{H}u'\right)(\theta) = (\mathscr{H}u)'(\theta) = \left(-\mathrm{i} \sum_{n=-\infty}^{\infty} \hat{u}(n) \operatorname{sign}(n)\, e^{\mathrm{i}n\theta}\right)' = \sum_{n=-\infty}^{\infty} |n| \hat{u}(n) e^{\mathrm{i}n\theta}\,. \tag{15.18}$$

On the other hand, if $v \in H^2(S^1)$, then

$$-\frac{d^2}{d\theta^2} v(\theta) = \sum_{n=-\infty}^{\infty} n^2 \hat{v}(n) e^{in\theta} \tag{15.19}$$

and hence if $u \in H^1(S^1)$,

$$\left(-\frac{d^2}{d\theta^2}\right)^{\frac{1}{2}} u(\theta) = \sum_{n=-\infty}^{\infty} |n| \hat{u}(n) e^{in\theta} = (\mathscr{H}u')(\theta) = \left(\left(\mathscr{H} \circ \frac{d}{d\theta}\right) u\right)(\theta) \tag{15.20}$$

by (15.18). By the previous properties we have $(\mathscr{H} \circ d/d\theta)^2 = -d^2/d\theta^2$, as expected; note that the extra term, which is the zero order Fourier coefficient, does not appear in this case, because the derivative eliminates it.

Now, if $\varphi = e^{\mathrm{i}f} \in \mathrm{L}(S^1)$, i.e., $\varphi(1) = 1$ and $f : [-\pi, \pi] \to \mathbb{R}$ is a periodic function, then $\hat{f}(0) = f(0) = 0$. Similarly, if $u \in \mathrm{L}(\mathbb{R})$, i.e., $u(1) = 0$ and we think of u as a periodic function $u : [-\pi, \pi] \to \mathbb{R}$, then $\hat{u}(0) = u(0) = 0$. This, and the properties of the Hilbert transform on the circle, imply: $\mathscr{H}\left(\mathrm{L}(\mathbb{R})\right) \subseteq \mathrm{L}(\mathbb{R})$, $\mathscr{H}$ is unitary on $\mathrm{L}(\mathbb{R})$ (relative to the H^s-inner product), $\mathscr{H} \circ \mathscr{H} = -I$ on $\mathrm{L}(\mathbb{R})$. Concretely, the Hilbert transform on $\mathrm{L}(\mathbb{R})$ has the form:

$$u(\theta) = \sum_{n \in \mathbb{Z}\setminus\{0\}} \hat{u}(n) e^{\mathrm{i}n\theta} \in \mathrm{L}(\mathbb{R}) \quad \Rightarrow \quad \mathscr{H}u(\theta) = -\mathrm{i} \sum_{n \in \mathbb{Z}\setminus\{0\}} \hat{u}(n) \operatorname{sign}(n)\, e^{\mathrm{i}n\theta} \in \mathrm{L}(\mathbb{R}).$$

Thus, $\mathscr{H}$ defines the structure of a complex Hilbert space on $\mathrm{L}(\mathbb{R})$, relative to the H^s inner product, $s \geq 1$. Hence, translating $\mathscr{H}$ to any tangent space of $\mathrm{L}(S^1)$, we obtain an invariant almost complex structure on the Hilbert Lie group $\mathrm{L}(S^1)$ which is, in fact, a complex structure. For general criteria how to obtain complex structures

on real Banach manifolds, see [8]; the argument above is a very special case of these general methods.

Finally, $\mathrm{L}(S^1)$ is a Kähler manifold, as proved in [7]. This is immediately seen by noting that

$$g(1)(u,v) := \omega(\mathscr{H}u, v) = \sum_{n=-\infty}^{\infty} |n|\hat{u}(n)\hat{v}(n) \tag{15.21}$$

is symmetric and positive definite and so, by translations, defines a weak Riemannian metric on $\mathrm{L}(S^1)$. Note that this metric is *not* the H^s metric for any $s \geq 1$. In fact, the metric g is incomplete, whereas the H^s metric is complete.

Concluding, $(\mathrm{L}(S^1), \omega, g, \mathscr{H})$ is a weak Kähler manifold and all structures are group invariant (see [7, 64, 65]).

15.3.4 Weak Riemannian Metrics on $\mathrm{L}(S^1)$

The three metrics discussed in Sect. 15.2 for $\mathrm{L}(S^1)$, viewed as a coadjoint orbit of its central extension, have been computed in [64]. We recall here relevant formulas.

The *induced metric* is defined by the natural inner product on $\mathrm{L}(\mathbb{R})$, which is the usual L^2-inner product. Hence, the induced metric is obtained by left (equivalently, right) translation of the inner product

$$b(1)(u,v) := \langle u, v\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}t\, u(t)v(t) \tag{15.22}$$

for any two functions $u, v \in \mathrm{L}(\mathbb{R})$.

Define the following inner products on $\mathrm{L}(\mathbb{R})$:

$$b_2(1)(u,v) := b(1)(u, \mathscr{H}v') = \langle u, \mathscr{H}v'\rangle\,, \quad \text{if} \quad u, v \in H^s(S^1),\ s \geq 1 \tag{15.23}$$

$$b_1(1)(u,v) := b(1)(u', v') = \langle u', v'\rangle\,, \quad \text{if} \quad u, v \in H^s(S^1),\ s \geq 1. \tag{15.24}$$

Bilinearity and symmetry of $b_1(1)$ and $b_2(1)$ are obvious. If $u \in \mathrm{L}(S^1)$, writing $u(\theta) = \sum_{n=-\infty}^{\infty} \hat{u}(n)e^{\mathrm{i}n\theta}$ with $\hat{u}(0) = 0$, we have $u'(\theta) = \mathrm{i}\sum_{n=-\infty}^{\infty} n\hat{u}(n)e^{\mathrm{i}n\theta}$. Since $\{e^{\mathrm{i}n\theta} \mid n \in \mathbb{Z}\}$ is an orthonormal Hilbert basis of $L^2(S^1)$, we get

$$b_1(1)(u,u) = \sum_{n=-\infty}^{\infty} n^2|\hat{u}(n)|^2 \geq 0.$$

In addition, $b_1(1)(u,u) = 0$ if and only if $\hat{u}(n) = 0$ for all $n \neq 0$, i.e., $u(\theta) = \hat{u}(0) = 0$. This shows that $b_1(1)$ is indeed an inner product on $\mathrm{L}(\mathbb{R})$ which coincides

with the H^1 inner product. Hence, if $\mathrm{L}(\mathbb{R})$ is endowed with the H^s topology for $s \geq 1$, this inner product is strong if $s = 1$ and weak if $s > 1$. Left translating this inner product to any tangent space of $\mathrm{L}(S^1)$ (endowed with the H^s topology for $s \geq 1$), yields a Riemannian metric on $\mathrm{L}(S^1)$ that is strong for $s = 1$ and weak for $s > 1$. This Riemannian metric is the *normal metric* on $\mathrm{L}(S^1)$.

The inner product $b_2(1)$ is identical to $g(1)$ by (15.21), (15.23), and (15.15). Thus, translating this inner product to the tangent space at every point of the Hilbert Lie group $\mathrm{L}(S^1)$, yields the *standard Kähler metric* $b_2 = g$ on $\mathrm{L}(S^1)$, endowed with the H^s topology for $s \geq 1$. Note that if $u \in \mathrm{L}(S^1)$, then

$$b_2(1)(u,u) = \sum_{n=-\infty}^{\infty} |n||\hat{u}(n)|^2$$

which shows that the Kähler metric b_2 coincides with the $H^{1/2}$ metric and is, therefore, a weak metric on $\mathrm{L}(S^1)$.

There are relations similar to (15.13) and (15.14), namely

$$b(1)(u,v) = b_1(1)(\mathscr{A}^2 u, v), \qquad b_2(1)(u,v) = b_1(1)(\mathscr{A} u, v),$$

where

$$(\mathscr{A}^2 u)(\theta) = \sum_{n=-\infty}^{\infty} n^2 \hat{u}(n) e^{\mathrm{i}n\theta}, \qquad (\mathscr{A} u)(\theta) = \sum_{n=-\infty}^{\infty} |n| \hat{u}(n) e^{\mathrm{i}n\theta}$$

if $u(\theta) = \sum_{n=-\infty}^{\infty} \hat{u}(n) e^{\mathrm{i}n\theta}$. However, note that the relation involving $\mathscr{A}^2$ requires that $u \in H^s(S^1)$ with $s \geq 2$.

15.3.5 *Vector Fields on* $\mathrm{L}(S^1)$ *and* $\mathrm{L}(\mathbb{R})$

Recall that the exponential map $\exp : \mathrm{L}(\mathbb{R}) \ni u \mapsto e^{\mathrm{i}u} \in \mathrm{L}(S^1)$ is a Lie group isomorphism [65, page 151, §8.9]. Here we identified the Lie algebra of S^1 with $\mathbb{R}$, even though, naturally, it is the imaginary axis, the tangent space at $1 \in S^1$ to S^1. This means that care must be taken when carrying out standard Lie group operations with the exponential map, interpreted as the exponential of a purely imaginary number. Since such computations affect our next results, we clarify these statements below.

The tangent space at the identity 1 to S^1 is the imaginary axis. This is the natural Lie algebra of the Lie group S^1 and the exponential map is given by $\exp : \mathrm{i}\mathbb{R} \ni (\mathrm{i}x) \mapsto e^{\mathrm{i}x} \in S^1$. Of course, traditionally, one identifies $\mathrm{i}\mathbb{R}$ with $\mathbb{R}$ by dividing by i and thinks of the exponential map as $\exp : \mathbb{R} \ni x \mapsto e^{\mathrm{i}x} \in S^1$. Unfortunately, this induces some problems. For example, since (left) translation is given by $L_{e^{\mathrm{i}x}} e^{\mathrm{i}y} := e^{\mathrm{i}x} e^{\mathrm{i}y}$, it follows that

$$T_1 L_{e^{\mathrm{i}x}}(\mathrm{i}y) := \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} L_{e^{\mathrm{i}x}} e^{\mathrm{i}\varepsilon y} = \frac{d}{d\varepsilon}\bigg|_{\varepsilon=0} e^{\mathrm{i}x} e^{\mathrm{i}\varepsilon y} = \mathrm{i}y e^{\mathrm{i}x}, \tag{15.25}$$

so the identification of the Lie algebra with $\mathbb{R}$ poses no problems and we have, dividing both sides by i,

$$T_1 L_{e^{\mathrm{i}x}}(y) = y e^{\mathrm{i}x}. \tag{15.26}$$

However, the definition of the exponential map for any Lie group G with Lie algebra $\mathfrak{g}$, yields

$$\frac{d}{dt}\exp(t\xi) = T_e L_{\exp(t\xi)}\xi, \quad \text{for all} \quad \xi \in \mathfrak{g}. \tag{15.27}$$

This formula works perfectly well if the Lie algebra of S^1 is $\mathrm{i}\mathbb{R}$. Indeed

$$\frac{d}{dt} e^{\mathrm{i}tx} = \mathrm{i}x e^{\mathrm{i}tx}$$

which coincides with (15.27) in view of (15.25). On the other hand, if the Lie algebra is thought of as $\mathbb{R}$, i.e., the right-hand side needs to be divided by i, then with the definition of $\exp(tx) = e^{\mathrm{i}tx}$ the identity above is no longer valid. What we should get is

$$\frac{d}{dt}\exp(tx) = x\exp(tx) = T_1 L_{\exp(tx)} x = x e^{\mathrm{i}tx}$$

by (15.26) if $\exp(tx) = e^{\mathrm{i}tx}$, but the righthand side gives $\mathrm{i}x e^{\mathrm{i}tx}$, as we saw above. In other words, if the Lie algebra of S^1 is thought of as $\mathbb{R}$, as is traditionally done, then we need a formula for the derivative of the Lie group exponential map in terms of the exponential map of purely imaginary numbers. In view of the previous discussion, this formula is

$$\frac{d}{dt}\exp(tx) := \frac{1}{\mathrm{i}}\frac{d}{dt} e^{\mathrm{i}tx} = x e^{\mathrm{i}tx}. \tag{15.28}$$

With these remarks in mind, we shall now compute the push-forward of a vector field on $\mathrm{L}(\mathbb{R})$ to $\mathrm{L}(S^1)$.

Proposition 15.2. *Let $X \in \mathfrak{X}(\mathrm{L}(\mathbb{R}))$ be an arbitrary vector field . Then its push-forward to $\mathrm{L}(S^1))$ has the expression*

$$(\exp_* X)\left(e^{\mathrm{i}u}\right) = X(u) e^{\mathrm{i}u}$$

for any $u \in \mathrm{L}(\mathbb{R})$.

Proof. By the definition of push forward of vector fields by a diffeomorphism, we have

$$\begin{aligned}(\exp_* X)\left(e^{\mathrm{i}u}\right) &= \left(T\exp\circ X\circ\exp^{-1}\right)\left(e^{\mathrm{i}u}\right) = T_u\exp\left(X(u)\right)\\ &= \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\exp\left(u+\varepsilon X(u)\right) = \left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\exp(u)\exp(\varepsilon X(u))\\ &= \left(\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}\exp(\varepsilon X(u))\right)\exp(u) \overset{(15.28)}{=} \left(\frac{1}{\mathrm{i}}\left.\frac{d}{d\varepsilon}\right|_{\varepsilon=0}e^{i\varepsilon X(u)}\right)e^{\mathrm{i}u}\\ &= X(u)e^{\mathrm{i}u}\end{aligned}$$

as stated. □

15.3.6 The Gradient Vector Fields in the Three Metrics of $\mathrm{L}(S^1)$

We compute now the gradients of a specific function using the three metrics.

Theorem 15.1. *The gradients of the smooth function $H : \mathrm{L}(S^1) \to \mathbb{R}$ given by*

$$H\left(e^{\mathrm{i}f}\right) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\mathrm{d}\theta\, f'(\theta)^2$$

are

(i) $\nabla^1 H\left(e^{\mathrm{i}f}\right) = fe^{\mathrm{i}f}$ *for the normal metric* b_1*;*
(ii) $\nabla H\left(e^{\mathrm{i}f}\right) = -f''e^{\mathrm{i}f}$ *with respect to the induced metric* b *for* $f \in H^s(S^1)$ *with* $s \geq 2$*;*
(iii) $\nabla^2 H\left(e^{\mathrm{i}f}\right) = (\mathscr{H}f')e^{\mathrm{i}f}$ *with respect to the weak Kähler metric* b_2*.*

Proof. (i) Since $T_1 L_{e^{\mathrm{i}f}} u = ue^{\mathrm{i}f}$ for any $u \in \mathrm{L}(\mathbb{R})$ and $e^{\mathrm{i}f} \in \mathrm{L}(S^1)$, invariance of b_1 yields

$$\begin{aligned}b_1(1)\left(e^{-\mathrm{i}f}\nabla^1 H\left(e^{\mathrm{i}f}\right), u\right) &= b_1\left(e^{\mathrm{i}f}\right)\left(\nabla^1 H\left(e^{\mathrm{i}f}\right), ue^{\mathrm{i}f}\right) = \mathbf{d}H\left(e^{\mathrm{i}f}\right)\left(ue^{\mathrm{i}f}\right)\\ &= \left.\frac{d}{dt}\right|_{t=0} H\left(e^{\mathrm{i}(f+tu)}\right)\\ &= \left.\frac{d}{dt}\right|_{t=0}\frac{1}{4\pi}\int_{-\pi}^{\pi}\mathrm{d}\theta\,\left(f'(\theta)+tu'(\theta)\right)^2\\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi}\mathrm{d}\theta\, f'(\theta)u'(\theta) = \left\langle f', u'\right\rangle \overset{(15.1)}{=} b_1(1)(f,u)\end{aligned}$$

which shows that $\nabla^1 H\left(e^{\mathrm{i}f}\right) = fe^{\mathrm{i}f}$.

(ii) Proceeding as above, using the same notations, and assuming that $f \in H^s(S^1)$ with $s \geq 2$, we have

$$\begin{aligned} b(1)\left(e^{-\mathrm{i}f}\nabla H\left(e^{\mathrm{i}f}\right), u\right) &= b\left(e^{\mathrm{i}f}\right)\left(\nabla H\left(e^{\mathrm{i}f}\right), ue^{\mathrm{i}f}\right) = \mathbf{d}H\left(e^{\mathrm{i}f}\right)\left(ue^{\mathrm{i}f}\right) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\, f'(\theta)u'(\theta) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\, f''(\theta)u(\theta) \\ &= \langle -f'', u\rangle \overset{(15.12)}{=} b(1)\left(-f'', u\right) \end{aligned}$$

which shows that $\nabla H\left(e^{\mathrm{i}f}\right) = -f''e^{\mathrm{i}f}$.

(iii) This computation uses the isometry property of $\mathscr{H}$ relative to the L^2 inner product. We have,

$$\begin{aligned} b_2(1)\left(e^{-\mathrm{i}f}\nabla^2 H\left(e^{\mathrm{i}f}\right), u\right) &= b_2\left(e^{\mathrm{i}f}\right)\left(\nabla^2 H\left(e^{\mathrm{i}f}\right), ue^{\mathrm{i}f}\right) = \mathbf{d}H\left(e^{\mathrm{i}f}\right)\left(ue^{\mathrm{i}f}\right) \\ &= \langle f', u'\rangle = \langle \mathscr{H}f', \mathscr{H}u'\rangle \overset{(15.23)}{=} b_2(1)\left(\mathscr{H}f', u\right) \end{aligned}$$

which shows that $\nabla^2 H\left(e^{\mathrm{i}f}\right) = (\mathscr{H}f')e^{\mathrm{i}f}$. □

Since

$$\omega\left(e^{\mathrm{i}f}\right)\left(\mathscr{H}\nabla^2 H\left(e^{\mathrm{i}f}\right), ue^{\mathrm{i}f}\right) \overset{(15.21)}{=} b_2\left(e^{\mathrm{i}f}\right)\left(\nabla^2 H\left(e^{\mathrm{i}f}\right), ue^{\mathrm{i}f}\right) = \mathbf{d}H\left(e^{\mathrm{i}f}\right)\left(ue^{\mathrm{i}f}\right)$$

it follows that the Hamiltonian vector field on $\left(\mathrm{L}(S^1), \omega\right)$ for the function H is $X_H = \mathscr{H}\nabla^2 H$. Since $\mathscr{H}$ commutes with the tangent lift to group translations, Theorem 15.1(iii) implies that

$$X_H\left(e^{\mathrm{i}f}\right) = \left(\mathscr{H}\nabla^2 H\right)\left(e^{\mathrm{i}f}\right) = \mathscr{H}\left(\nabla^2 H\left(e^{\mathrm{i}f}\right)\right) = \mathscr{H}\left(\left(\mathscr{H}f'\right)e^{\mathrm{i}f}\right) = -f'e^{\mathrm{i}f}.$$

This proves the first part of the following statement.

Corollary 15.1. *The Hamiltonian vector field of H relative to the translation invariant symplectic form ω on $\mathrm{L}(S^1)$ whose value at the identity element is given by (15.15) has the expression $X_H\left(e^{\mathrm{i}f}\right) = -f'e^{\mathrm{i}f}$. Its flow is the rotation*

$$\left(F_t\left(e^{\mathrm{i}f}\right)\right)(\theta) = e^{-\mathrm{i}(f(t+\theta)-f(t))}.$$

Proof. Since $\mathrm{L}(\mathbb{R}) \ni u \longmapsto e^{\mathrm{i}u} \in \mathrm{L}(S^1)$ is the exponential map and we think of $\mathbb{R}$ as the Lie algebra of S^1 (and not the imaginary axis), we write $de^{\mathrm{i}tu}/dt = ue^{\mathrm{i}tu}$ without the factor of i in front (see (15.28)). The verification that F_t is indeed the flow of X_H is straightforward:

$$\frac{d}{dt}\left(F_t\left(e^{\mathrm{i}f}\right)\right)(\theta) = \frac{d}{dt}e^{-\mathrm{i}(f(t+\theta)-f(t))} = -(f'(t+\theta)-f'(t))e^{-\mathrm{i}(f(t+\theta)-f(t))}$$
$$= X_H\left(F_t\left(e^{\mathrm{i}f}\right)\right)(\theta)$$

as required. □

We recover thus [64, Proposition 3.1] (up to a sign which is due to different conventions calibrating ω, $\mathscr{H}$, and b_2).

Applying Proposition 15.2 to Theorem 15.1, we get the following result:

Corollary 15.2. *The three gradient vector fields for the smooth function* $H_1 : \mathrm{L}(\mathbb{R}) \to \mathbb{R}$ *given by*

$$H_1(u) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\,(u')^2$$

are

(i) $\nabla^1 H_1(u) = u$ *for the weak inner product* $b_1(1)$ *defining the normal metric;*
(ii) $\nabla H_1(u) = -u''$ *for the weak inner product* $b(1)$ *defining the induced metric, where for* $u \in H^s(\mathbb{R})$ *with* $s \geq 2$*;*
(iii) $\nabla^2 H_1(u) = \mathscr{H}u'$ *for the weak inner product* $b_2(1)$ *defining the Kähler metric.*

Since the exponential map is a Lie group isomorphism and the three metrics coincide with the respective inner products at the identity, their left invariance guarantees that the three inner products on $\mathrm{L}(\mathbb{R})$ correspond to the three invariant metrics on $\mathrm{L}(S^1)$.

Applying Proposition 15.2 to Corollary 15.1, we conclude:

Corollary 15.3. *The Hamiltonian vector field of* H_1 *relative to the symplectic form* ω *given by* (15.15) *has the expression* $X_H(u) = -u'$*. Its flow is* $(F_t(u))(\theta) = u(\theta - t)$.

The verification of the statement about the flow is immediate:

$$\frac{d}{dt}(F_t(u))(\theta) = \frac{d}{dt}u(\theta - t) = -u'(\theta - t) = (X_H(F_t(u)))(\theta).$$

If one is willing to put more stringent hypotheses on the functional, it is possible to obtain a general result.

Theorem 15.2. *Let* $H : \mathrm{L}(S^1) \to \mathbb{R}$ *be a smooth function (with* $\mathrm{L}(S^1)$ *endowed, as usual, with the* H^s *topology for* $s \geq 1$*) and assume that the functional derivative* $\delta H/\delta u \in \mathrm{L}(S^1)$ *exists. Then the gradient vector fields are*

(i) $\nabla H(u) = \frac{\delta H}{\delta u}$ *with respect the weak inner product* $b(1)$ *defining the induced metric;*

(ii) $\left(\nabla^1 H(u)\right)(\theta) = -\int_0^\theta \mathrm{d}\varphi \left(\int_0^\varphi \mathrm{d}\psi \, \frac{\delta H}{\delta u}(\psi)\right)$ *with respect to the (weak) inner product* $b_1(1)$ *defining the normal metric, provided both* $\int_0^\theta \mathrm{d}\varphi \, \frac{\delta H}{\delta u}(\varphi)$ *as well as* $\int_0^\theta \mathrm{d}\varphi \left(\int_0^\varphi \mathrm{d}\psi \, \frac{\delta H}{\delta u}(\psi)\right)$ *are periodic;*

(iii) $\left(\nabla^2 H(u)\right)(\theta) = -\mathscr{H} \int_0^\theta \mathrm{d}\varphi \, \frac{\delta H}{\delta u}(\varphi)$ *with respect to the weak inner product* $b_2(1)$ *defining the Kähler metric, provided* $\int_0^\theta \mathrm{d}\varphi \, \frac{\delta H}{\delta u}(\varphi)$ *is periodic.*

Proof. (i) For the inner product $b(1)$ on $\mathrm{L}(S^1)$ defining the induced metric, if $u, v \in \mathrm{L}(\mathbb{R})$, we have by periodicity of u, v,

$$b(1)\left(\nabla H(u), v\right) = \mathbf{D}H(u) \cdot v = \left\langle \frac{\delta H}{\delta u}, v \right\rangle \overset{(15.12)}{=} b(1)\left(\frac{\delta H}{\delta u}, v\right).$$

This shows that $\nabla H(u) = \frac{\delta H}{\delta u}$.

(ii) For the inner product $b_1(1)$ on $\mathrm{L}(S^1)$ defining the normal metric, if $u, v \in \mathrm{L}(\mathbb{R})$, we have by periodicity of $\int_0^\theta \mathrm{d}\varphi \, \frac{\delta H}{\delta u}(\varphi)$ and $\int_0^\theta \mathrm{d}\varphi \left(\int_0^\varphi \mathrm{d}\psi \, \frac{\delta H}{\delta u}(\psi)\right)$,

$$\begin{aligned}
b_1(1)(\nabla^1 H(u), v) = \mathbf{D}H(u) \cdot v &= \left\langle \frac{\delta H}{\delta u}, v \right\rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, \frac{\delta H}{\delta u}(\theta) v(\theta) \\
&= \frac{1}{2\pi} \left(\int_0^\theta \mathrm{d}\varphi \, \frac{\delta H}{\delta u}(\varphi) \right) v(\theta) \Bigg|_{-\pi}^{\pi} \\
&\quad - \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \left(\int_0^\theta \mathrm{d}\varphi \, \frac{\delta H}{\delta u}(\varphi) \right) v'(\theta) \\
&= -\frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, \frac{d}{d\theta} \left(\int_0^\theta \mathrm{d}\varphi \left(\int_0^\varphi \mathrm{d}\psi \, \frac{\delta H}{\delta u}(\psi) \right) \right) v'(\theta) \\
&= -\left\langle \frac{d}{d\theta} \left(\int_0^\theta \mathrm{d}\varphi \left(\int_0^\varphi \mathrm{d}\psi \, \frac{\delta H}{\delta u}(\psi) \right) \right), v' \right\rangle \\
&\overset{(15.1)}{=} b_1 \left(- \int_0^\theta \mathrm{d}\varphi \left(\int_0^\varphi \mathrm{d}\psi \, \frac{\delta H}{\delta u}(\psi) \right), v \right)
\end{aligned}$$

which shows that $(\nabla^1 H(u))(\theta) = -\int_0^\theta \mathrm{d}\varphi \left(\int_0^\varphi \mathrm{d}\psi \, \frac{\delta H}{\delta u}(\psi)\right)$.

(iii) For the inner product $b_2(1)$ on $\mathrm{L}(S^1)$ defining the Kähler metric, if $u, v \in \mathrm{L}(\mathbb{R})$, we have by periodicity of $\int_0^\theta \mathrm{d}\varphi \, \frac{\delta H}{\delta u}(\varphi)$ and the isometry property of $\mathscr{H}$,

$$
\begin{aligned}
b_2(1)(\nabla^2 H(u), v) = \mathbf{D}H(u)\cdot v &= \left\langle \frac{\delta H}{\delta u}, v \right\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\, \frac{\delta H}{\delta u}(\theta) v(\theta) \\
&= \frac{1}{2\pi}\left(\int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi) \right) v(\theta)\Big|_{-\pi}^{\pi} \\
&\quad - \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta \left(\int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi) \right) v'(\theta) \\
&= -\left\langle \int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi), v' \right\rangle = -\left\langle \mathscr{H} \int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi), \mathscr{H} v' \right\rangle \\
&\overset{(15.23)}{=} b_2(1)\left(-\mathscr{H} \int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi), v \right)
\end{aligned}
$$

which shows that $\left(\nabla^2 H(u)\right)(\theta) = -\mathscr{H}\int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi)$. □

Corollary 15.4. *Under the same hypothesis as in Theorem 15.2*(iii)*, the Hamiltonian vector field of the smooth function* $H : \mathrm{L}(S^1) \to \mathbb{R}$ *relative to the symplectic form* ω *on* $\mathrm{L}(\mathbb{R})$ *given by* (15.15) *has the expression* $X_H(u) = \int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi)$

Proof. We have $X_H(u) = \mathscr{H}\nabla^2 H(u) \overset{\text{(iii)}}{=} \int_0^{\theta} \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi)$. □

Of course, using Proposition 15.2, there are immediate counterparts of Theorem 15.2 and Corollary 15.4 on the loop group $\mathrm{L}(S^1)$, which we shall not spell out explicitly.

The hypotheses guaranteeing the existence of the functional derivative of H relative to the weakly non-degenerate L^2 pairing are quite severe. For example, the theorem can be applied to the functional H_1 in Corollary 15.2, but one needs additional smoothness. Indeed, the first thing to check is if this functional has a functional derivative. In fact, it does not, unless we assume that $u \in H^s(S^1)$ for $s \geq 2$, in which case we have

$$
\begin{aligned}
\mathbf{D}H_1(u)\cdot v &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}s\, u'(s)v'(s) = \frac{1}{2\pi}u'(s)v(s)\Big|_{-\pi}^{\pi} - \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}s\, u''(s)v(s) \\
&= \langle -u'', v\rangle\,,
\end{aligned}
$$

i.e., $\delta H/\delta u = -u''$. With this additional hypothesis, the gradient flow with respect to the weak inner product $b(1)$ defining the induced metric is given by $u_t = -u''$.

Therefore, to continue computing the other two gradients of H_1, we need to assume that $u \in H^s(S^1)$ for $s \geq 2$. Provided this holds, to find the gradient relative to the (weak) inner product $b_1(1)$ defining the normal metric, we have to check that both

$$\int_0^\theta \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi) = -\int_0^\theta \mathrm{d}\varphi\, u''(\varphi) = -u'(\theta) + u'(0)$$

$$\int_0^\theta \mathrm{d}\varphi\, \left(\int_0^\varphi \mathrm{d}\psi\, \frac{\delta H}{\delta u}(\psi) \right) = -\int_0^\theta \mathrm{d}\varphi\, (u'(\varphi) - u'(0)) = -u(\theta) + u'(0)\theta$$

are periodic. While the first one is periodic, the second one is not unless we assume that $u'(0) = 0$. With this additional hypothesis, the gradient is given by $u_t = u$. However, we know from Corollary 15.2 that neither $s \geq 2$, nor $u'(0) = 0$ is needed. In addition, this can also be seen directly, as follows. For any $u, v \in \mathrm{L}(\mathbb{R})$, we have

$$b_1(1)(\nabla^1 H(u), v) = \mathbf{D}H(u) \cdot v = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}s\, u'(s) v'(s) = \langle u', v' \rangle \overset{(15.1)}{=} b_1(u, v)$$

which shows that $\nabla^1 H(u) = u$.

The same situation occurs in the computation of the third gradient. In the hypotheses of the theorem, we have

$$\left(\nabla^2 H(u)\right)(\theta) = -\mathscr{H} \int_0^\theta \mathrm{d}\varphi\, \frac{\delta H}{\delta u}(\varphi) = \mathscr{H}(u' - u'(0)) = \mathscr{H} u'$$

because the Hilbert transform of a constant is zero. Thus, the gradient flow is given in this case by

$$u_t = \mathscr{H} u' \overset{(15.20)}{=} \left(-\frac{d^2}{d\theta^2} \right)^{\frac{1}{2}} u.$$

As before, the same result can be obtained easier and without any additional hypotheses in the following way:

$$b_2(1)(\nabla^2 H(u), v) = \mathbf{D}H(u) \cdot v = \langle u', v' \rangle = \langle \mathscr{H} u', \mathscr{H} v' \rangle \overset{(15.23)}{=} b_2(1)(\mathscr{H} u', v).$$

15.3.7 Symplectic Structure on Periodic Functions

The form of the periodic Korteweg–de Vries (KdV) equation we shall use is

$$u_t - 6uu_\theta + u_{\theta\theta\theta} = 0, \tag{15.29}$$

where $u(t, \theta)$ is a real valued function of $t \in \mathbb{R}$ and $\theta \in [-\pi, \pi]$, periodic in θ, and $u_\theta := \partial u / \partial \theta$. The KdV equation is, of course, a famous integrable infinite dimensional Hamiltonian system. It is Hamiltonian on the Poisson manifold of all periodic functions relative to the Gardner bracket [29]

$$\{F, G\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, \frac{\delta F}{\delta u} \frac{d}{d\theta} \frac{\delta G}{\delta u} \,, \tag{15.30}$$

where

$$F(u) = \int_{S^1} \mathrm{d}\theta \, f(u, u_\theta, u_{\theta\theta}, \ldots)$$

and similarly for G; the functional derivative $\delta F/\delta u$ is the usual one relative to the $L^2(S^1)$ inner product, i.e.,

$$\frac{\delta F}{\delta u} = \frac{\partial f}{\partial u} - \frac{d}{d\theta} \left(\frac{\partial f}{\partial u_\theta} \right) + \frac{d^2}{d\theta^2} \left(\frac{\partial f}{\partial u_{\theta\theta}} \right) - \cdots .$$

The Hamiltonian vector field of $H(u) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, h(u, u_\theta, u_{\theta\theta}, \ldots)$ has the expression

$$X_H(u) = \frac{d}{d\theta} \left(\frac{\delta H}{\delta u} \right).$$

For the KdV equation one takes

$$H(u) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, \left(u^3 + \frac{1}{2} u_\theta^2 \right). \tag{15.31}$$

The Casimir functions of the Gardner bracket are all smooth functionals C for which $\delta C/\delta u = c$ is a constant function, i.e.,

$$C(u) = \langle c, u \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, cu(\theta) = c\hat{u}(0).$$

Thus $C^{-1}(0)$ is a candidate weak symplectic leaf in the phase space of all periodic functions. The situation in infinite dimensions is not as clear as in finite dimensions, where this would be a conclusion, because there is no general stratification theorem and one cannot expect, in general, more than a weak symplectic form. However, in our case, this actually holds, as shown in [73]. Indeed,

$$\begin{aligned} \sigma(u_1, u_2) : &= \frac{1}{4\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, \left(\int_0^\theta \mathrm{d}\varphi \, \left(u_1(\varphi) u_2(\theta) - u_2(\varphi) u_1(\theta) \right) \right) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{d}\theta \, \left(\int_0^\theta \mathrm{d}\varphi \, u_1(\varphi) \right) u_2(\theta) = \left\langle \int_0^\theta \mathrm{d}\varphi \, u_1(\varphi), u_2 \right\rangle \end{aligned} \tag{15.32}$$

defines a weak symplectic form on $\mathrm{L}(\mathbb{R})$ whose formal Poisson bracket is (15.30). This immediately shows that there is a tight relationship with the symplectic form ω of the complex Hilbert space $\mathrm{L}(\mathbb{R})$, the Lie algebra of the based loop groups, given by (15.15), namely

$$\sigma\left(\frac{d^2}{d\theta^2}u, v\right) = \omega(u, v)$$

for all $u, v \in \mathrm{L}(\mathbb{R})$ of class H^s, $s \geq 2$. Defining

$$\left(\frac{d}{d\theta}\right)^{-1} u := \int_0^\theta \mathrm{d}\varphi\, u(\varphi),$$

the KdV symplectic form σ has the suggestive expression (see (15.28))

$$\sigma(u_1, u_2) = \left\langle \left(\frac{d}{d\theta}\right)^{-1} u_1, u_2 \right\rangle ,$$

which is well defined on $H^{-\frac{1}{2}}(S^1, \mathbb{R})$.

On the other hand, the Poisson bracket given by the Kähler symplectic form (15.15) on $\mathrm{L}(\mathbb{R})$ is

$$\{F, G\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\, \frac{\delta F}{\delta u}\left(\frac{d}{d\theta}\right)^{-1}\frac{\delta G}{\delta u}, \tag{15.33}$$

which is similarly well defined on $H^{-\frac{1}{2}}$, and the Hamiltonian vector field defined by this bracket is given by Corollary 15.4, i.e.,

$$u_t = X_H(u) = \left(\frac{d}{d\theta}\right)^{-1}\frac{\delta H}{\delta u}. \tag{15.34}$$

Now, the gradient vector field for the corresponding Kähler metric, as computed in Theorem 15.2(iii), is written as

$$u_t = -\mathscr{H}\left(\frac{d}{d\theta}\right)^{-1}\frac{\delta H}{\delta u}. \tag{15.35}$$

15.4 Metriplectic Systems

In this section we define metriplectic systems and show how to construct general classes of such systems in terms of triple brackets for both finite- and infinite-dimensional theories. We use some of the machinery developed above to address specific examples.

15.4.1 Definition and Consequences

A *metriplectic system* consists of a smooth manifold P, two smooth vector bundle maps $\pi, \kappa : T^*P \to TP$ covering the identity, and two functions $H, S \in C^\infty(P)$, the *Hamiltonian* or *total energy* and the *entropy* of the system, such that

(i) $\{F, G\} := \langle \mathbf{d}F, \pi(\mathbf{d}G)\rangle$ is a Poisson bracket; in particular $\pi^* = -\pi$;
(ii) $(F, G) := \langle \mathbf{d}F, \kappa(\mathbf{d}G)\rangle$ is a positive semidefinite symmetric bracket, i.e., $(\,,\,)$ is $\mathbb{R}$-bilinear and symmetric, so $\kappa^* = \kappa$, and $(F, F) \geq 0$ for every $F \in C^\infty(P)$;
(iii) $\{S, F\} = 0$ and $(H, F) = 0$ for all $F \in C^\infty(P) \iff \pi(\mathbf{d}S) = \kappa(\mathbf{d}H) = 0$.

The *metriplectic dynamics* of the system is given in terms of the two brackets by

$$\frac{d}{dt}F = \{F, H + S\} + (F, H + S) = \{F, H\} + (F, S), \quad \text{for all} \quad F \in C^\infty(P), \tag{15.36}$$

or, equivalently, as an ordinary differential equation, by

$$\frac{d}{dt}c(t) = \pi(c(t))\mathbf{d}H(c(t)) + \kappa(c(t))\mathbf{d}S(c(t)). \tag{15.37}$$

The Hamiltonian vector field $X_H := \pi(\mathbf{d}H) \in \mathfrak{X}(P)$ represents the *conservative* or *Hamiltonian part*, whereas $Y_S := \kappa(\mathbf{d}S) \in \mathfrak{X}(P)$ the *dissipative part* of the full metriplectic dynamics (15.36) or (15.37).

As far as we know, the first attempts to introduce such a structure were given in adjacent papers in [42, 51]. (See also [41].) Kaufman [42] imposed, instead of (iii), the weaker condition $\{H, S\} = (H, S) = 0$, which is enough, as will become apparent below, to deduce the First and Second Laws of Thermodynamics. In the plasma examples presented, he used (iii) for a large class of functions. All three axioms, including the degeneracy condition of (iii), were stated explicitly by Morrison in [51, 52]. The former treated the same kinetic example as [42] along with additional formalism, while the latter presented the metriplectic formalism for the compressible Navier–Stokes equations with entropy production. All three axioms were restated in [53], where the terminology metriplectic was introduced and a detailed physical motivation for the introduction of (iii) is presented along with other examples such as a dissipative free rigid body equation and the Vlasov–Poisson equation with a collision term that generalizes the Landau and Balescu–Lenard equations. In [31], under the name GENERIC (General Equations for Non-Equilibrium Reversible Irreversible Coupling), the same geometric structure was used to analyze many other equations; due to this paper and subsequent work of these authors, the metriplectic formalism has been popularized. For a very interesting modern application of this structure see [49] and for further discussion about avenues for generalization see [55].

The definition of metriplectic systems has three immediate important consequences. Let $c(t)$ be an integral curve of the system (15.37).

1. *Energy conservation*:

$$\frac{d}{dt}H(c(t)) = \{H, H\}(c(t)) + (H, S)(c(t)) = 0. \tag{15.38}$$

2. *Entropy production*:

$$\frac{d}{dt}S(c(t)) = \{S, H\}(c(t)) + (S, S)(c(t)) \geq 0. \tag{15.39}$$

3. *Maximum entropy principle yields equilibria*: Suppose that there are n functions $C_1, \ldots, C_n \in C^\infty(P)$ such that $\{F, C_i\} = (F, C_i) = 0$ for all $F \in C^\infty(P)$, i.e., these functions are simultaneously conserved by the conservative and dissipative part of the metriplectic dynamics. Let $p_0 \in P$ be a maximum of the entropy S subject to the constraints $H^{-1}(h) \cap C_1^{-1}(c_1) \cap \ldots C_n^{-1}(c_n)$, for given regular values $h, c_1, \ldots, c_n \in \mathbb{R}$ of $H, C_1, \ldots, C_n$, respectively. By the Lagrange Multiplier Theorem, there exist $\alpha, \beta_1, \ldots, \beta_n \in \mathbb{R}$ such that

$$\mathbf{d}S(p_0) = \alpha \mathbf{d}H(p_0) + \beta_1 \mathbf{d}C_1(p_0) + \cdots + \mathbf{d}C_n(p_0).$$

But then, assuming that $\alpha \neq 0$, for every $F \in C^\infty(P)$, we have

$$\begin{aligned}
&\{F, H\}(p_0) + (F, S)(p_0) \\
&= \langle \mathbf{d}F(p_0), \pi(p_0)\,(\mathbf{d}H(p_0))\rangle + \langle \mathbf{d}F(p_0), \kappa(p_0)\,(\mathbf{d}S(p_0))\rangle \\
&= \big\langle \mathbf{d}F(p_0), \tfrac{1}{\alpha}\pi(p_0)\,(\mathbf{d}S(p_0) - \beta_1 \mathbf{d}C_1(p_0) - \cdots - \mathbf{d}C_n(p_0))\big\rangle \\
&\qquad + \langle \mathbf{d}F(p_0), \kappa(p_0)\,(\alpha \mathbf{d}H(p_0) + \beta_1 \mathbf{d}C_1(p_0) + \cdots + \mathbf{d}C_n(p_0))\rangle \\
&= \tfrac{1}{\alpha}\{F, S\}(p_0) - \tfrac{\beta_1}{\alpha}\{F, C_1\}(p_0) - \cdots - \tfrac{\beta_n}{\alpha}\{F, C_n\}(p_0) \\
&\qquad + \alpha(F, H)(p_0) + \beta_1(F, C_1)(p_0) + \cdots + \beta_n(F, C_n)(p_0) = 0
\end{aligned}$$

which means that p_0 is an equilibrium of the metriplectic dynamics (15.36) or (15.37). This is akin to the free energy extremization of thermodynamics, as noted in [52, 53] where it was suggested that one can build in degeneracies associated with Hamiltonian "dynamical constraints." (See also [49].)

Suppose that $K \in C^\infty(P)$ is a conserved quantity for the Hamiltonian part of the metriplectic dynamics, i.e., $\{K, H\} = 0$. Then, if $c(t)$ is an integral curve of the metriplectic dynamics, we have

$$\begin{aligned}
\frac{d}{dt}K(c(t)) = \mathbf{d}K(c(t))\,(\dot{c}(t)) &= \langle \mathbf{d}F(c(t)), \pi(c(t))\,(\mathbf{d}H(c(t)))\rangle \\
&\quad + \langle \mathbf{d}F(c(t)), \kappa(c(t))\,(\mathbf{d}S(c(t)))\rangle \\
&= \{K, H\}(c(t)) + (K, S)(c(t)) = (K, S)(c(t)).
\end{aligned}$$

As pointed out in [53], this immediately implies that a function that is simultaneously conserved for the full metriplectic dynamics and its Hamiltonian part, is

necessarily conserved for the dissipative part. Physically, it is advantageous for general metriplectic systems to conserve dynamical constraints, i.e., conserved quantitates of its Hamiltonian part and the examples given in [42, 51–53] satisfy this condition.

15.4.2 Metriplectic Systems Based on Lie Algebra Triple Brackets

Associated with any quadratic Lie algebra (i.e., a Lie algebra admitting a bilinear symmetric invariant form) is a natural completely antisymmetric triple bracket. This is used to construct Lie algebra-based metriplectic systems. The algebra $\mathfrak{so}(3)$ is worked out explicitly and examples are given.

15.4.2.1 General Theory

A quadratic Lie algebra is, by definition, a Lie algebra admitting a bilinear symmetric non-degenerate invariant form $\kappa : \mathfrak{g} \times \mathfrak{g} \to \mathbb{R}$ (the letter κ is meant to remind one of the Killing form in a semisimple Lie algebra). Recall that invariance means that $\kappa([\xi, \eta], \zeta) = \kappa(\xi, [\eta, \zeta])$ for all $\xi, \eta, \zeta \in \mathfrak{g}$ or, equivalently, that the adjoint operators ad_η for all $\eta \in \mathfrak{g}$ are antisymmetric relative to κ. Non-degeneracy (strong) means that the map $\mathfrak{g} \ni \xi \mapsto \kappa(\xi, \cdot) \in \mathfrak{g}^*$ is an isomorphism. Finite dimensional quadratic Lie algebras have been completely classified in [48]. For finite dimensional Lie algebras, non-degeneracy is equivalent to the following statement: $\kappa(\xi, \eta) = 0$ for all $\eta \in \mathfrak{g}$ if and only if $\xi = 0$. In infinite dimensions this condition is called weak non-degeneracy and it is implied by non-degeneracy but the converse is, in general, false.

For example, let $\mathfrak{g}$ be an arbitrary finite dimensional Lie algebra. Recall that the Killing form is defined by $\kappa(\xi, \eta) := \mathrm{Trace}(\mathrm{ad}_\xi \circ \mathrm{ad}_\eta)$. If $\{e_i\}$, $i = 1, \ldots \dim \mathfrak{g}$, is an arbitrary basis of $\mathfrak{g}$ and ${c^p}_{ij}$ are the structure constants of $\mathfrak{g}$, i.e., $[e_i, e_j] = {c^p}_{ij} e_p$, then

$$\kappa(\xi, \eta) = \xi^i {c^p}_{iq} \eta^j {c^q}_{jp}$$

and hence the components of κ in the basis $\{e_i\}$, $i = 1, \ldots \dim \mathfrak{g}$, are given by

$$\kappa_{ij} = \kappa(e_i, e_j) = {c^p}_{iq} {c^q}_{jp}.$$

The Killing form is bilinear symmetric and invariant; it is non-degenerate if and only if $\mathfrak{g}$ is semisimple. Moreover, $-\kappa$ is a positive definite inner product if and only if the Lie algebra $\mathfrak{g}$ is compact (i.e., it is the Lie algebra of a compact Lie group).

In general, let κ be a bilinear symmetric non-degenerate invariant form and define the completely antisymmetric covariant 3-tensor

$$c(\xi,\eta,\zeta) := \kappa(\xi,[\eta,\zeta]) = -c(\xi,\zeta,\eta) = -c(\eta,\xi,\zeta) = -c(\zeta,\eta,\xi).$$

In the coordinates given by the basis $\{e_i\}$, $i = 1,\dots \dim \mathfrak{g}$, the components of c are

$$c_{ijk} := \kappa_{im} c^m{}_{jk} = -c_{ikj} = -c_{jik} = -c_{kji}.$$

This construction immediately leads to the triple bracket introduced in [10] (see also [54]), $\{\cdot,\cdot,\cdot\} : C^\infty(\mathfrak{g}) \times C^\infty(\mathfrak{g}) \times C^\infty(\mathfrak{g}) \to C^\infty(\mathfrak{g})$ defined by

$$\{f,g,h\}(\xi) := c(\nabla f(\xi), \nabla g(\xi), \nabla h(\xi)) := \kappa\left(\nabla f(\xi), [\nabla g(\xi), \nabla h(\xi)]\right), \tag{15.40}$$

where the gradient is taken relative to the non-degenerate bilinear form κ, i.e., for any $\xi \in \mathfrak{g}$ we have

$$\kappa(\nabla f(\xi), \cdot) := \mathbf{d} f(\xi),$$

or, in coordinates

$$\nabla^i f(\xi) = \kappa^{ij} \frac{\partial f}{\partial \xi^i},$$

where $[\kappa^{ij}] = [\kappa_{kl}]^{-1}$, i.e., $\kappa^{ij}\kappa_{jk} = \delta^i_k$. This triple bracket is trilinear over $\mathbb{R}$, completely antisymmetric, and satisfies the Leibniz rule in any of its variables. In coordinates it is given by

$$\begin{aligned}\{f,g,h\} &= c_{ijk}\nabla^i f \nabla^j g \nabla^k h = \kappa_{im} c^m{}_{jk} \kappa^{ip} \frac{\partial f}{\partial \xi^p} \kappa^{jq} \frac{\partial g}{\partial \xi^q} \kappa^{kr} \frac{\partial h}{\partial \xi^r} \\ &= c^p{}_{jk} \kappa^{jq} \kappa^{kr} \frac{\partial f}{\partial \xi^p} \frac{\partial g}{\partial \xi^q} \frac{\partial h}{\partial \xi^r} = c^{pqr} \frac{\partial f}{\partial \xi^p} \frac{\partial g}{\partial \xi^q} \frac{\partial h}{\partial \xi^r},\end{aligned}$$

where c^{pqr} are the components of the contravariant completely antisymmetric 3-tensor $\bar{c}$ associated with c by raising its indices with the non-degenerate symmetric bilinear form κ, i.e., for any $\xi,\eta,\zeta \in \mathfrak{g}$, we have

$$\bar{c}\left(\kappa(\xi,\cdot), \kappa(\eta,\cdot), \kappa(\gamma,\cdot)\right) := c(\xi,\eta,\zeta).$$

This construction extends the bracket due to Nambu [57] to a Lie algebra setting. Nambu considered ordinary vectors in $\mathbb{R}^3$ and defined

$$\{f,g,h\}_{\text{Nambu}}(\boldsymbol{\Pi}) = \nabla f(\boldsymbol{\Pi}) \cdot \left(\nabla g(\boldsymbol{\Pi}) \times \nabla h(\boldsymbol{\Pi})\right), \tag{15.41}$$

where "$\cdot$" and "$\times$" are the ordinary dot and cross products. Thus, the Nambu bracket is a special case of the triple bracket (15.40) in the case of $\mathfrak{g} = \mathfrak{so}(3)$, whose structure constants are the completely antisymmetric Levi–Civita symbols ϵ_{ijk}. Such "modified rigid body brackets" were also described in [19, 34, 47].

If $\mathfrak{g}$ is an arbitrary quadratic Lie algebra with bilinear symmetric non-degenerate invariant form κ, the quadratic function

$$C_2(\xi) := \tfrac{1}{2}\kappa(\xi, \xi) \tag{15.42}$$

is a Casimir function for the Lie–Poisson bracket on $\mathfrak{g}$, identified with $\mathfrak{g}^*$ via κ, i.e.,

$$\{f, g\}_{\pm}(\xi) = \pm\kappa\left(\xi, [\nabla f(\xi), \nabla g(\xi)]\right), \tag{15.43}$$

as an easy verification shows, since $\nabla C_2(\xi) = \xi$. In view of (15.43), the following identity is obvious

$$\{f, g\}_{+} = \{C_2, f, g\}$$

(this was first pointed out in [10]). For example, if $\mathfrak{g} = \mathfrak{so}(3)$, the $(-)$Lie–Poisson bracket

$$\{f, g\}_{-}^{\mathfrak{so}(3)}(\boldsymbol{\Pi}) = -\{C_2, f, g\}_{\text{Nambu}}(\boldsymbol{\Pi}) = -\boldsymbol{\Pi} \cdot (\nabla f(\boldsymbol{\Pi}) \times \nabla g(\boldsymbol{\Pi})) \tag{15.44}$$

is the rigid body bracket, i.e., if $h(\boldsymbol{\Pi}) = \frac{1}{2}\boldsymbol{\Pi} \cdot \boldsymbol{\Omega}$, where $\boldsymbol{\Pi}_i = I_i\boldsymbol{\Omega}_i$, $I_i > 0$, $i = 1, 2, 3$, and I_i are the principal moments of inertia of the body, then Hamilton's equations $\frac{d}{dt}F(\boldsymbol{\Pi}) = \{f, h\}_{-}^{\mathfrak{so}(3)}(\boldsymbol{\Pi})$ are equivalent to Euler's equations $\dot{\boldsymbol{\Pi}} = \boldsymbol{\Pi} \times \boldsymbol{\Omega}$.

Note that given any two functions, $f, g \in C^\infty(\mathfrak{g})$, because the triple bracket satisfies the Leibniz identity in every factor, the map $C^\infty(\mathfrak{g}) \ni h \mapsto \{h, f, g\} \in C^\infty(\mathfrak{g})$ is a derivation and hence defines a vector field on $\mathfrak{g}$, denoted by $X_{f,g} : \mathfrak{g} \to \mathfrak{g}$, i.e.,

$$\left\langle \mathbf{d}h(\xi), X_{f,g}(\xi) \right\rangle = \kappa\left(\nabla h(\xi), X_{f,g}(\xi)\right) = \{h, f, g\}(\xi) \qquad \text{for all} \qquad h \in C^\infty(\mathfrak{g}). \tag{15.45}$$

Note that $X_{f,f} = 0$. Thus, for triple brackets, two functions define a vector field, analogous to the Hamiltonian vector field defined by a single function associated with a standard Poisson bracket.

From (15.40) we have the following result.

Proposition 15.3. *The vector field $X_{f,g}$ on $\mathfrak{g}$ corresponding to the pair of functions f, g is given by*

$$X_{f,g}(\xi) = [\nabla f(\xi), \nabla g(\xi)]. \tag{15.46}$$

Triple brackets of the form (15.40) can be used to construct metriplectic systems on a quadratic Lie algebra $\mathfrak{g}$ in the following manner. Let κ be the bilinear symmetric non-degenerate form on $\mathfrak{g}$ defining the quadratic structure and fix some $h \in C^\infty(\mathfrak{g})$. Define the symmetric bracket

$$(f,g)_h^\kappa(\xi) := -\kappa\left(X_{h,f}(\xi), X_{h,g}(\xi)\right). \tag{15.47}$$

Assume that $-\kappa$ is a positive definite inner product. Then $(f,f) \geq 0$. Thus we have the manifold $\mathfrak{g}$ endowed with the Lie–Poisson bracket (15.43), the symmetric bracket (15.47), the Hamiltonian h, and for the entropy S we take any Casimir function of the Lie–Poisson bracket. Then the conditions (i)–(iii) of Sect. 15.4.1 are all satisfied, because $(h,g)_h^\kappa = -\kappa(X_{h,h}, X_{h,g}) = -\kappa(0, X_{h,g}) = 0$ for any $g \in C^\infty(\mathfrak{g})$. The equations of motion (15.36) are in this case given by

$$\begin{aligned}\frac{d}{dt}f(\xi) &= \kappa\left(\nabla f(\xi), \frac{d}{dt}\xi\right) = \{f,h\}_\pm(\xi) + (f,S)(\xi)\\ &= \pm\kappa\left(\xi, [\nabla f(\xi), \nabla h(\xi)]\right) - \kappa\left(X_{h,f}(\xi), X_{h,S}(\xi)\right)\\ &= \mp\kappa\left(\nabla f(\xi), [\xi, \nabla h(\xi)]\right) - \kappa\left([\nabla h(\xi), \nabla f(\xi)], [\nabla h(\xi), \nabla S(\xi)]\right)\end{aligned}$$

for any $f \in C^\infty(\mathfrak{g})$.

This gives the equations of motion

$$\dot{\xi} = \pm[\xi, \nabla h(\xi)] + [\nabla h(\xi), [\nabla h(\xi), \nabla S(\xi)]]. \tag{15.48}$$

Note that the flow corresponding to S is a generalized double bracket flow. Observe also that this flow reduces to a double bracket flow and is tangent to an orbit of the group if $\nabla h(\xi) = \xi$. Indeed if $h = \frac{1}{2}\kappa(\xi,\xi)$ the symmetric bracket (15.47) reduces to the symmetric bracket induced from the normal metric.

We remark that flows with a similar structure are discussed in the fluid dynamics setting in [30]. We discuss the case of PDEs below.

15.4.2.2 Special Case of $\mathfrak{so}(3)$

If the quadratic Lie algebra is $\mathfrak{so}(3)$, we identify it with $\mathbb{R}^3$ with the cross product as Lie bracket via the Lie algebra isomorphism $\hat{}: \mathbb{R}^3 \to \mathfrak{so}(3)$ given by $\hat{\mathbf{u}}\mathbf{v} := \mathbf{u} \times \mathbf{v}$ for all $\mathbf{u}, \mathbf{v} \in \mathbb{R}^3$. Since $\mathrm{Ad}_A\,\hat{\mathbf{u}} = \widehat{A\mathbf{u}}$, for any $A \in SO(3)$ and $\mathbf{u} \in \mathbb{R}^3$, we conclude that the usual inner product on $\mathbb{R}^3$ is an invariant inner product. In terms of elements of $\mathfrak{so}(3)$ we have $\mathbf{u}\cdot\mathbf{v} = -\frac{1}{2}\,\mathrm{Trace}\,(\hat{\mathbf{u}}\hat{\mathbf{v}})$. We shall show below that the metriplectic structure on $\mathbb{R}^3$ is precisely the one given in [53].

Recall that the Nambu bracket is given for $\mathfrak{so}(3)$ by (15.41) and hence the symmetric bracket (15.47) has the form

$$\begin{aligned}\kappa(\{\Pi,h,f\},\{\Pi,h,g\}) &= \epsilon^{imn}\frac{\partial h}{\partial \Pi^m}\frac{\partial f}{\partial \Pi^n}\,\delta_{ij}\,\epsilon^{jst}\frac{\partial h}{\partial \Pi^s}\frac{\partial g}{\partial \Pi^t}\\ &= \epsilon^{imn}\,\epsilon_i^{\,st}\,\frac{\partial h}{\partial \Pi^m}\frac{\partial f}{\partial \Pi^n}\frac{\partial h}{\partial \Pi^s}\frac{\partial g}{\partial \Pi^t}\\ &= \|\nabla h\|^2\nabla g\cdot\nabla f-(\nabla f\cdot\nabla h)(\nabla g\cdot\nabla h)\end{aligned}\tag{15.49}$$

where in the third equality we have used the identity $\epsilon^{imn}\epsilon_i^{\,st}=\delta^{ms}\delta^{nt}-\delta^{mt}\delta^{ns}$. This coincides with [53, equation (31)].

With the choice $S(\boldsymbol{\Pi})=\|\boldsymbol{\Pi}\|^2/2$ and the usual rigid body Hamiltonian, the equations of motion (15.48) are those for the relaxing rigid body given in [53].

Comments

- In three dimensions any Poisson bracket can be written as

$$\{f,g\}=J^{ij}\frac{\partial f}{\partial \Pi^i}\frac{\partial g}{\partial \Pi^j}=\epsilon^{ij}{}_kV^k(\boldsymbol{\Pi})\frac{\partial f}{\partial \Pi^i}\frac{\partial g}{\partial \Pi^j}\tag{15.50}$$

  where $i,j,k=1,2,3$, and $V\in\mathbb{R}^3$. The last equality follows from the identification of 3×3 antisymmetric matrices with vectors (the hat map discussed above). Using the well-known fact (which is easy to show directly) that brackets of the form of (15.50) satisfy the Jacobi identity if

$$V\cdot\nabla\times V=0\,,\tag{15.51}$$

  we conclude that

$$\{F,G\}_f=\{f,F,G\}_{\text{Nambu}}\tag{15.52}$$

  satisfies the Jacobi identity for *any* smooth function f; i.e., unlike the general case where the theorem of [10] requires f to be the quadratic Casimir, one obtains a good Poisson bracket for any f. Thus, for the special case of three dimensions, one can interchange the roles of Hamiltonian and entropy in the metriplectic formalism.
- Thinking in terms of $\mathfrak{so}(3)^*$, the setting arising from reduction (see, e.g., [47]), this construction leads to a natural geometric interpretation of a metriplectic system on the manifold $P=\mathbb{R}^3$. With the Poisson bracket on $\mathbb{R}^3$ of (15.52), the bundle map $\pi: T^*\mathbb{R}^3\to T\mathbb{R}^3$ has the expression

$$\pi_f(x,\Pi)=\left(x,\nabla f(\Pi)\times(\cdot)^\top\right)$$

since $\mathbf{d}H(\Pi)^\top = \nabla H(\Pi)$ ($\mathbf{d}H(\Pi)$ is a row vector and $\nabla H(\Pi)$ is its transpose, a column vector). Now the triple bracket associated with Eq. (15.48) can be used to generate a symmetric bracket given in [17] as follows:

$$\begin{aligned}(F,G)_{BKMR}(\Pi) &= (F,G)_C^\kappa = \kappa(\{\Pi, C, F\}, \{\Pi, C, G\}) \\ &= (\Pi \times \nabla F(\Pi)) \cdot (\Pi \times \nabla G(\Pi)) \ . \end{aligned} \tag{15.53}$$

where now $C = ||\Pi||^2/2$. Hence the bundle map $\kappa : T^*\mathbb{R}^3 \to T\mathbb{R}^3$ has the expression

$$\kappa(x, \Pi) = -\Pi \times \left(\Pi \times (\cdot)^\top\right) .$$

Thus, with the freedom to choose any quantity $S = f$ as an entropy, with the assurance that (15.51) will be satisfied because $\nabla \times V = \nabla \times \nabla f = 0$, we can take $H = C$ and have $\{F, S\}_f = 0$ and $(F, H) = 0$ for all $F \in C^\infty(\mathbb{R}^3)$. The equations of motion for this metriplectic system are

$$\dot{\Pi} = -\Pi \times \nabla f(\Pi) - \Pi \times (\Pi \times \nabla f(\Pi)). \tag{15.54}$$

The symmetric bracket is the inner product of the two Hamiltonian vector fields on each concentric sphere. As discussed in [17], this symmetric bracket can be defined on any compact Lie algebra by taking the normal metric on each coadjoint orbit.

- The following set of equations were given in [26]:

$$\dot{\boldsymbol{\Pi}} = \nabla S(\boldsymbol{\Pi}) \times \nabla H(\boldsymbol{\Pi}) - \nabla H(\boldsymbol{\Pi}) \times (\nabla H(\boldsymbol{\Pi}) \times \nabla S(\boldsymbol{\Pi})). \tag{15.55}$$

Yet, this metriplectic system is identical to that obtained from (15.48), using (15.49), viz.

$$\dot{\boldsymbol{\Pi}} = \{\boldsymbol{\Pi}, S, H\} + \kappa\left(\{\boldsymbol{\Pi}, H, \Pi\}, \{\boldsymbol{\Pi}, H, S\}\right) , \tag{15.56}$$

Replacing H by g in (15.49) gives

$$\begin{aligned}(F,G)_g((\boldsymbol{\Pi})) &= \kappa\left(\{(\boldsymbol{\Pi}), g, F\}, \{(\boldsymbol{\Pi}), g, G\}\right) \\ &= (\nabla g(\Pi) \times \nabla F(\Pi)) \cdot (\nabla g(\Pi) \times \nabla G(\Pi)). \end{aligned} \tag{15.57}$$

Thus, the bundle map $\kappa : T^*\mathbb{R}^3 \to T\mathbb{R}^3$ has the expression

$$\kappa_g(x, \boldsymbol{\Pi}) = -\nabla g(\boldsymbol{\Pi}) \times \left(\nabla \boldsymbol{\Pi} \times (\cdot)^\top\right) .$$

Examples. Two special cases of the equation (15.55) are of interest.

(i) If we take $H = \frac{1}{2}\|\boldsymbol{\Pi}\|^2$ and $S = c \cdot \boldsymbol{\Pi}$, c a constant vector, we obtain

$$\dot{\boldsymbol{\Pi}} = c \times \boldsymbol{\Pi} - \boldsymbol{\Pi} \times (\boldsymbol{\Pi} \times c). \tag{15.58}$$

(ii) If we take $S = \frac{1}{2}\|\boldsymbol{\Pi}\|^2$ and $H = c \cdot \boldsymbol{\Pi}$, c a constant, we obtain

$$\dot{\boldsymbol{\Pi}} = \boldsymbol{\Pi} \times c - c \times (c \times \boldsymbol{\Pi})\,. \tag{15.59}$$

The equation of motion (15.58) is an instance of double bracket damping, where the damping is due to the normal metric, whereas (15.59) gives linear damping of the sort arising in quantum systems.

- As the above examples indicate, given a Hamiltonian (or Poisson) structure on a Riemannian manifold M, it is natural to construct a symmetric bracket of the form $(F, G) = g(X_F, X_G)$ where X_F denotes the Hamiltonian vector field corresponding to the function F and g is the metric as discussed in [28].

15.4.3 The Toda System Revisited

15.4.3.1 The Toda Lattice Equation Revisited

We note that the Toda lattice equation fits into the metriplectic picture in a degenerate but interesting fashion since it has a dual Hamiltonian and gradient character which may be seen by writing it in the double bracket form (15.2).

It may be viewed either as the Hamiltonian part or the dissipative part of a metriplectic system with Hamiltonian $H = \frac{1}{2}\operatorname{Tr} L^2$ or entropy function $S = \operatorname{Tr} LN$, respectively, with the Toda lattice equations in the corresponding form (15.7) or (15.2), as discussed in Sect. 15.2. This observation may be extended to the Toda lattice flow on the normal form of any complex semisimple Lie algebra as can be seen in [14].

15.4.3.2 Full Toda with Dissipation

It is possible to construct an interesting metriplectic system which incorporates the full Toda dynamics.

We consider again the flow on the vector space of symmetric matrices $\mathfrak{k}^{\perp} = \mathfrak{sym}(n)$ but now restrict the flow to a generic orbit as discussed in [25] where it was shown that the flow is integrable. The Hamiltonian is again $\frac{1}{2}\operatorname{Tr} L^2$ and the flow on full symmetric matrices is given by

$$\dot{L} = [\pi_{\mathfrak{s}} L, L] \tag{15.60}$$

with $\pi_{\mathfrak{s}}$ being the projection onto the skew symmetric matrices in the lower triangular–skew decomposition of a matrix. In this setting there are nontrivial Casimir functions of the bracket (15.9). These are given as follows. For L an $n \times n$ symmetric matrix set for $0 \leq k \leq [\frac{1}{2}n]$

$$\det(L - \lambda)_k = \sum_{r-0}^{n-2k} E_{rk}(L)\lambda^{n-2k-r} \tag{15.61}$$

where the subscript k denotes the matrix obtained by deleting the first k rows and the last k columns. Then $I_{1k}(L) = E_{1k}(L)/E_{0k}(L)$ are Casimir functions of the generic orbit in $\mathfrak{sym}(n)$ as shown in [25].

Thus we obtain the metriplectic systems

$$\dot{L} = [\pi_{\mathfrak{s}} L, L] + [L, [L, \nabla I_{1k}]] \tag{15.62}$$

where the metric is the normal metric on orbits of $\mathfrak{su}(n)$ restricted to the symmetric matrices (identified with i times the symmetric matrices) as in [14]. Here $H = \frac{1}{2}\,\mathrm{Tr}\; L^2$ and $S = I_{1k}$.

15.4.4 Metriplectic Systems for PDEs: Metriplectic Brackets and Examples

First we construct a class of metriplectic brackets based on triple brackets for infinite systems, then we consider in detail an example based on Gardner's bracket on S^1. Lastly, we mention various generalizations.

15.4.4.1 Symmetric Brackets for PDEs Based on Triple Brackets

Similary to Sect. 15.4.2, we can construct metriplectic flows for infinite-dimensional systems from completely antisymmetric triple brackets of the form

$$\begin{aligned}&\{E, F, G\}\\ &= \textstyle\int_{S^1} \mathrm{d}\theta_1 \int_{S^1} \mathrm{d}\theta_2 \int_{S^1} \mathrm{d}\theta_3\, \mathscr{C}_{ijk}(\theta_1, \theta_2, \theta_3)\, (\mathscr{P}^i E_u)(\theta_1)\, (\mathscr{P}^j F_u)(\theta_2)\, (\mathscr{P}^k G_u)(\theta_3)\end{aligned} \tag{15.63}$$

where E, F, and G are smooth functions on S^1, $\mathscr{C}_{ijk}$ is a smooth function on $S^1 \times S^1 \times S^1$ which is completely antisymmetric in its arguments, so as to assure complete antisymmetry of $\{E, F, G\}$. In addition, we denote $E_u := \delta E/\delta u$, etc. Let $\mathscr{P}^i$, $i = 1, 2, 3$, be pseudo-differential operators. Evidently, the triple bracket of (15.63) is trilinear and completely antisymmetric in E, F, G.

From (15.63) and a Hamiltonian H, we construct a symmetric bracket as follows:

$$(F,G)_H = \int_{S^1} \mathrm{d}\theta' \int_{S^1} \mathrm{d}\theta'' \, \{U(\theta'), H, F\} \, \mathscr{G}(\theta', \theta'') \, \{U(\theta''), H, G\} \,, \tag{15.64}$$

where $U(\theta)$ in (15.64) denotes the functional

$$U(\theta) : u \mapsto \int_{S^1} \mathrm{d}\theta' u(\theta') \delta(\theta - \theta'). \tag{15.65}$$

We shall use this notation in subsequent expressions below. The "metric" $\mathscr{G}$ is assumed to be symmetric and positive semidefinite, i.e., the smooth function $\mathscr{G} : S^1 \times S^1 \to \mathbb{R}$ satisfies $\mathscr{G}(\theta', \theta'') = \mathscr{G}(\theta'', \theta')$ and

$$\int_{S^1} \mathrm{d}\theta' \int_{S^1} \mathrm{d}\theta'' \, \mathscr{G}(\theta', \theta'') f(\theta') f(\theta'') \geq 0 \tag{15.66}$$

for all functions $f \in C^\infty(S^1)$. Therefore, by construction, it is clear that (15.64) satisfies the following:

(i) $(F,G)_H = (G,F)_H$ for all F, G,
(ii) $(F,H)_H = 0$ for all F, and
(iii) $(F,F)_H \geq 0$ for all F.

As a special case, suppose $\mathscr{P}^i = \mathscr{P}$ for all $i = 1, 2, 3$; then (15.63) becomes

$$\{E,F,G\} = \int_{S^1} \mathrm{d}\theta_1 \int_{S^1} \mathrm{d}\theta_2 \int_{S^1} \mathrm{d}\theta_3 \, \mathscr{C}(\theta_1, \theta_2, \theta_3) \, \mathscr{P}(\theta_1) E_u \; \mathscr{P}(\theta_2) F_u \; \mathscr{P}(\theta_3) G_u \,. \tag{15.67}$$

As a further specialization, suppose $\mathscr{C}(\theta_1, \theta_2, \theta_3)$ is given by

$$\mathscr{C}(\theta_1, \theta_2, \theta_3) = A(\theta_1, \theta_2) + A(\theta_2, \theta_3) + A(\theta_3, \theta_1), \tag{15.68}$$

where A is any antisymmetric function, i.e.,

$$A(\theta_1, \theta_2) = -A(\theta_2, \theta_1) \,. \tag{15.69}$$

The form (15.68), assuming (15.69), assures complete antisymmetry of $\mathscr{C}$.

Finally, a particularly interesting self-contained case would be to suppose that the A's come from some Poisson bracket, according to

$$A(\theta_1, \theta_2) = \{U(\theta_1), U(\theta_2)\} \,. \tag{15.70}$$

It would be quite natural to choose the entropy, S, to be a Casimir function of this bracket and to choose this bracket as the Hamiltonian part of the metriplectic system

with symmetric bracket given by (15.64). We give an example of this construction in Sect. 15.4.4.2.

It is evident that one can construct a wide variety of symmetric brackets based on triple brackets. For example, one can choose the pseudo-differential operators from the list $\{\mathscr{I}, d/d\theta, (d/d\theta)^{-1}, \mathscr{H}\}$, where $\mathscr{I}$ is the identity operator, and the Hamiltonian, H, and entropy (Casimir) C could be one of the following functionals:

$$H_0 = \int_{S^1} \mathrm{d}\theta\, u \tag{15.71}$$

$$H_2 = \int_{S^1} \mathrm{d}\theta\, u^2/2 \tag{15.72}$$

$$H_1 = \int_{S^1} \mathrm{d}\theta\, u'^2/2 \tag{15.73}$$

$$H_{KdV} = \int_{S^1} \mathrm{d}\theta\, \left(u^3 + u'^2/2\right). \tag{15.74}$$

In Sect. 15.4.4.2 we will construct a metriplectic system based on the Gardner bracket (15.30) of Sect. 15.3.7. To avoid complications, we choose a simple example, yet one that displays general features of a large class of $1 + 1$ energy conserving dissipative system.

15.4.4.2 Metriplectic Systems Based on the Gardner Bracket

For simplicity we choose $\mathscr{P}_i = \mathscr{I}$ for all i, and as mentioned above, we suppose $A(\theta_1, \theta_2)$ is generated from the Gardner bracket (15.30), i.e.,

$$A(\theta_1, \theta_2) := \{U(\theta_1), U(\theta_2)\} = \int_{S^1} d\theta\, \delta(\theta - \theta_1) \frac{d}{d\theta} \delta(\theta - \theta_2) = \delta'(\theta_1 - \theta_2)\,, \tag{15.75}$$

where prime denotes differentiation with respect to argument and $\delta'(\theta_1 - \theta_2)$ is defined by

$$\begin{aligned}
\int_{S^1} \mathrm{d}\theta_1 \int_{S^1} \mathrm{d}\theta_2 \delta'(\theta_1 - \theta_2) f(\theta_1) g(\theta_2) &= -\int_{S^1} \mathrm{d}\theta_1 \int_{S^1} \mathrm{d}s\, \delta'(s) f(\theta_1) g(\theta_1 - s) \\
&= \int_{S^1} \mathrm{d}\theta_1 f(\theta_1) g'(\theta_1) \\
&= -\int_{S^1} \mathrm{d}\theta_1 \int_{S^1} \mathrm{d}\theta_2 \delta'(\theta_2 - \theta_1) f(\theta_1) g(\theta_2)
\end{aligned}$$

for any $f, g \in C^\infty(S^1)$, which shows that $\delta'(\theta_2 - \theta_1) = -\delta'(\theta_1 - \theta_2)$. With this choice for A we obtain

$$\mathscr{C}(\theta_1, \theta_2, \theta_3) = \delta'(\theta_1 - \theta_2) + \delta'(\theta_2 - \theta_3) + \delta'(\theta_3 - \theta_1)\,,$$

and Eq. (15.67) becomes

$$\begin{aligned}\{E, F, G\} = \int_{S^1} \mathrm{d}\theta_1 \int_{S^1} \mathrm{d}\theta_2 \int_{S^1} \mathrm{d}\theta_3 \, [\delta'(\theta_1 - \theta_2) + \delta'(\theta_2 - \theta_3) + \delta'(\theta_3 - \theta_1)] \cdot \\ \cdot E_u(\theta_1)\, F_u(\theta_2)\, G_u(\theta_3) \\ = \left(\int_{S^1} \mathrm{d}\bar\theta\, G_u(\bar\theta)\right) \int_{S^1} \mathrm{d}\theta\, F_u(\theta) E'_u(\theta) \\ + \left(\int_{S^1} \mathrm{d}\bar\theta\, E_u(\bar\theta)\right) \int_{S^1} \mathrm{d}\theta\, G_u(\theta) F'_u(\theta) \\ + \left(\int_{S^1} \mathrm{d}\bar\theta\, F_u(\bar\theta)\right) \int_{S^1} \mathrm{d}\theta\, E_u(\theta) G'_u(\theta). \end{aligned} \tag{15.76}$$

We shall construct a metriplectic system of the form

$$\dot F = \{H, F, G\} + \int_{S^1} \mathrm{d}\theta' \int_{S^1} \mathrm{d}\theta'' \, \{U(\theta'), S, F\}\, \mathscr{G}(\theta', \theta'')\, \{U(\theta''), S, G\}\,,$$

using the Gardner bracket (15.30).

Note that if $F = H_0$, the Casimir for the Gardner bracket (15.30), then, since $\delta H_0/\delta u = 1$, we obtain

$$\{F, H_0, G\} = \int_{S^1} \mathrm{d}\theta\, F_u G'_u \tag{15.77}$$

which is precisely the Gardner bracket. To see this, let us compute, for example, the integral in the third term of (15.76). Changing variables $s = \theta_3 - \theta_1$ we get

$$\begin{aligned}&\textstyle\int_{S^1} \mathrm{d}\theta_1 \int_{S^1} \mathrm{d}\theta_2 \int_{S^1} \mathrm{d}\theta_3\, \delta'(\theta_3 - \theta_1)\, E_u(\theta_1)\, G_u(\theta_3) \\ &= -\textstyle\int_{S^1} \mathrm{d}s \int_{S^1} \mathrm{d}\theta_3 \delta'(s)\, E_u(\theta_3 - s)\, G_u(\theta_3) = \int_{S^1} \mathrm{d}\theta_3\, E'_u(\theta_3) G_u(\theta_3).\end{aligned}$$

A similar computation shows that the first and second terms vanish.

In order to construct the symmetric bracket in (15.64), we need the following, computed using (15.76):

$$\begin{aligned}\{U(\theta), H, G\} = -\left(\int_{S^1} \mathrm{d}\bar\theta\, G_u(\bar\theta)\right) H'_u(\theta) + \int_{S^1} \mathrm{d}\bar\theta\, G_u(\bar\theta) H'_u(\bar\theta) \\ + \left(\int_{S^1} \mathrm{d}\bar\theta\, H_u(\bar\theta)\right) G'_u(\theta). \end{aligned} \tag{15.78}$$

Now with the counterpart of (15.78) for the functional F with $U(\theta')$, a choice for H, and a choice for $\mathscr{G}$, we can construct $(F, G)_H$. We make the following choices:

$$H_2(u) = \int_{S^1} \mathrm{d}\theta \, \frac{u^2}{2}, \qquad S(u) := H_0(u) = \int_{S^1} \mathrm{d}\theta \, u \tag{15.79}$$

$$\mathscr{G}(\theta', \theta'') = \delta(\theta' - \theta'') \,. \tag{15.80}$$

Now choose H_2 from (15.79) and insert it into (15.78) which gives

$$\{U(\theta), H_2, G\} = -\left(\int_{S^1} \mathrm{d}\bar{\theta} \, G_u(\bar{\theta})\right) u'(\theta) + \int_{S^1} \mathrm{d}\bar{\theta} \, G_u(\bar{\theta}) u'(\bar{\theta}) + S G_u'(\theta) \tag{15.81}$$

and to construct the symmetric bracket (15.64), we need

$$\{U(\theta''), H_2, S\} = -u'(\theta''). \tag{15.82}$$

Thus, the equations of motion are

$$\frac{d}{dt} F = \{F, H_0, H_2\} + (F, S)_{H_2}$$

where

$$(F, S)_{H_2} = \int_{S^1} \mathrm{d}\theta' \int_{S^1} \mathrm{d}\theta'' \, \{U(\theta'), H_2, F\} \, \mathscr{G}(\theta', \theta'') \, \{U(\theta''), H_2, S\} \,. \tag{15.83}$$

This yields

$$u_t - u_\theta = S \, u_{\theta\theta} + Q \qquad \text{with} \qquad Q := \int_{S^1} \mathrm{d}\theta' |u_{\theta'}|^2 \,. \tag{15.84}$$

Equation (15.84) has several interesting features. For fixed given constant S and Q, it is a linear equation composed of the heat equation with a source and with the inclusion of a linear advection term. One can proceed to solve this equation by the usual method of constructing a temporal Green's function out of the heat kernel and expanding in a Fourier series. After such a solution is constructed, one must enforce the fact that the global quantities S and Q are both time dependent and, importantly, dependent on the solution so constructed. Only after these constraints are enforced would one actually have a solution. Pursuing this construction, although interesting, is outside the scope of this paper and will be treated elsewhere.

We observe that the equation (15.84) is metriplectic. Indeed, by construction, we have a Poisson bracket (15.77) (the Gardner bracket) and a symmetric bracket (15.83). Since these were constructed out of triple brackets, property (iii) of

Definition in Sect. 15.4.1 holds. Positive semidefiniteness of the symmetric bracket follows from (15.81).

The nature of the dissipation of (15.84) is of particular interest in that it involves the global quantities S and Q. This is reminiscent of collision operators such as the Boltzmann, generalized Fokker–Planck, Landau, Lenard–Balescu, and other such operators (see, e.g., [53]). (Note also that non-local terms occur in the dissipative brackets discussed in, e.g., [37,39].) However, the usual dissipation in $1+1$ systems is local in nature (see Sect. 15.4.5) and dissipates energy, while (15.84) conserves energy like any good collision operator and does so in a system with a single spatial dimension. Thus the metriplectic construction of this section has pointed to a quite natural type of dynamical system that has dynamical versions of both the first and second laws of thermodynamics. The pathway for constructing other systems with nonlinear and dispersive Hamiltonian components, other kinds of dissipation, etc. is now cleared, and some will be considered in future publications.

15.4.4.3 Some Metriplectic Generalizations

It is evident that many generalizations are possible. We mention a few.

- Without destroying the symmetries or formal metriplectic bracket properties, we could allow one or both of the functions C and $\mathscr{G}$ to depend on the field variable u or even contain pseudodifferential operations. In fact, such ideas were used in similar brackets in [28] to facilitate numerical computation.
- It is clear how to generalize (15.64) to preserve more constraints, say $I_1, I_2, \ldots$, in addition to H. One first constructs the completely antisymmetric multilinear brackets $\{E, F, G, H, \ldots\}$ paralleling (15.63), and then, analogous to (15.64), constructs

$$(F,G)_{H,I_1,I_2,\ldots} = \int_{S^1} \mathrm{d}\theta' \int_{S^1} \mathrm{d}\theta'' \, \{U(\theta'), H, I_1, I_2, \ldots, F\} \mathscr{G}(\theta', \theta'') \cdot \\ \cdot \{U(\theta''), H, I_1, I_2, \ldots, G\} \, . \tag{15.85}$$

  The bracket $(F,G)_{H,I_1,I_2,\ldots}$ is guaranteed to be symmetric, conserve the invariants, and be positive semidefinite.
- It is of general interest to have metriplectic systems of the form

$$\dot{F} = \{H, F, G\} + \int_{S^1} \mathrm{d}\theta' \int_{S^1} \mathrm{d}\theta'' \, \{U(\theta'), S, F\} \mathscr{G}(\theta', \theta'') \{U(\theta''), S, G\}$$

  (such as our example of Sect. 15.4.4.2) for a suitably chosen function G; here H is the Hamiltonian and S is the entropy. Exploring the mathematics of when this is possible is an area to pursue.
- The construction here is easily extendable to higher spatial dimensions. For example, consider the following triple bracket given in [10]:

$$\{E, F, G\} = \int_{\mathscr{D}} d^6 z \; E_f \left[F_f, G_f\right] \,, \tag{15.86}$$

where $z = (q, p)$ is a canonical six-dimensional phase space variable, $f(z, t)$ is a phase space density, as in Vlasov theory, and the "inner" Poisson bracket is defined by

$$[f, g] = f_q \cdot g_p - f_p \cdot g_q \,. \tag{15.87}$$

We assume that the domain $\mathscr{D}$ with boundary conditions enables us to set all surface terms obtained by integrations by parts to zero, thereby assuring complete antisymmetry. Inserting the quadratic Casimir $C_2 := \int_{\mathscr{D}} d^6 z \, f^2/2$ into (15.86) gives

$$\{F, G\}_{VP} = \{C_2, F, G\} = \int_{\mathscr{D}} d^6 z \, f \, \left[F_f, G_f\right] \,,$$

the Lie–Poisson bracket for the Vlasov–Poisson system, as given in [50]. Thus, this bracket with the quadratic Casimir is formally akin to the construction given in Sect. 15.4.2.1 (although we note it reduces to a good bracket for any Casimir and in this way is like the case of $\mathfrak{so}(3)$ of Sect. 15.4.2.2). The triple bracket of (15.86) can be used in a generalization of the bracket of (15.64) to obtain a variety of energy conserving collision operators, with a wide choice of Casimirs as entropies.

15.4.5 Hybrid Dissipative Structures

Even if a system is not metriplectic, it is of interest to see if it can be obtained from an equation which consists of a Hamiltonian part and a gradient part with respect to a suitable Poisson bracket and metric, respectively.

For KdV-like equations, energy (the Hamiltonian) is generally not conserved when dissipation is added to the system. This is common for physical systems, but a more complete model would conserve energy while accounting for heat loss, i.e., entropy production. In the terminology of [55], models that lose energy, such as those treated here and those described by the double bracket formalism of Sect. 15.2.1, are *incomplete*, while those that do represent dynamical models of the laws of thermodynamics, such as metriplectic systems, are termed complete. Although incomplete systems do not conserve energy, they may conserve other invariants, and building this in represents an advantage of various bracket formulations. Thus, we construct incomplete hybrid Hamiltonian and dissipative dynamics by combining a Hamiltonian and a gradient vector field according to the prescription

$$u_t = \{u, H\} + (u, S), \tag{15.88}$$

where $u \mapsto \{u, H\}$ is a Hamiltonian vector field generated by H and $u \mapsto (u, S)$ is a gradient vector field generated by S (which could be H). Thus, $(\ ,\ )$ is, up to a sign, an inner product on the space of functions u.

Consider the following examples:

- With the usual KdV Hamiltonian of (15.31) and the Gardner bracket of (15.30) describing the Hamiltonian vector field, together with the choice

$$S(u) = H_1(u) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \mathrm{d}\theta\, (u_\theta)^2$$

we obtain for the gradients of Corollary 15.2

(i) $u_t = \{u, H\} - \nabla^1 H_1 = -u_{\theta\theta\theta} + 6uu_\theta - u$
(ii) $u_t = \{u, H\} - \nabla H_1 = -u_{\theta\theta\theta} + 6uu_\theta + u_{\theta\theta}$
(iii) $u_t = \{u, H\} - \nabla^2 H_1 = -u_{\theta\theta\theta} + 6uu_\theta - \mathscr{H}(u_\theta)$

which is the KdV equation of (15.29) with the inclusion of a new term that describes dissipation. Case (i) corresponds to simple linear damping, case (ii) to "viscous" diffusion, and case (iii) to the equation of [61] which adds a term to the KdV equation that describes Landau damping. For these systems the KdV invariant $\int_{-\pi}^{\pi} \mathrm{d}\theta\, u^2$ serves as a Lyapunov function.

- Choosing $H = S = H_1$, the Kähler Hamiltonian flow of (15.34) together with the dissipative flow generated by (15.21), yields

$$u_t = \{u, H_1\} - \nabla^2 H_1 = -u_\theta - \mathscr{H}(u_\theta)$$

which describes simple advection with Landau damping. This equation possesses the damped traveling wave solution.

- We note that we can derive the heat equation from a symmetric bracket of the form (15.64), again with $\mathscr{G}(\theta', \theta'') = \delta(\theta' - \theta'')$. Using this $\mathscr{G}$ and noting $\{U(\theta), H_0, F\} = G'_u(\theta)$, we obtain

$$(F, G)_{H_0} = \int_{S^1} d\theta\, F'_u G'_u\,. \tag{15.89}$$

Let us compute, for example, $\dot{F}(u) = (F, -H_2)_{H_0}$ (see (15.72)). Since $\delta H_2/\delta u = -u$, we obtain

$$\int_{S^1} \mathrm{d}\,\theta F_u\, \dot{u} = \frac{d}{dt}F(u) = (F, H_2)_{H_0} = -\int_{S^1} \mathrm{d}\theta\, F'_u u' = \int_{S^1} \mathrm{d}\theta\, F_u u''.$$

This yields

$$u_t = u_{xx}$$

which is the heat equation.

From these examples, it is clear how a variety of hybrid Hamiltonian and dissipative flows can be constructed from the machinery we have developed. For example, if we replace the KdV Hamiltonian by $H(u) = \int_{S^1} \mathrm{d}\theta \left(\frac{1}{2} u \mathscr{H}(u_\theta) + \frac{1}{3} u^3\right)$, we obtain the Benjamin-Ono equation with the various dissipative terms. Related ideas applied to fluid dynamics may be found in [30].

Acknowledgements AMB was partially supported by NSF grants DMS-0907949 and DMS-1207693. PJM was supported by U.S. Department of Energy contract DE-FG05-80ET-53088. TSR was partially supported by Swiss NSF grant 200021-140238, and by the government grant of the Russian Federation for support of research projects implemented by leading scientists, Lomonosov Moscow State University under the agreement No. 11.G34.31.0054. We thank Darryl Holm and the referees for useful comments.

References

1. Abraham, R., Marsden, J.E.: Foundations of Mechanics, 2nd edn, revised and enlarged. With the assistance of Tudor Ratiu and Richard Cushman. Benjamin/Cummings Publishing Co., Inc., Advanced Book Program, Reading (1978). Reprinted by Perseus Press, 1997 and AMS Chelsea, 2009
2. Adams, R.A., Fournier, J.J.F.: Sobolev Spaces, 2nd edn. Pure and Applied Mathematics, vol. 140. Elsevier/Academic, Amsterdam (2003)
3. Adler, M.: On a trace functional for formal pseudo differential operators and the symplectic structure of the Korteweg–de Vries type equations. Invent. Math. **50**(3), 219–248 (1979)
4. Arnold, V.I.: Matematicheskie Metody Klassicheskoi Mekhaniki. Isdat. "Nauka" Moscow (1974, in Russian). English translation Mathematical Methods of Classical Mechanics, Graduate Texts in Mathematics, vol. 60. Springer, New York (1978)
5. Arnold, V.I., Avez, A.: Problèmes ergodiques de la mécanique classique. Monographies Internationales de Mathématiques Modernes, No. 9. Gauthier-Villars, Éditeur, Paris (1967). English translation Egodic Problems of Classical Mechanics. W.A. Benjamin, Inc., New York (1968)
6. Atiyah, M.F.: Convexity and commuting Hamiltonians. Bull. Lond. Math. Soc. **14**, 305–315 (1982)
7. Atiyah, M.F., Pressley, A.N.: Convexity and loop groups. In: Arithmetic and Geometry, vol. II. Progress in Mathematics, vol. 36, pp. 33–63. Birkhäuser Boston, Boston (1983)
8. Beltiţă, D.: Integrability of analytic almost complex structures on Banach manifolds. Ann. Global Anal. Geom. **28**, 59–73 (2005)
9. Besse, A.L.: Einstein Manifolds. Reprint of the 1987 edition. Classics in Mathematics. Springer, Berlin (2008)
10. Bialynicki-Birula, I., Morrison, P.J.: Quantum mechanics as a generalization of Nambu dynamics to the Weyl-Wigner formalism. Phys. Lett. A **158**, 453–457 (1991)
11. Bloch, A.M.: Steepest descent, linear programming and Hamiltonian flows. Contemp. Math. AMS **114**, 77–88 (1990)
12. Bloch, A.M.: Nonholonomic Mechanics and Control. Springer, New York (2003)
13. Bloch, A.M., Brockett, R.W., Ratiu, T.S.: A new formulation of the generalized Toda Lattice equations and their fixed point analysis via the momentum map. Bull. Am. Math. Soc. **23**, 477–485 (1990)
14. Bloch, A.M., Brockett, R.W., Ratiu, T.S.: Completely integrable gradient flows. Commun. Math. Phys. **147**, 57–74 (1992)

15. Bloch, A.M., Flaschka, H., Ratiu, T.S.: A convexity theorem for isospectral manifolds of Jacobi matrices in a compact Lie algebra. Duke Math. J. **61**, 41–65 (1990)
16. Bloch, A.M., Iserles, A.: Aspects of generalized double bracket flows. In: Proc. Centre de Recherche Montreal, AMS, Group Theory and Numerical Analysis, vol. 39, pp. 65–76 (2005)
17. Bloch, A.M., Krishnaprasad, P.S., Marsden, J.E., Ratiu, T.S.: Dissipation induced instabilities. Ann. Inst. H. Poincaré Anal. Nonlineare **11**, 37–90 (1994)
18. Bloch, A.M., Krishnaprasad, P.S., Marsden, J.E., Ratiu, T.S.: The Euler–Poincaré equations and double bracket dissipation. Commun. Math. Phys. **175**, 1–42 (1996)
19. Bloch, A.M., Marsden, J.E.: Stabilization of rigid body dynamics by the energy-Casimir method. Syst. Control Lett. **14**, 341–346 (1990)
20. Bourbaki, N.: Lie Groups and Lie Algebras, Chapters 1–3. Springer, Berlin (1998). Translated from the 1971 French edition
21. Brockett, R.: Dynamical systems that sort lists, solve linear programming problems and diagonalize symmetric matrices. In: Proc. 1988 IEEE Conference on Decision and Control. Linear Algebra Appl., vol. 146, pp. 79–91 (1991)
22. Brockett, R.: The double bracket equation as the solution of a variational problem. In: Hamiltonian and Gradient Flows, Algorithms and Control. Fields Institute Communications, vol. 3, pp. 69–76. American Mathematical Society, Providence (1994)
23. Chow, B., Knopf, D.: The Ricci Flow: An Introduction. American Mathematical Society, Providence (2004)
24. Deift, P., Nanda, T., Tomei, C.: Differential equations for the symmetric eigenvalue problem. SIAM J. Numer. Anal. **20**, 1–22 (1983)
25. Deift, P., Li, L.C., Nanda, T., Tomei, C.: The Toda flow on a generic orbit is integrable. Mem. Am. Math. Soc. **100** (1992)
26. Fish, D.: Metriplectic Systems. Ph.D. Thesis. Portland State University (2005)
27. Flaschka, H.: The Toda lattice. Phys. Rev. B **9**, 1924–1925 (1974)
28. Flierl, G.R., Morrison, P.J.: Hamiltonian–Dirac simulated annealing: application to the calculation of vortex states. Physica D **240**, 212–232 (2011)
29. Gardner, C.S.: Korteweg–de Vries equation and generalizations. IV. The Korteweg–de Vries equation as a Hamiltonian systems. J. Math. Phys. **12**, 1548–1551 (1971)
30. Gay-Balmaz, F., Holm, D.D.: Parameterizing interaction of disparate scales: selective decay by Casimir dissipation in fluids. Preprint (2012). arXiv:1206.2607v1
31. Grmela, M., Öttinger, H.C.: Dynamics and thermodynamics of complex fluids. I. Development of a general formalism. Phys. Rev. E (3) **56**(6), 6620–6632 (1997)
32. Hamilton, R.S.: The inverse function theorem of Nash and Moser. Bull. Am. Math. Soc. (N.S.) **7**(1), 65–222 (1982)
33. Hamilton, R.S.: Three-manifolds with positive Ricci curvature. J. Differ. Geom. **17**, 255–306 (1982)
34. Holm, D.D., Marsden, J.E.: The rotor and the pendulum. In: Donato, P., et al. (eds.) Symplectic Geometry and Mathematical Physics, pp. 189–203. Birkhauser, Boston (1991)
35. Holm, D.D., Putkaradze, V.: Aggregation of finite size particles with variable mobility. Phys. Rev. Lett. **95**, 226106 (2005)
36. Holm, D.D., Putkaradze, V.: Formation and evolution of singularities in anisotropic geometric continua. Physica D **235**, 33–47 (2007)
37. Holm, D.D., Putkaradze, V., Tronci, C.: Geometric dissipation in kinetic equations. Comp. Rend. Acad. Sci. Sér. I **345**, 297–302 (2007)
38. Holm, D.D., Putkaradze, V., Tronci, C.: Geometric gradient-flow dynamics with singular solutions. Physica D **237**, 2952–2965 (2008)
39. Holm, D.D., Putkaradze, V., Tronci, C.: Double bracket dissipation in kinetic theory for particles with anisotropic interactions. Proc. R. Soc. A **466**, 2991–3012 (2010)
40. Kandrup, H.E.: The secular instability of axisymmetric collisionless star cluster. Astrophys. J. **380**, 511–514 (1991)
41. Kaufman, A.N., Morrison, P.J.: Algebraic structure of the plasma quasilinear equations. Phys. Lett. A **88**, 405–406 (1982)

42. Kaufman, A.N.: Dissipative Hamiltonian systems: a unifying principle. Phys. Lett. A **100**, 419–422 (1984)
43. King, F.W.: Hilbert Transforms, 2 vols. Encyclopedia of Mathematics and its Applications, vols. 124 and 125. Cambridge University Press, Cambridge (2009)
44. Kostant, B.: The solution to a generalized Toda lattice and representation theory. Adv. Math. **34**(3), 195–338 (1979)
45. Kriegl, A., Michor, P.W.: The Convenient Setting of Global Analysis. Mathematical Surveys and Monographs, vol. 53. American Mathematical Society, Providence (1997)
46. Liero, M., Mielke, A.: Gradient structures and geodesic convexity for reaction-diffusion systems. Preprint (2012)
47. Marsden, J.E., Ratiu, T.S.: Introduction to Mechanics and Symmetry. Texts in Applied Mathematics, vol. 17, 2nd edn. Springer, Berlin (1999)
48. Medina, A., Revoy, Ph.: Algèbres de Lie et produit scalaire invariant. Ann. Sci. Ec. Norm Super. 4^e série **18**, 553–561 (1985)
49. Mielke, A.: Formulation of thermoelastic dissipative material using GENERIC. Contin. Mech. Thermodyn. **23**, 233–256 (2011)
50. Morrison, P.J.: The Maxwell–Vlasov equations as a continuous Hamiltonian system. Phys. Lett. A **80**, 383–386 (1980)
51. Morrison, P.J.: Bracket formulation for irreversible classical fields. Phys. Lett. A **100**, 423–427 (1984)
52. Morrison, P.J.: Some observations regarding brackets and dissipation. Center for Pure and Applied Mathematics Report PAMD228. University of California, Berkeley (1984)
53. Morrison, P.J.: A paradigm for joined Hamiltonian and dissipative systems. Physica D **18**, 410–419 (1986)
54. Morrison, P.J.: Hamiltonian description of the ideal fluid. Rev. Mod. Phys. **70**, 467–521 (1998)
55. Morrison, P.J.: Thoughts on brackets and dissipation: old and new. J. Phys: Conf. Ser. **169**, 1–12 (2009)
56. Moser, J.: Finitely many mass points on the line under the influence of an exponential potential – an integrable system. In: Dynamical Systems, Theory and Applications (Rencontres, Battelle Res. Inst., Seattle, Wash., 1974). Lecture Notes in Physics, vol. 38, pp. 467–497. Springer, Berlin (1975)
57. Nambu, Y.: Generalized Hamiltonian dynamics. Phys. Rev. D **7**, 2405–2412 (1971)
58. Neeb, K.-H.: Infinite-dimensional groups and their representations. In: Lie Theory. Progress in Mathematics, vol. 228, pp. 213–328. Birkhäuser Boston, Boston (2004)
59. Neeb, K.-H.: Towards a Lie theory for infinite-dimensional groups. Jap. J. Math. 3rd Ser. **1**(2), 291–468 (2006)
60. Oettinger, H.C.: Beyond Equilibrium Thermodynamics. Wiley, New York (2006)
61. Ott, E., Sudan, R.N.: Nonlinear theory of ion acoustic waves with Landau damping. Phys. Fluids **12**, 2388–2394 (1969)
62. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**, 101–174 (2001)
63. Palais, R.S.: Foundations of Global Non-linear Analysis. Benjamin/Cummins Publishing Co., Reading (1968)
64. Pressley, A.N.: The energy flow on the loop space of a compact Lie group. J. Lond. Math. Soc. (2) **26**(3), 557–566 (1982)
65. Pressley, A., Segal, G.: Loop Groups. Oxford University Press, Oxford (1986)
66. Ratiu, T.: Involution theorems. In: Geometric Methods in Mathematical Physics (Proc. NSF-CBMS Conf., Univ. Lowell, Lowell, Mass., 1979), pp. 219–257. Lecture Notes in Math., vol. 775. Springer, Berlin (1980)
67. Segal, G.: Unitary representations of some infinite dimensional groups. Commun. Math. Phys. **80**, 301–342 (1981)
68. Souriau, J.-M.: Structure des Systèmes Dynamiques. Dunod, Paris (1970)
69. Symes, W.W.: Hamiltonian group actions and integrable systems. Physica D **1**, 339–374 (1980)

70. Symes, W.W.: Systems of Toda type, inverse spectral problems, and representation theory. Invent. Math. **59**(1), 13–51 (1980)
71. Symes, W.W.: The QR algorithm and scattering for the nonperiodic Toda lattice. Physica D **4**, 275–280 (1982)
72. Vallis, G.K., Carnevale, G., Young, W.R.: Extremal energy properties and construction of stable solutions of the Euler equations. J. Fluid Mech. **207**, 133–152 (1989)
73. Zaharov, V.E., Faddeev, L.D.: The Korteweg–de Vries equation is a fully integrable Hamiltonian system. Funkcional. Anal. Priložen. **5**(4), 18–27 (1971, in Russian)

Chapter 16
Boundary Tracking and Obstacle Avoidance Using Gyroscopic Control

Fumin Zhang, Eric W. Justh, and P.S. Krishnaprasad

Abstract For some time now, control-theoretic studies of collective motion of particles have shown the effectiveness of gyroscopic interactions in establishing stable spatiotemporal patterns in free space. This paper is concerned with strategies (respectively feedback laws) that prescribe (respectively execute) gyroscopic interaction of a single unit-speed particle with a fixed obstacle in space. The purpose of such interaction is to avoid collision with the obstacle and track associated (boundary) curves. Working in a planar setting, using the language of natural frames, we construct a steering law for the particle, based on sensing of curvature of the boundary, to track a *Bertrand mate* of the same. The curvature data is sensed at the closest point (image particle) on the boundary curve from the current location of the unit-speed particle. This construction extends to the three-dimensional case in which a unit-speed particle tracks a prescribed curve on a spherical obstacle. The tracking results exploit in an essential way, the method of reduction to shape space, and stability analysis of dynamics in shape space.

F. Zhang
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: fumin@gatech.edu

E.W. Justh
Naval Research Laboratory, Washington, DC 20375, USA

P.S. Krishnaprasad (✉)
Institute for Systems Research and Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA
e-mail: krishna@umd.edu

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_16,

16.1 Introduction

Tracking the boundary of a fixed structure (obstacle) is a biologically plausible mechanism for guidance and has applications in robotics. Here we are concerned specifically with animals (or vehicles) whose motion remains within certain upper and lower constraints on speed. For locomoting animals (running, flying, or swimming), there are certain (gait-dependent) ranges of speeds for which the *efficiency* of motion is optimized. It is then reasonable to hypothesize that the primary control input available to the animal to achieve a certain objective through motion is *steering* control. Precisely the same situation occurs in vehicles such as UAVs (unmanned aerial vehicles): speeds are kept within a restricted range to optimize distance covered or duration in the air, while desired trajectories are achieved through steering control.

Designing boundary-tracking steering laws in the planar setting is possible using simple systems of equations involving distances and angles. However, there are several major advantages to considering instead the differential geometric techniques described in this paper. First, using the language of curves and moving frames, the concepts can be generalized to three-dimensional motion in a straightforward and appealing way. Second, the filtering and interpretation of biological motion data (particularly three-dimensional trajectory data) are facilitated by the same techniques [5, 13]. Finally, the problem of boundary tracking can be placed within the same framework as successful (and biologically plausible) formulations of steering laws for formation control and pursuit [6–8, 14].

Here we consider a particle (representing an animal or vehicle) moving at unit speed in the presence of a fixed rigid body (i.e., an obstacle), and the prototype problem we consider is boundary tracking with collision avoidance. In the plane, the moving particle is subject to steering (i.e., curvature) control. From the point of view of mechanics, we are considering particle motion subject to *gyroscopic forces*; i.e., forces which change the direction of motion of the particle without altering its kinetic energy (and hence its speed). Recently, the idea of using gyroscopic forces for obstacle avoidance in robotics has gained renewed attention [3, 16, 20], and similar ideas have also started appearing in behavioral psychology [4]. In this paper, we draw strong analogies with recent work on two-vehicle interaction laws for formation control [6, 7] to develop a novel formulation of a gyroscopic boundary tracking and collision avoidance law. In the planar setting, we prove a global convergence result for regular obstacle boundaries enclosing convex regions. The key calculations are also shown to generalize to the three-dimensional setting, where we consider the interaction of a unit-speed particle with a prescribed curve on a spherical obstacle surface.

In robotics, two problems related to boundary tracking are curve tracking and trajectory tracking (where trajectory tracking has the additional requirement that points on the prescribed curve be reached at specific times). The distinctions between boundary tracking and curve/trajectory tracking are both mathematical (due to the noncollision requirement) and practical (due to the differences in sensing

and processing). Also, formulations of robotic curve/trajectory tracking problems often use a more detailed model of the vehicle than the unit-speed particle model we use. Nonetheless, our planar boundary-tracking approach can be viewed as building on earlier work in planar robotic curve/trajectory tracking (e.g., [11, 15]). For simultaneously tracking a spherical curve and avoiding collision with the sphere—the problem treated in Sect. 16.4—features of both boundary tracking and curve tracking are evident.

This paper is organized as follows. In Sect. 16.2 we introduce the planar model for boundary tracking in which a unit-speed particle (the "free particle") interacts with the closest point (assumed to be unique) on a convex obstacle boundary. In Sect. 16.3 the concept of a Bertrand mate to a curve is used to describe the boundary tracking *strategy*. Next, the *steering law* employed to achieve planar boundary tracking is presented and interpreted, along with the Lyapunov function used to prove convergence. In the convergence proof, a particular choice of variables is used, foreshadowing the convergence proof for the spherical curve tracking law of Sect. 16.4. A key step in both the planar and three-dimensional setting is selecting the boundary tracking law so as to make the closed-loop "shape dynamics" autonomous, which enables us to use the usual LaSalle invariance principle for autonomous systems to prove convergence. The spherical curve tracking law analyzed in Sect. 16.4 can be viewed as a restricted three-body problem in which the free particle responds to the closest point (guide point) on the spherical curve (guiding curve), as well as the closest point (shadow point) on the sphere (obstacle). The free particle thus simultaneously tracks the guiding curve and avoids collision. Furthermore, the form of the steering law suggests a biologically plausible mechanism for achieving the simultaneous objectives of curve tracking and obstacle avoidance in three dimensions.

16.2 Planar Boundary Tracking

It is useful to distinguish between *strategies* and the *feedback laws* used to execute strategies. In particular, when analyzing biological data, it is often desirable to test hypotheses about what strategy a moving animal is using without having to perform the modeling and data extraction needed to test hypotheses about specific feedback laws. The mathematical distinction between strategies and feedback laws is made precise in Sect. 16.3 (for the planar setting) and Sect. 16.4 (for the spherical obstacle setting).

16.2.1 Models

In the planar setting, consider a particle (which we will refer to as the "free particle") moving at unit speed (and subject to steering control) in the presence of a single

obstacle (i.e., the region enclosed by a simple closed curve) whose boundary is twice continuously differentiable. Suppose that at each instant of time, the point on the obstacle boundary which is closest to (i.e., the minimum Euclidean distance from) the moving particle is unique. This closest point on the obstacle boundary, which we call the "shadow point," moves along the boundary curve. (We assume uniqueness of the closest point in order to streamline the discussion and bring out the key ideas. Of course, in dealing with real-world obstacles, nonuniqueness of the closest point is an important issue.)

Let $\mathbf{r}_1$ denote the position of the shadow point, let $\mathbf{x}_1$ denote the unit tangent vector to the boundary curve at the shadow point, and let $\mathbf{y}_1$ denote the unit normal vector. We use the convention that a unit normal vector completes a right-handed orthonormal frame with the corresponding unit tangent vector. In terms of the arc-length parameterization, the boundary curve can be described by

$$\begin{aligned} \mathbf{r}_1' &= \mathbf{x}_1, \\ \mathbf{x}_1' &= \mathbf{y}_1 \kappa, \\ \mathbf{y}_1' &= -\mathbf{x}_1 \kappa, \end{aligned} \tag{16.1}$$

where the prime denotes differentiation with respect to arc-length parameter s; $\mathbf{r}_1$, $\mathbf{x}_1$, and $\mathbf{y}_1$ are $\mathbb{R}^2$-valued functions of s; and κ is the plane curvature function for the boundary curve. Using the chain rule, we can express the time-evolution of the shadow point as

$$\begin{aligned} \dot{\mathbf{r}}_1 &= \nu \mathbf{x}_1, \\ \dot{\mathbf{x}}_1 &= \nu \mathbf{y}_1 \kappa, \\ \dot{\mathbf{y}}_1 &= -\nu \mathbf{x}_1 \kappa, \end{aligned} \tag{16.2}$$

where $\nu = ds/dt$. Because the shadow point depends on the motion of the free particle, ν depends on both the boundary curve and on the trajectory of the free particle.

Letting $\mathbf{r}_2$ denote the position of the free particle, $\mathbf{x}_2$ the unit tangent vector to its trajectory, $\mathbf{y}_2$ the unit normal vector, and u the steering control for the free particle, we have the following system of equations for the "formation" consisting of the free particle and the shadow point:

$$\begin{aligned} \dot{\mathbf{r}}_1 &= \nu \mathbf{x}_1, & \dot{\mathbf{r}}_2 &= \mathbf{x}_2, \\ \dot{\mathbf{x}}_1 &= \nu \mathbf{y}_1 \kappa, & \dot{\mathbf{x}}_2 &= \mathbf{y}_2 u, \\ \dot{\mathbf{y}}_1 &= -\nu \mathbf{x}_1 \kappa, & \dot{\mathbf{y}}_2 &= -\mathbf{x}_2 u. \end{aligned} \tag{16.3}$$

Note that $\mathbf{x}_2$ describes the *heading* of the free particle, i.e., its (instantaneous) direction of motion. In (16.3), κ may be considered given (in practice, κ may be derived from sensor measurements); ν is a deterministic function of $(\mathbf{r}_1, \mathbf{x}_1, \mathbf{y}_1)$,

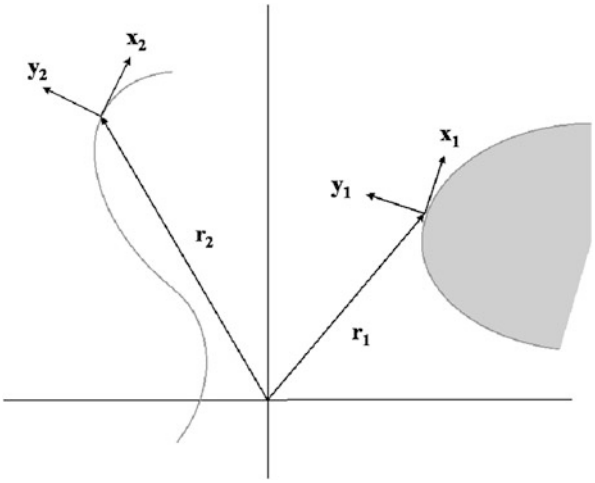

Fig. 16.1 Positions and frames for the trajectory of the moving particle ($\mathbf{r}_2$, $\mathbf{x}_2$, $\mathbf{y}_2$) and for the closest point on the boundary curve ($\mathbf{r}_1$, $\mathbf{x}_1$, $\mathbf{y}_1$) (from [20])

$(\mathbf{r}_2, \mathbf{x}_2, \mathbf{y}_2)$, and κ; and u is the control input we apply to avoid colliding with the obstacle and to achieve boundary tracking (see Fig. 16.1). A control law is a feedback function u of $(\mathbf{r}_1, \mathbf{x}_1, \mathbf{y}_1)$, $(\mathbf{r}_2, \mathbf{x}_2, \mathbf{y}_2)$, and κ. We seek to prove analytically that collision avoidance and boundary tracking are achieved for particular choices of the control law.

16.2.2 Boundary-Curve Frame Convention

We define $\mathbf{r} = \mathbf{r}_2 - \mathbf{r}_1$ to be the vector from the shadow point on the boundary curve to the free particle. We assume that initially

$$|\mathbf{r}| > 0, \tag{16.4}$$

and we will prove that our boundary-tracking steering law then guarantees (16.4) for all future time. Using $|\mathbf{r}| = (\mathbf{r} \cdot \mathbf{r})^{1/2}$, we compute

$$\frac{d}{dt}|\mathbf{r}| = \frac{\mathbf{r} \cdot \dot{\mathbf{r}}}{|\mathbf{r}|} = \frac{\mathbf{r}}{|\mathbf{r}|} \cdot (\mathbf{x}_2 - \nu \mathbf{x}_1) = \left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{x}_2\right) - \nu \left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{x}_1\right). \tag{16.5}$$

The first-order necessary conditions for the shadow point to be an extremum of the Euclidean distance from the free particle to the curve are

$$\mathbf{r} \cdot \mathbf{x}_1 \equiv 0, \tag{16.6}$$

and

$$\mathbf{y}_1 \equiv \pm \frac{\mathbf{r}}{|\mathbf{r}|}, \tag{16.7}$$

where the correct choice of sign depends on whether the boundary curve is to the right or left of the free particle, and on what convention is chosen for the positive direction along the boundary curve.

In order to fix a convention for the positive direction of the boundary curve, we assume that initially

$$\mathbf{r} \cdot \mathbf{y}_2 \neq 0, \tag{16.8}$$

i.e., that the free particle is not initially heading directly toward or directly away from the shadow point on the boundary curve. We will prove that under this assumption (and other appropriate hypotheses), our boundary-tracking steering law derived below guarantees (16.8) for all future time. We may thus choose the positive direction of the boundary curve such that

$$\mathbf{x}_1 \cdot \mathbf{x}_2 > 0. \tag{16.9}$$

With this convention, the choice of sign in (16.7) is determined by whether the boundary curve is to the left or right of the free particle. Furthermore, because of our convention that $(\mathbf{x}_1, \mathbf{y}_1)$ forms a right-handed frame, (16.9) implies that the sign of κ depends not just on the boundary curve, but on the position and heading of the free particle, as well. (Although curvature is a property of an oriented boundary curve alone, we are allowing the *orientation* of the boundary curve to depend on the initial relative position and heading of the free particle.)

We can derive an expression for ν by differentiating (16.6) with respect to time, to obtain

$$\begin{aligned} \frac{d}{dt}(\mathbf{r} \cdot \mathbf{x}_1) &= \dot{\mathbf{r}} \cdot \mathbf{x}_1 + \mathbf{r} \cdot \dot{\mathbf{x}}_1 \\ &= (\mathbf{x}_2 - \nu \mathbf{x}_1) \cdot \mathbf{x}_1 + (\mathbf{r} \cdot \mathbf{y}_1)\, \nu \kappa \\ &= \mathbf{x}_1 \cdot \mathbf{x}_2 - \nu + (\mathbf{r} \cdot \mathbf{y}_1)\, \nu \kappa \\ &= 0. \end{aligned} \tag{16.10}$$

We then have

$$\nu = \frac{\mathbf{x}_1 \cdot \mathbf{x}_2}{1 - (\mathbf{r} \cdot \mathbf{y}_1)\, \kappa}, \tag{16.11}$$

where we assume that $1 - (\mathbf{r} \cdot \mathbf{y}_1)\, \kappa > 0$. If $(\mathbf{r} \cdot \mathbf{y}_1)\, \kappa < 0$, we say that the boundary "curves away from" the free particle, and if $(\mathbf{r} \cdot \mathbf{y}_1)\, \kappa > 0$, we say that the

boundary "curves toward" the free particle. (The intermediate case corresponds to $\kappa = 0$.) We note that there is a singularity in the expression for ν when $|\mathbf{r}| = 1/|\kappa|$ and the boundary curves inward toward the free particle.

16.3 Planar Bertrand Mate Strategy

In [17] the concept of *strategy* in the context of pursuit is defined as a particular submanifold of the configuration space of a pair of particles. Specific steering laws which leave this submanifold invariant (or approximately invariant), and which drive the system toward this submanifold from general initial conditions, are also discussed. Taking this viewpoint, we can identify the boundary-tracking strategy with the submanifold

$$M_{2D}(r_0) = \left\{ \left(\begin{bmatrix} \mathbf{x}_1 & \mathbf{y}_1 & \mathbf{r}_1 \\ 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} \mathbf{x}_2 & \mathbf{y}_2 & \mathbf{r}_2 \\ 0 & 0 & 1 \end{bmatrix} \right) \in SE(2) \times SE(2) \middle| \right.$$
$$\left. |\mathbf{r}| = |\mathbf{r}_2 - \mathbf{r}_1| = r_0, \;\; \mathbf{x}_1 = \mathbf{x}_2, \; \mathbf{y}_1 = \mathbf{y}_2, \; \mathbf{r} \cdot \mathbf{x}_1 = 0 \right\}, \quad (16.12)$$

i.e., the free particle moves along a *Bertrand mate* of the boundary curve [1,12], with a separation r_0. The control law presented below leaves $M_{2D}(r_0)$ invariant, and furthermore, drives the system to $M_{2D}(r_0)$ for a large set of initial conditions. (For a regular curve $\mathbf{r}_1$ parameterized by s as in Eq. (16.1), a Bertrand mate, when one exists, is a regular curve $\mathbf{r}_2$ satisfying

$$\mathbf{r}_2 = \mathbf{r}_1 + \lambda \mathbf{y}_1, \;\; \mathbf{x}_2 = \mathbf{x}_1, \;\; \mathbf{y}_2 = \mathbf{y}_1, \quad (16.13)$$

where $\lambda > 0$ is a constant. The two curves $\mathbf{r}_1$ and $\mathbf{r}_2$ are then referred to as a Bertrand pair. The notion of Bertrand pairs generalizes the concept of parallel straight lines to "parallel" curves.)

16.3.1 Lyapunov Function and Steering Law

Consider the Lyapunov function candidate (from [20])

$$V_{2D} = -\ln(\mathbf{x}_1 \cdot \mathbf{x}_2) + h(\mathbf{r} \cdot \mathbf{y}_1), \quad (16.14)$$

where $h(\cdot)$ is continuously differentiable and satisfies certain hypotheses, to be specified below. Since $\mathbf{y}_1 = \pm \mathbf{r}/|\mathbf{r}|$, we have $h(\mathbf{r} \cdot \mathbf{y}_1) = h(\pm|\mathbf{r}|)$. There are thus two cases that need to be treated separately. We will focus on the $\mathbf{y}_1 = \mathbf{r}/|\mathbf{r}|$ case, and note that analogous results hold for $\mathbf{y}_1 = -\mathbf{r}/|\mathbf{r}|$.

The term in (16.14) involving $\mathbf{x}_1 \cdot \mathbf{x}_2$ penalizes heading misalignment between the free particle and the shadow point moving on the boundary curve. The term $h(|\mathbf{r}|)$ in (16.14) involves the separation between the free particle and the shadow point. We assume that

$$\lim_{\rho \to 0} h(\rho) = \infty, \tag{16.15}$$

so that V blows up as the free particle approaches collision with the boundary curve.

For $\mathbf{y}_1 = \mathbf{r}/|\mathbf{r}|$, following [20], we take as our steering law

$$u = \mu(\mathbf{x}_1 \cdot \mathbf{y}_2) - f(|\mathbf{r}|)\left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{y}_2\right) + \nu\kappa, \tag{16.16}$$

where ν is given by (16.11) and $f(\rho) = dh/d\rho$ (Fig. 16.2). One possible choice for $f(\cdot)$ is

$$f(|\mathbf{r}|) = \alpha\left[1 - \left(\frac{r_0}{|\mathbf{r}|}\right)^2\right], \tag{16.17}$$

where α and r_0 are positive constants, and r_0 represents the desired separation between the free particle and the boundary curve during boundary tracking. The term in (16.16) involving $f(\cdot)$ serves to steer the free particle toward or away from the boundary curve to achieve the desired separation.

Then differentiating V_{2D} along trajectories of (16.3) for steering law u given by (16.16) yields

$$\begin{aligned}
\dot{V}_{2D} &= -\frac{(\dot{\mathbf{x}}_1 \cdot \mathbf{x}_2 + \mathbf{x}_1 \cdot \dot{\mathbf{x}}_2)}{\mathbf{x}_1 \cdot \mathbf{x}_2} + f(|\mathbf{r}|)\frac{d}{dt}|\mathbf{r}| \\
&= -\frac{[\dot{\mathbf{x}}_1 \cdot \mathbf{x}_2 + (\mathbf{x}_1 \cdot \mathbf{y}_2)u]}{\mathbf{x}_1 \cdot \mathbf{x}_2} + f(|\mathbf{r}|)\left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{x}_2\right) \\
&= -\frac{\mathbf{x}_1 \cdot \mathbf{y}_2}{\mathbf{x}_1 \cdot \mathbf{x}_2}\left[u - \nu\kappa + f(|\mathbf{r}|)\left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{y}_2\right)\right] \\
&= -\frac{\mu(\mathbf{x}_1 \cdot \mathbf{y}_2)^2}{\mathbf{x}_1 \cdot \mathbf{x}_2},
\end{aligned} \tag{16.18}$$

where we have used $\mathbf{x}_2 \cdot \mathbf{y}_1 = -\mathbf{x}_1 \cdot \mathbf{y}_2$, and

$$0 = \frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{x}_1 = \left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{x}_2\right)(\mathbf{x}_1 \cdot \mathbf{x}_2) + \left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{y}_2\right)(\mathbf{x}_1 \cdot \mathbf{y}_2). \tag{16.19}$$

Thus, for $\mathbf{y}_1 = \mathbf{r}/|\mathbf{r}|$, we have $\dot{V}_{2D} \leq 0$ and $\dot{V}_{2D} = 0$ if and only if $\left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{x}_2\right) = 0$.

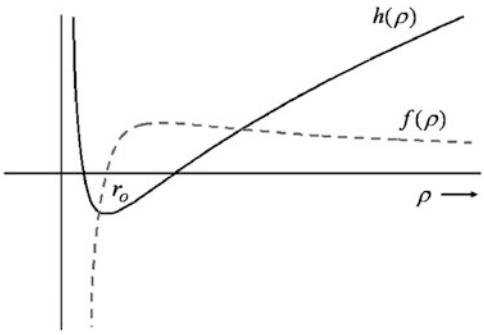

Fig. 16.2 Examples of functions $h(\cdot)$ (*solid*) and $f(\cdot)$ (*dashed*) meeting the requirements $(A1)$, $(A2)$, and $(A3)$ in **Proposition 16.1** (from [20])

Expression (16.16) is simple to interpret. The term $\mu(\mathbf{x}_1 \cdot \mathbf{y}_2)$ serves to align the heading of the free particle with the tangent vector to the shadow point trajectory on the boundary curve. The term involving $f(\cdot)$ serves to steer the free particle toward or away from the boundary curve to achieve the desired separation. Finally, the term involving κ enables the free particle to respond to the nonzero curvature of the boundary curve.

This way of thinking about obstacle avoidance and boundary tracking leads to the idea that the role of the sensors on a moving vehicle (or animal), represented by the free particle, is to identify the closest point on the boundary curve, and to estimate its relative position and the curvature of the boundary at that point (the shadow point). Once the shadow point is initially identified, it can in principle be tracked as the interaction of the free particle and shadow point evolves. However, there is a requirement to periodically scan in other directions, in case the local minimizer being tracked ceases to be a global minimizer at some point in time.

16.3.2 Shape Variables

Observe that under control (16.16), we have (for $\mathbf{y}_1 = \mathbf{r}/|\mathbf{r}|$)

$$\frac{d}{dt}(\mathbf{r} \cdot \mathbf{y}_1) = \dot{\mathbf{r}} \cdot \mathbf{y}_1 + \mathbf{r} \cdot \dot{\mathbf{y}}_1 = (\mathbf{x}_2 - \nu \mathbf{x}_1) \cdot \mathbf{y}_1 - \nu\kappa(\mathbf{r} \cdot \mathbf{x}_1) = \mathbf{x}_2 \cdot \mathbf{y}_1, \quad (16.20)$$

$$\begin{aligned} \frac{d}{dt}(\mathbf{x}_2 \cdot \mathbf{x}_1) &= \dot{\mathbf{x}}_2 \cdot \mathbf{x}_1 + \mathbf{x}_2 \cdot \dot{\mathbf{x}}_1 = (\mathbf{y}_2 \cdot \mathbf{x}_1)u + (\mathbf{x}_2 \cdot \mathbf{y}_1)\nu\kappa \\ &= \mu(\mathbf{x}_1 \cdot \mathbf{y}_2)^2 - (\mathbf{x}_1 \cdot \mathbf{y}_2)\left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{y}_2\right) f(|\mathbf{r}|), \end{aligned} \quad (16.21)$$

$$\begin{aligned} \frac{d}{dt}(\mathbf{x}_2 \cdot \mathbf{y}_1) &= \dot{\mathbf{x}}_2 \cdot \mathbf{y}_1 + \mathbf{x}_2 \cdot \dot{\mathbf{y}}_1 = (\mathbf{y}_2 \cdot \mathbf{y}_1)u - (\mathbf{x}_2 \cdot \mathbf{x}_1)\nu\kappa \\ &= \mu(\mathbf{x}_2 \cdot \mathbf{x}_1)(\mathbf{x}_1 \cdot \mathbf{y}_2) - (\mathbf{x}_2 \cdot \mathbf{x}_1)\left(\frac{\mathbf{r}}{|\mathbf{r}|} \cdot \mathbf{y}_2\right) f(|\mathbf{r}|). \end{aligned} \quad (16.22)$$

Also, note that

$$\mathbf{r}\cdot\mathbf{y}_2 = (\mathbf{r}\cdot\mathbf{x}_1)(\mathbf{x}_1\cdot\mathbf{y}_2)+(\mathbf{r}\cdot\mathbf{y}_1)(\mathbf{y}_1\cdot\mathbf{y}_2) = (\mathbf{x}_2\cdot\mathbf{x}_1)(\mathbf{r}\cdot\mathbf{y}_1) = (\mathbf{x}_2\cdot\mathbf{x}_1)|\mathbf{r}|. \tag{16.23}$$

Thus, if we define the shape variables

$$\begin{aligned} 0 &= \mathbf{r}\cdot\mathbf{x}_1, \\ z_1 &= \mathbf{r}\cdot\mathbf{y}_1 = |\mathbf{r}|, \\ z_2 &= \mathbf{x}_2\cdot\mathbf{x}_1, \\ z_3 &= \mathbf{x}_2\cdot\mathbf{y}_1, \end{aligned} \tag{16.24}$$

system (16.20)–(16.22) becomes

$$\begin{aligned} \dot{z}_1 &= z_3, \\ \dot{z}_2 &= \mu z_3^2 + z_3 z_2 f(z_1), \\ \dot{z}_3 &= -\mu z_2 z_3 - z_2^2 f(z_1), \end{aligned} \tag{16.25}$$

which is an autonomous system. Observe that if initially $z_2^2 + z_3^2 = 1$ in (16.25), then $z_2^2 + z_3^2 = 1$ for all time. Furthermore,

$$V_{2D} = -\ln(z_2) + h(z_1), \tag{16.26}$$

and

$$\dot{V}_{2D} = -\frac{\mu z_3^2}{z_2} = -\frac{\mu}{z_2}\left(1 - z_2^2\right). \tag{16.27}$$

Although (16.24) is not the only possible way to define shape variables, it is the approach that generalizes most naturally to three dimensions (see Sect. 16.4 below).

16.3.3 Convergence Result

Remark 16.1. Equilibria of (16.25) can be viewed as "relative equilibria" for the non-autonomous system (16.3) with control law (16.16).

Remark 16.2. For the special case of a circular obstacle, κ is constant, and (16.3) with control law (16.16) is an autonomous system. This special case is treated in [20].

Remark 16.3. It has been recently shown in [10] that the control law (16.16) achieves input-to-state stability (ISS) for the relative equilibria of the nonlinear

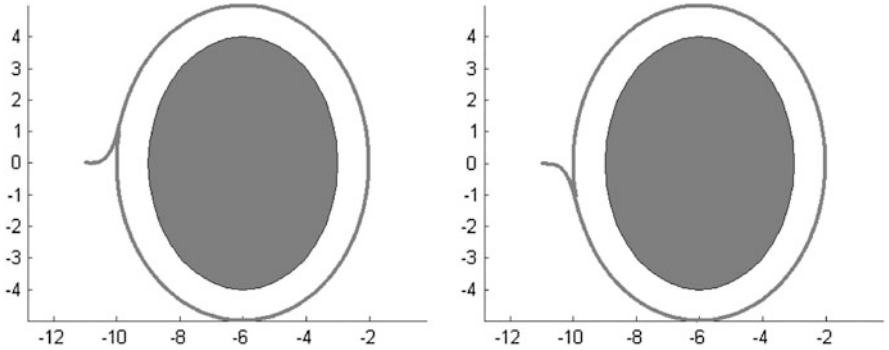

Fig. 16.3 Simulation of a free particle tracking an elliptical boundary curve in the plane using control law (16.16). *Left plot*: the free particle starts at coordinates $(-11,0)$ with an initial heading of $+10^\circ$ measured counterclockwise from the x-axis. The free particle follows the boundary curve to its right, circling the obstacle clockwise. *Right plot*: the free particle starts at the same initial position, but with an initial heading of -10°. Then the free particle follows the boundary curve to its left, circling the obstacle counterclockwise. In both cases, the desired separation between the free particle and the boundary curve is set to one distance unit (from [20])

system dynamics (16.3). This ISS property justifies that control law (16.16) has a certain robustness to disturbances, which agrees with observations in robotic applications [19].

Proposition 16.1. *Consider the planar dynamics (16.3) with control law (16.16) for a unit-speed free particle and the closest point to it on a boundary curve. Assume that the boundary curve encloses a convex region in the plane, the free particle initially lies outside this convex region, and initially (16.9) is satisfied with* $\mathbf{y}_1 = \mathbf{r}/|\mathbf{r}|$. *Let* (z_1, z_2, z_3) *be defined by (16.24). Assume also that*

(A1) $dh/d\rho = f(\rho)$, *where* $f(\rho)$ *is a Lipschitz continuous function on* $(0, \infty)$, *so that* $h(\rho)$ *is continuously differentiable on* $(0, \infty)$;
(A2) $f(\rho) = 0$ *at a finite number of isolated points;*
(A3) $\lim_{\rho \to 0} h(\rho) = \infty$, $\lim_{\rho \to \infty} h(\rho) = \infty$, *and* $\exists \tilde{\rho}$ *such that* $h(\tilde{\rho}) = 0$;
(A4) $\mu > 0$ *is a Lipschitz continuous function of* (z_1, z_2, z_3).

Then for arbitrary initial conditions satisfying $0 < z_2 \leq 1$ *and* $z_1 \neq 0$, *system (16.25) converges to the set of its equilibria for which* $z_2 = 1$ *and* $z_3 = 0$.

Proof. Note that $V_{2D} \to \infty$ as $z_1 \to \infty$, $z_1 \to 0$, or $z_2 \to 0$. Also, $z_2^2 + z_3^2 \equiv 1$. Therefore, the sublevel sets of V_{2D} are compact. Hence, $\dot{V}_{2D}$ always exists, satisfies $\dot{V}_{2D} \leq 0$, and $\dot{V}_{2D} = 0$ if and only if $z_2 = 1$.

Then by LaSalle's Invariance Principle (for autonomous systems) [9], we conclude that system (16.25) converges to the largest invariant set with $z_2 = 1$.

But $z_2 = 1$ implies $z_3 = 0$, and (16.25) reduces to

$$\dot{z}_1 = 0, \quad \dot{z}_2 = 0, \quad \dot{z}_3 = -f(z_1). \tag{16.28}$$

Thus, (16.25) converges to the set of its equilibria for which $z_2 = 1$, $z_3 = 0$, and $f(z_1) = 0$. By assumption (A2), these equilibria are isolated, and therefore we can conclude that (z_1, z_2, z_3) converge to an equilibrium. This equilibrium in the shape variables (z_1, z_2, z_3) corresponds to the free particle following a trajectory which is a Bertrand mate to the boundary curve, and at a distance from the boundary curve given by one of the zeros of the function $f(\cdot)$. □

Figure 16.3 shows a simulation of a free particle tracking an elliptical boundary curve in the plane using control law (16.16).

16.4 Curve Tracking with Obstacle Avoidance in Three Dimensions

One of the major strengths of our planar boundary tracking law is that it can be generalized to the three-dimensional setting. Suppose there is a free particle moving in three-dimensional space, and there is a fixed, smooth, two-dimensional obstacle surface (e.g., a sphere). As the free particle moves at unit speed, the closest point to it on the obstacle surface (which we assume is unique) also moves along a three-dimensional trajectory, which is constrained to lie on the obstacle surface. We can then, as in the planar problem, consider the coupled dynamics of the free particle and the closest point (shadow point). This two-particle problem with a spherical obstacle is discussed in [20]).

Here we consider a three-particle problem: a free particle moving at unit speed in three-dimensional space interacts with the closest point on a prescribed *spherical curve* (i.e., a curve constrained to lie on the surface of a sphere of radius $R > 0$), which we will refer to as the "guiding curve." But the free particle also interacts with the closest point on the sphere, which may or may not coincide with the closest point on the guiding curve. We will refer to the closest point on the sphere as the "shadow point" and the closest point on the guiding curve as the "guide point." The guiding curve is meant to guide the free particle, while the sphere itself is an obstacle that the free particle must avoid colliding with. As the free particle approaches the desired motion state (a non-autonomous relative equilibrium corresponding to the free particle tracing the projection of the guiding curve onto a concentric sphere with prescribed radius $R + r_0$, $r_0 > 0$), the shadow point approaches the guide point. This problem is most effectively formulated using the natural Frenet framing of three-dimensional curves, as described next.

16.4.1 Curves and Moving Frames

A single particle moving in three-dimensional space traces out a trajectory $\boldsymbol{\gamma} : [0, \infty) \to \mathbb{R}^3$, which we assume to be at least twice continuously differentiable, satisfying[1] $|\boldsymbol{\gamma}'(s)| = 1$, $\forall s$; i.e., s is the arc-length parameter of the curve (and the prime denotes differentiation with respect to s). The direction of motion of the particle at s is the unit tangent vector to the trajectory, $\mathbf{T}(s) = \boldsymbol{\gamma}'(s)$. The gyroscopic force vector always lies in the plane perpendicular to $\mathbf{T}$, so to describe the effects of this force, we are compelled to introduce orthonormal unit vectors which span this *normal plane*. Taken together with $\mathbf{T}$, these unit vectors constitute a *framing* of the curve $\boldsymbol{\gamma}$ representing the particle trajectory.

One choice for framing the curve is the natural Frenet frame, which is also referred to as the Fermi–Walker frame or Relatively Parallel Adapted Frame (RPAF) [1]:

$$\begin{aligned} \boldsymbol{\gamma}'(s) &= \mathbf{T}(s), \\ \mathbf{T}'(s) &= k_1(s)\mathbf{M}_1(s) + k_2(s)\mathbf{M}_2(s), \\ \mathbf{M}_1'(s) &= -k_1(s)\mathbf{T}(s), \\ \mathbf{M}_2'(s) &= -k_2(s)\mathbf{T}(s). \end{aligned} \tag{16.29}$$

In (16.29), $\mathbf{M}_1(s)$ and $\mathbf{M}_2(s)$ are unit normal vectors which (along with $\mathbf{T}(s)$) complete a right-handed orthonormal frame. However, there is freedom in the choice of initial conditions $\mathbf{M}_1(0)$ and $\mathbf{M}_2(0)$; once these are specified, the corresponding natural Frenet frame for a twice-continuously differentiable curve $\boldsymbol{\gamma}$ is unique. The benefits of using natural Frenet frames for describing interacting systems of particles in three dimensions are discussed in [7, 14]. A deeper discussion of the theory underlying natural Frenet frames can be found in [2].

16.4.2 Spherical Curves and Natural Frames

Spherical curves, i.e., curves constrained to lie on the surface of a sphere in $\mathbb{R}^3$, are readily described using natural Frenet frames [1]. If the curve $\boldsymbol{\gamma}(s)$ lies on a sphere of radius $R > 0$ centered at $\mathbf{p}$, then for all s,

$$(\boldsymbol{\gamma} - \mathbf{p}) \cdot (\boldsymbol{\gamma} - \mathbf{p}) = R^2. \tag{16.30}$$

Differentiating (16.30) gives

$$(\boldsymbol{\gamma} - \mathbf{p}) \cdot \boldsymbol{\gamma}' = (\boldsymbol{\gamma} - \mathbf{p}) \cdot \mathbf{T} = 0, \tag{16.31}$$

[1] Here and in the rest of the paper, $|\mathbf{w}| = (\mathbf{w} \cdot \mathbf{w})^{1/2}$ for any $\mathbf{w} \in \mathbb{R}^3$.

and differentiating a second time yields

$$\mathbf{T} \cdot \mathbf{T} + (\boldsymbol{\gamma} - \mathbf{p}) \cdot \mathbf{T}' = 0, \tag{16.32}$$

or

$$(\boldsymbol{\gamma} - \mathbf{p}) \cdot \mathbf{T}' = -1. \tag{16.33}$$

From $\mathbf{M}_1' = -k_1 \mathbf{T}$, we obtain

$$\mathbf{M}_1' \cdot (\boldsymbol{\gamma} - \mathbf{p}) = -k_1 \mathbf{T} \cdot (\boldsymbol{\gamma} - \mathbf{p}) = 0, \tag{16.34}$$

and similarly, $\mathbf{M}_2' \cdot (\boldsymbol{\gamma} - \mathbf{p}) = 0$.

We then observe that

$$(\mathbf{M}_1 \cdot (\boldsymbol{\gamma} - \mathbf{p}))' = \mathbf{M}_1' \cdot (\boldsymbol{\gamma} - \mathbf{p}) + \mathbf{M}_1 \cdot \mathbf{T} = 0, \tag{16.35}$$

and similarly, $(\mathbf{M}_2 \cdot (\boldsymbol{\gamma} - \mathbf{p}))' = 0$, from which we conclude that

$$\mathbf{M}_1 \cdot (\boldsymbol{\gamma} - \mathbf{p}) = \text{constant}, \quad \mathbf{M}_2 \cdot (\boldsymbol{\gamma} - \mathbf{p}) = \text{constant}, \tag{16.36}$$

for all s. Suppose that at $s = 0$, we take

$$\mathbf{M}_1(0) = \frac{\boldsymbol{\gamma}(0) - \mathbf{p}}{|\boldsymbol{\gamma}(0) - \mathbf{p}|}, \tag{16.37}$$

where we note that $|\boldsymbol{\gamma} - \mathbf{p}| = R$, for all s. Then from (16.36) and $\mathbf{M}_1 \cdot \mathbf{M}_2 = 0$, it follows that

$$\mathbf{M}_1 \cdot (\boldsymbol{\gamma} - \mathbf{p}) = R, \quad \mathbf{M}_2 \cdot (\boldsymbol{\gamma} - \mathbf{p}) = 0. \tag{16.38}$$

Using $\mathbf{T}' = k_1 \mathbf{M}_1 + k_2 \mathbf{M}_2$ and (16.33),

$$\mathbf{T}' \cdot (\boldsymbol{\gamma} - \mathbf{p}) = (k_1 \mathbf{M}_1 + k_2 \mathbf{M}_2) \cdot (\boldsymbol{\gamma} - \mathbf{p}) = k_1 R = -1, \tag{16.39}$$

so that $k_1 = -1/R$. Thus, a spherical curve is described by (16.29), where $\mathbf{M}_1$ is the outward-pointing normal to the surface of the sphere, $k_1 = -1/R$, $\mathbf{M}_2 = \mathbf{T} \times \mathbf{M}_1$, and k_2 can be interpreted as a "steering control" for the curve evolving with respect to the arc-length parameter along the surface of the sphere (Fig. 16.4).

16.4.3 Free-Particle Interaction with the Spherical Curve

The motion of a free particle moving at unit speed can be described by (16.29), where the arc-length parameter s is identified with time t. To distinguish the free

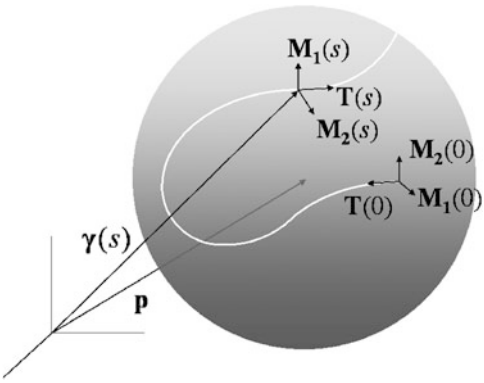

Fig. 16.4 A spherical curve $\boldsymbol{\gamma}(s)$ and its natural frame $\{\mathbf{T}, \mathbf{M}_1, \mathbf{M}_2\}$. The center of the sphere is denoted by $\mathbf{p}$

particle from the guide point $\boldsymbol{\gamma}$ confined to the surface of the sphere, we introduce notation

$$\begin{aligned} \dot{\boldsymbol{\sigma}} &= \boldsymbol{\tau}, \\ \dot{\boldsymbol{\tau}} &= u\mathbf{m}_1 + v\mathbf{m}_2, \\ \dot{\mathbf{m}}_1 &= -u\boldsymbol{\tau}, \\ \dot{\mathbf{m}}_2 &= -v\boldsymbol{\tau}, \end{aligned} \tag{16.40}$$

where $\boldsymbol{\sigma}$ is the position of the free particle (as a function of time), $\boldsymbol{\tau}$ is the unit tangent vector to the trajectory of the free particle, $(\mathbf{m}_1, \mathbf{m}_2)$ are corresponding unit normal vectors (which complete a right-handed orthonormal frame), and (u, v) are natural curvatures (the three-dimensional analog to steering controls) for the free particle (Fig. 16.5).

The induced dynamics of the guide point are then

$$\begin{aligned} \dot{\boldsymbol{\gamma}} &= \nu\mathbf{T}, \\ \dot{\mathbf{T}} &= \nu[(-1/R)\mathbf{M}_1 + k_2\mathbf{M}_2], \\ \dot{\mathbf{M}}_1 &= -\nu(-1/R)\mathbf{T}, \\ \dot{\mathbf{M}}_2 &= -\nu k_2\mathbf{T}, \end{aligned} \tag{16.41}$$

where $\nu = ds/dt$ is the speed with which the guide point moves along the guiding curve. As before, we define $\mathbf{r} = \boldsymbol{\sigma} - \boldsymbol{\gamma}$ to be the vector from the guide point to the free particle. The first-order necessary condition for the guide point is $\mathbf{r} \cdot \mathbf{T} \equiv 0$.

16.4.4 Lyapunov Function and Control Law Derivation

System (16.40), for prescribed (u, v) (and prescribed $\mathbf{m}_1(0)$, $\mathbf{m}_2(0)$), describes the evolution of the free particle trajectory. However, for the free particle interacting

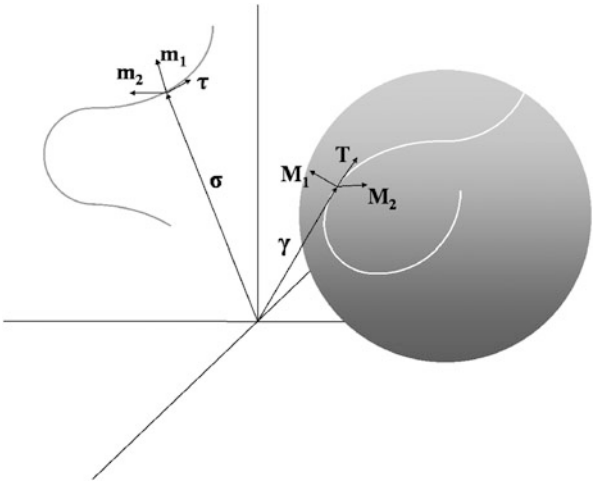

Fig. 16.5 Positions and frames for the trajectory of the free particle and for the closest point on the spherical curve

with a spherical curve, (u, v) are considered to be feedback functions of certain *shape variables*, to be identified below. To simplify the derivation of the control law, it is convenient to express (u, v) as

$$u = \mathbf{a} \cdot \mathbf{m_1}, \quad v = \mathbf{a} \cdot \mathbf{m_2}, \tag{16.42}$$

so that $\dot{\boldsymbol{\tau}} = \mathbf{a} - (\mathbf{a} \cdot \boldsymbol{\tau})\boldsymbol{\tau}$, where the vector $\mathbf{a}$ is a feedback function of the shape variables. (Expressing (u, v) in the form (16.42) respects the requirement that the trajectory of the free particle under feedback control be independent of the specific choice of $\mathbf{m}_1(0)$ and $\mathbf{m}_2(0)$.) The projection of the vector $\mathbf{a}$ onto the $\mathbf{m}_1$-$\mathbf{m}_2$ plane can be interpreted as the acceleration of the free particle (which is assumed to have unit mass): this projection is perpendicular to the direction of motion of the free particle, and therefore leaves its speed unchanged.

Consider the Lyapunov function candidate

$$V_{3D} = -\ln(\mathbf{T} \cdot \boldsymbol{\tau}) + h(|\boldsymbol{\sigma} - \mathbf{p}|) + h_1(\mathbf{r} \cdot \mathbf{M}_1) + h_2(\mathbf{r} \cdot \mathbf{M}_2), \tag{16.43}$$

where $h(\cdot)$ satisfies

(B1) $dh/d\rho = f(\rho)$, where $f(\rho)$ is a Lipschitz continuous function on (R, ∞), so that $h(\rho)$ is continuously differentiable on (R, ∞);
(B2) $f(R + r_0) = 0$ at a unique point $0 < r_0 < \infty$;
(B3) $\lim_{\rho \to R} h(\rho) = \infty$, $\lim_{\rho \to \infty} h(\rho) = \infty$, and $h(r_0) < \infty$;

(analogous to hypotheses (A1), (A2), and (A3) of **Proposition 16.1**), and $h_1(\cdot)$ and $h_2(\cdot)$ are continuously differentiable functions satisfying

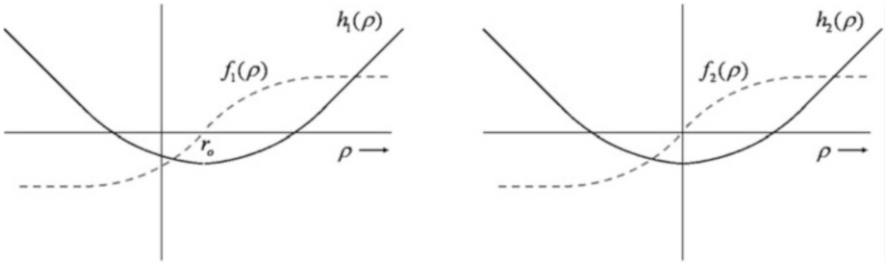

Fig. 16.6 *Left plot*: examples of functions $h_1(\cdot)$ and $f_1(\cdot)$. *Right plot*: examples of functions $h_2(\cdot)$ and $f_2(\cdot)$

(B4) $dh_j/d\rho = f_j(\rho)$, where $f_j(\rho)$ is a Lipschitz continuous function on $(-\infty, \infty)$, so that $h_j(\rho)$ is continuously differentiable on $(-\infty, \infty)$ for $j = 1, 2$;

(B5) $f_j(\rho_j) = 0$ at a single point ρ_j, for $j = 1, 2$; furthermore, $\rho_1 > 0$ and $\rho_2 = 0$;

(B6) $\lim_{\rho\to\infty} h_j(\rho) = \infty$, $\lim_{\rho\to-\infty} h_j(\rho) = \infty$, and $h(\rho_j)$ is finite, for $j = 1, 2$.

Figure 16.6 shows examples of suitable functions h_1 and h_2 with the corresponding f_1 and f_2.

Differentiating V_{3D} with respect to time along trajectories of (16.40) and (16.41) gives

$$\dot{V}_{3D} = -\frac{\dot{\mathbf{T}}\cdot\boldsymbol{\tau} + \mathbf{T}\cdot\dot{\boldsymbol{\tau}}}{\mathbf{T}\cdot\boldsymbol{\tau}} + \left(\frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}\cdot\boldsymbol{\tau}\right) f(|\boldsymbol{\sigma}-\mathbf{p}|) + (\dot{\mathbf{r}}\cdot\mathbf{M}_1 + \mathbf{r}\cdot\dot{\mathbf{M}}_1) f_1(\mathbf{r}\cdot\mathbf{M}_1) + (\dot{\mathbf{r}}\cdot\mathbf{M}_2 + \mathbf{r}\cdot\dot{\mathbf{M}}_2) f_2(\mathbf{r}\cdot\mathbf{M}_2), \quad (16.44)$$

where we have used $dh/d\rho = f(\rho)$, and $dh_j/d\rho = f_j(\rho)$, $j = 1, 2$. Noting that

$$\dot{\mathbf{r}} = \dot{\boldsymbol{\sigma}} - \dot{\boldsymbol{\gamma}} = \boldsymbol{\tau} - \nu\mathbf{T}, \quad (16.45)$$

along with the first-order necessary condition $\mathbf{r}\cdot\mathbf{T} = 0$ for the guide point, we further compute

$$\begin{aligned}
\dot{V}_{3D} &= -\frac{\nu(k_1\mathbf{M}_1 + k_2\mathbf{M}_2)\cdot\boldsymbol{\tau} + \mathbf{T}\cdot(u\mathbf{m}_1 + v\mathbf{m}_2)}{\mathbf{T}\cdot\boldsymbol{\tau}} + \left(\frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}\cdot\boldsymbol{\tau}\right) f(|\boldsymbol{\sigma}-\mathbf{p}|) \\
&\quad + (\boldsymbol{\tau}\cdot\mathbf{M}_1) f_1(\mathbf{r}\cdot\mathbf{M}_1) + (\boldsymbol{\tau}\cdot\mathbf{M}_2) f_2(\mathbf{r}\cdot\mathbf{M}_2) \\
&= -\frac{1}{\mathbf{T}\cdot\boldsymbol{\tau}}\Big[\nu k_1(\boldsymbol{\tau}\cdot\mathbf{M}_1) + \nu k_2(\boldsymbol{\tau}\cdot\mathbf{M}_2) + u(\mathbf{T}\cdot\mathbf{m}_1) + v(\mathbf{T}\cdot\mathbf{m}_2) - (\mathbf{b}\cdot\boldsymbol{\tau})(\mathbf{T}\cdot\boldsymbol{\tau})\Big] \\
&= -\frac{1}{\mathbf{T}\cdot\boldsymbol{\tau}}\Big[\nu k_1(\boldsymbol{\tau}\cdot\mathbf{M}_1) + \nu k_2(\boldsymbol{\tau}\cdot\mathbf{M}_2) + (\mathbf{a}\cdot\mathbf{T}) - (\mathbf{a}\cdot\boldsymbol{\tau})(\mathbf{T}\cdot\boldsymbol{\tau}) - (\mathbf{b}\cdot\boldsymbol{\tau})(\mathbf{T}\cdot\boldsymbol{\tau})\Big],
\end{aligned} \quad (16.46)$$

where

$$\mathbf{b} = f(|\boldsymbol{\sigma} - \mathbf{p}|)\frac{\boldsymbol{\sigma} - \mathbf{p}}{|\boldsymbol{\sigma} - \mathbf{p}|} + f_1(\mathbf{r} \cdot \mathbf{M}_1)\mathbf{M}_1 + f_2(\mathbf{r} \cdot \mathbf{M}_2)\mathbf{M}_2, \tag{16.47}$$

and we have used the identity

$$(\mathbf{a} \cdot \mathbf{m}_1)(\mathbf{T} \cdot \mathbf{m}_1) + (\mathbf{a} \cdot \mathbf{m}_2)(\mathbf{T} \cdot \mathbf{m}_2) = \mathbf{a} \cdot \mathbf{T} - (\mathbf{a} \cdot \boldsymbol{\tau})(\mathbf{T} \cdot \boldsymbol{\tau}). \tag{16.48}$$

Suppose that we take

$$\mathbf{a} = \tilde{\mathbf{a}} - \mathbf{b}, \tag{16.49}$$

where $\tilde{\mathbf{a}}$ consists of terms in the control law for the free particle which are yet to be specified. Then using $\mathbf{b} \cdot \mathbf{T} = 0$, which follows from $\mathbf{M}_1 \cdot \mathbf{T} = 0$, $\mathbf{M}_2 \cdot \mathbf{T} = 0$, and

$$(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{T} = (\mathbf{r} + (\boldsymbol{\gamma} - \mathbf{p})) \cdot \mathbf{T} = 0, \tag{16.50}$$

we obtain

$$\dot{V}_{3D} = -\frac{1}{\mathbf{T} \cdot \boldsymbol{\tau}}\left[\nu k_1(\boldsymbol{\tau} \cdot \mathbf{M}_1) + \nu k_2(\boldsymbol{\tau} \cdot \mathbf{M}_2) + (\tilde{\mathbf{a}} \cdot \mathbf{T}) - (\tilde{\mathbf{a}} \cdot \boldsymbol{\tau})(\mathbf{T} \cdot \boldsymbol{\tau})\right]. \tag{16.51}$$

Furthermore, for $\mathbf{T} \cdot \boldsymbol{\tau} \neq 0$ we can take

$$\tilde{\mathbf{a}} = \mu\mathbf{T} + \frac{\nu k_1 \mathbf{M}_1 + \nu k_2 \mathbf{M}_2}{\mathbf{T} \cdot \boldsymbol{\tau}} + \eta\left[\mathbf{m}_2(\mathbf{T} \cdot \mathbf{m}_1) - \mathbf{m}_1(\mathbf{T} \cdot \mathbf{m}_2)\right], \tag{16.52}$$

where $\mu > 0$ is a constant (or else a function of the shape variables defined below), and η is an arbitrary scalar-valued function, to obtain

$$\dot{V}_{3D} = -\frac{\mu}{\mathbf{T} \cdot \boldsymbol{\tau}}\left[1 - (\mathbf{T} \cdot \boldsymbol{\tau})^2\right]. \tag{16.53}$$

It is clear from (16.53) that $\dot{V}_{3D} \leq 0$ for all t, and $\dot{V}_{3D} = 0$ if and only if $\mathbf{T} = \boldsymbol{\tau}$ (provided $\mathbf{T} \cdot \boldsymbol{\tau} > 0$). The term involving η in (16.52) will be used below to ensure that with the feedback law applied, the resulting system of shape variables (i.e., an appropriate reduced system) is autonomous.

16.4.5 Control Law Interpretation

To summarize (16.47), (16.49), and (16.52), the control law we propose for the free particle moving in three dimensions while interacting with a spherical curve (i.e., guiding curve) is given by (16.42) where

$$\mathbf{a} = \mu\mathbf{T} - f(|\boldsymbol{\sigma} - \mathbf{p}|)\frac{\boldsymbol{\sigma} - \mathbf{p}}{|\boldsymbol{\sigma} - \mathbf{p}|} - f_1(\mathbf{r}\cdot\mathbf{M}_1)\mathbf{M}_1 - f_2(\mathbf{r}\cdot\mathbf{M}_2)\mathbf{M}_2 + \frac{\nu k_1\mathbf{M}_1 + \nu k_2\mathbf{M}_2}{\mathbf{T}\cdot\boldsymbol{\tau}} + \eta\left[\mathbf{m}_2(\mathbf{T}\cdot\mathbf{m}_1) - \mathbf{m}_1(\mathbf{T}\cdot\mathbf{m}_2)\right], \tag{16.54}$$

with $k_1 = -1/R$, and the speed ν of the guide point is

$$\nu = \frac{\mathbf{T}\cdot\boldsymbol{\tau}}{1 - \mathbf{r}\cdot(k_1\mathbf{M}_1 + k_2\mathbf{M}_2)} \tag{16.55}$$

(which follows from differentiating $\mathbf{r}\cdot\mathbf{T} \equiv 0$ with respect to t).

We can give the terms in $\mathbf{a}$ with the following physical interpretations. The term $\mu\mathbf{T}$ serves to align the heading of the free particle with the heading of the guide point. The term involving $f(\cdot)$ provides collision avoidance: it serves to steer the free particle toward the sphere if the free particle is too far away, and steer the free particle away from the sphere if the free particle comes too close to it. The terms involving $f_1(\cdot)$ and $f_2(\cdot)$ serve to steer the free particle toward a sphere concentric with the sphere on which the guiding curve lies. Finally, the term involving ν, $k_1 = -1/R$, and k_2 enables the free particle to respond to the curvature of the guiding curve. (As mentioned in the previous subsection, the term involving η is needed to make the shape dynamics autonomous.)

16.4.6 Strategy and Invariant Submanifold

Similar to (16.12), we can identify the spherical curve-tracking strategy with the submanifold

$$M_{3D} = \left\{ \left(\begin{bmatrix} \mathbf{T} & \mathbf{M}_1 & \mathbf{M}_2 & \boldsymbol{\gamma} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\tau} & \mathbf{m}_1 & \mathbf{m}_2 & \boldsymbol{\sigma} \\ 0 & 0 & 0 & 1 \end{bmatrix} \right) \in SE(3)\times SE(3) \,\middle|\, \mathbf{T} = \boldsymbol{\tau},\ \mathbf{r}\cdot\mathbf{T} = 0,\ \ \mathbf{r} = \boldsymbol{\sigma} - \boldsymbol{\gamma} \right\}. \tag{16.56}$$

Note that in contrast to definition (16.12) of $M_{2D}(r_0)$, this definition of M_{3D} does not incorporate a separation parameter r_0. While $\mathbf{T} \equiv \boldsymbol{\tau}$ and $\mathbf{r}\cdot\mathbf{T} \equiv 0$ together imply that $|\mathbf{r}| \equiv$ constant, we shall see below that the particular constant is not determined exclusively by the choice of steering law (specifically, f, f_1, and f_2). Initial conditions also play a role.

While (16.12) was invariant under the planar boundary-tracking steering law, (16.56) is not invariant under the three-dimensional boundary-tracking law (16.54). Therefore, we define the additional submanifold

$$M_{3DI} = \left\{ \left(\begin{bmatrix} \mathbf{T} & \mathbf{M}_1 & \mathbf{M}_2 & \boldsymbol{\gamma} \\ 0 & 0 & 0 & 1 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\tau} & \mathbf{m}_1 & \mathbf{m}_2 & \boldsymbol{\sigma} \\ 0 & 0 & 0 & 1 \end{bmatrix} \right) \in M_{3D} \right| \\ f(|\boldsymbol{\sigma} - \mathbf{p}|) \left(\frac{\boldsymbol{\sigma} - \mathbf{p}}{|\boldsymbol{\sigma} - \mathbf{p}|} \cdot \mathbf{M}_1 \right) = -f_1(\mathbf{r} \cdot \mathbf{M}_1), \\ f(|\boldsymbol{\sigma} - \mathbf{p}|) \left(\frac{\boldsymbol{\sigma} - \mathbf{p}}{|\boldsymbol{\sigma} - \mathbf{p}|} \cdot \mathbf{M}_2 \right) = -f_2(\mathbf{r} \cdot \mathbf{M}_2) \Big\}, \quad (16.57)$$

where $\mathbf{p}$ is the prescribed constant vector denoting the center of the spherical obstacle. On M_{3DI}, (16.54) becomes

$$\mathbf{a} = \mu \mathbf{T} + \nu k_1 \mathbf{M}_1 + \nu k_2 \mathbf{M}_2, \quad (16.58)$$

which gives

$$\dot{\boldsymbol{\tau}} = \mathbf{a} - (\mathbf{a} \cdot \boldsymbol{\tau})\boldsymbol{\tau} = \nu k_1 \mathbf{M}_1 + \nu k_2 \mathbf{M}_2 = \dot{\mathbf{T}}, \quad (16.59)$$

enforcing $\mathbf{T} = \boldsymbol{\tau}$ for all future time. Furthermore, it is easily verified that $\frac{d}{dt}|\boldsymbol{\sigma} - \mathbf{p}| = 0$ on M_{3DI}, along with $\frac{d}{dt}[(\boldsymbol{\sigma} - \boldsymbol{\gamma}) \cdot \mathbf{M}_i] = 0$ and $\frac{d}{dt}\left(\frac{\boldsymbol{\sigma} - \mathbf{p}}{|\boldsymbol{\sigma} - \mathbf{p}|} \cdot \mathbf{M}_i \right) = 0$, $i = 1, 2$, Thus, control law (16.54) leaves M_{3DI} invariant. Below, we show that (16.54) also gives convergence to M_{3DI} (from initial conditions specified by a sublevel set of the Lyapunov function V_{3D}), by which we conclude that (16.54) realizes a curve tracking strategy for spherical guiding curves.

16.4.7 Shape Variables

To apply LaSalle's Invariance Principle [9] and prove a convergence result for the dynamics of the free particle interacting with the spherical curve, we need to first identify appropriate shape variables. Next, we need to derive the shape dynamics, and choose η such that the shape dynamics are a self-contained, autonomous system (so that we can apply the usual LaSalle invariance principle for autonomous systems).

From (16.43), the definition of V_{3D}, we identify the following as being among the shape variables: $(\mathbf{T} \cdot \boldsymbol{\tau})$, $|\boldsymbol{\sigma} - \mathbf{p}|$, $(\mathbf{r} \cdot \mathbf{M}_1)$, and $(\mathbf{r} \cdot \mathbf{M}_2)$. Further calculations, shown in the Appendix, suggest that a good collection of shape variables to use is

$$0 = \mathbf{r} \cdot \mathbf{T},$$
$$z_1 = \mathbf{r} \cdot \mathbf{M}_1,$$
$$z_2 = \mathbf{r} \cdot \mathbf{M}_2,$$
$$z_3 = \boldsymbol{\tau} \cdot \mathbf{T},$$

$$
\begin{aligned}
z_4 &= \boldsymbol{\tau} \cdot \mathbf{M}_1, \\
z_5 &= \boldsymbol{\tau} \cdot \mathbf{M}_2, \\
0 &= (\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{T}, \\
z_6 &= (\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_1, \\
z_7 &= (\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_2,
\end{aligned} \tag{16.60}
$$

where they have been arranged so as to emphasize that the shape variables are simply $\mathbf{r}$, $\boldsymbol{\tau}$, and $(\boldsymbol{\sigma} - \mathbf{p})$ expressed in the (time-varying) frame of the guide point.

We define η by

$$
\eta = \frac{-\nu k_1 (\mathbf{M}_2 \cdot \boldsymbol{\tau}) + \nu k_2 (\mathbf{M}_1 \cdot \boldsymbol{\tau})}{\mathbf{T} \cdot \boldsymbol{\tau}}, \tag{16.61}
$$

so that the shape dynamics become autonomous. The shape dynamics can then be expressed (see the Appendix) as

$$
\begin{aligned}
\dot{z}_1 &= z_4, \\
\dot{z}_2 &= z_5, \\
\dot{z}_3 &= \mu \left(1 - z_3^2\right) + z_3 \chi, \\
\dot{z}_4 &= -f\left(\sqrt{z_6^2 + z_7^2}\right) \frac{z_6}{\sqrt{z_6^2 + z_7^2}} - f_1(z_1) + z_4(-\mu z_3 + \chi), \\
\dot{z}_5 &= -f\left(\sqrt{z_6^2 + z_7^2}\right) \frac{z_7}{\sqrt{z_6^2 + z_7^2}} - f_2(z_2) + z_5(-\mu z_3 + \chi), \\
\dot{z}_6 &= z_4, \\
\dot{z}_7 &= z_5,
\end{aligned} \tag{16.62}
$$

where

$$
\chi = f\left(\sqrt{z_6^2 + z_7^2}\right) \frac{(z_4 z_6 + z_5 z_7)}{\sqrt{z_6^2 + z_7^2}} + f_1(z_1) z_4 + f_2(z_2) z_5. \tag{16.63}
$$

Note that if initially $z_3^2 + z_4^2 + z_5^2 = 1$, then under (16.62), this condition holds for all future time. Furthermore,

$$
V_{3D} = -\ln(z_3) + h\left(\sqrt{z_6^2 + z_7^2}\right) + h_1(z_1) + h_2(z_2), \tag{16.64}
$$

and

$$
\dot{V}_{3D} = -\frac{\mu}{z_3}\left(1 - z_3^2\right). \tag{16.65}
$$

Thus, closing the loop with the feedback law given by (16.42) and (16.54), we see that system (16.62) for the shape variables is autonomous, and in addition, the Lyapunov function can be expressed in terms of shape variables alone.

Observe that $z_3 \to 0$, which causes $\dot{V}_{3D}$ to become undefined, also implies that $V_{3D} \to \infty$. Therefore, the $z_3 \to 0$ singularity, which would also prevent η in (16.61) from being well defined, can be avoided by verifying that V_{3D} is radially unbounded, and then restricting attention to sublevel sets of V_{3D}.

However, in order for system (16.62) to be properly defined, we also require that ν be well defined for all time. From (16.55), we see that

$$\mathbf{r} \cdot (k_1 \mathbf{M}_1 + k_2 \mathbf{M}_2) = 1, \tag{16.66}$$

or equivalently

$$k_2(\mathbf{r} \cdot \mathbf{M}_2) = 1 + (1/R)(\mathbf{r} \cdot \mathbf{M}_1), \tag{16.67}$$

causes ν to be infinite. Expressed in shape variables, (16.67) becomes

$$k_2 z_2 = 1 + (1/R) z_1, \tag{16.68}$$

which can be avoided if z_2 is sufficiently small, $|k_2|$ is bounded, and $z_1 > 0$ is sufficiently away from zero.

16.4.8 Convergence Result

Proposition 16.2. *Suppose that hypotheses (B1)–(B6) are satisfied, along with*

(B7) $\mu > 0$ *is a Lipschitz continuous function of* $(z_1, \ldots, z_7)$*;*
(B8) $|k_2|$ *is bounded;*
(B9) $0 < R < \infty$.

Also, suppose that initially $z_3^2 + z_4^2 + z_5^2 = 1$ *with* $0 < z_3 \leq 1$, *and* $\sqrt{z_6^2 + z_7^2} > R$. *Then there exists a finite neighborhood* Λ *of a global minimizer for* V_{3D} *such that for arbitrary initial conditions in* Λ, *system (16.62) converges to the set of its equilibria for which* $z_3 = 1$ *and* $z_4 = z_5 = 0$.

Proof. Note that $V_{3D} \to \infty$ as $|z_1| \to \infty$, $|z_2| \to \infty$, $|z_6| \to \infty$, $|z_7| \to \infty$, $\sqrt{z_6^2 + z_7^2} \to R$, or $z_3 \to 0$. Also, $z_3^2 + z_4^2 + z_5^2 \equiv 1$. Therefore, the sublevel sets of V_{3D} are compact.

From the form of V_{3D} in (16.64), it is clear that V_{3D} is bounded below, and is minimized when $z_1 = \rho_1$, $z_2 = 0$, $z_3 = 1$, and $\sqrt{z_6^2 + z_7^2} = r_0$. Thus, there exists a constant $\xi > 0$ such that

$$\forall \mathbf{z} \in \Lambda = \left\{ \mathbf{z} \;\middle|\; V_{3D}(\mathbf{z}) \leq V_{3D}(\mathbf{z}_{min}) + \xi \right\}, \quad z_2 < \frac{1 + (1/R) z_1}{k_{2max}}, \tag{16.69}$$

where $\mathbf{z} = (z_1, z_2, z_3, z_4, z_5, z_6)$, $\mathbf{z}_{min}$ is a global minimizer of V_{3D} and $k_{2max} > 0$ is an upper bound on $|k_2|$.

Hence, $\forall \mathbf{z} \in \Lambda$, $\dot{V}_{3D}$ exists, satisfies $\dot{V}_{3D} \leq 0$, and $\dot{V}_{3D} = 0$ if and only if $z_3 = 1$.

Then by LaSalle's Invariance Principle (for autonomous systems) [9], we conclude that system (16.62) converges to the largest invariant set with $z_3 = 1$. But $z_3 = 1$ implies $z_4 = z_5 = 0$, and (16.62) reduces to

$$\begin{aligned}
\dot{z}_1 &= 0, \\
\dot{z}_2 &= 0, \\
\dot{z}_3 &= 0, \\
\dot{z}_4 &= -f\left(\sqrt{z_6^2 + z_7^2}\right) \frac{z_6}{\sqrt{z_6^2 + z_7^2}} - f_1(z_1) = 0, \\
\dot{z}_5 &= -f\left(\sqrt{z_6^2 + z_7^2}\right) \frac{z_7}{\sqrt{z_6^2 + z_7^2}} - f_2(z_2) = 0, \\
\dot{z}_6 &= 0, \\
\dot{z}_7 &= 0.
\end{aligned} \tag{16.70}$$

Thus, (16.62) converges to the set of its equilibria for which $z_3 = 1$, $z_4 = z_5 = 0$,

$$f_1(z_1) = -f\left(\sqrt{z_6^2 + z_7^2}\right) \frac{z_6}{\sqrt{z_6^2 + z_7^2}}, \tag{16.71}$$

and

$$f_2(z_2) = -f\left(\sqrt{z_6^2 + z_7^2}\right) \frac{z_7}{\sqrt{z_6^2 + z_7^2}}. \tag{16.72}$$

□

Remark 16.4. If we square both sides of (16.71) and (16.72), and sum the result, we obtain

$$f_1^2(z_1) + f_2^2(z_2) = f^2\left(\sqrt{z_6^2 + z_7^2}\right), \tag{16.73}$$

with $z_1^2 + z_2^2 = \text{constant} = r_0^2$, where r_0 denotes the distance between the free particle and the closest point on the guiding curve (which is not necessarily the same as the r_0 depicted in Figs. 16.2 and 16.6), and $z_6^2 + z_7^2 = \text{constant} = R_0^2$, where $R_0 > R$ denotes the distance between the free particle and the center of the spherical obstacle. We can thus introduce an angle variable ϕ and write (16.73) as

$$f_1^2(r_0 \cos\phi) + f_2^2(r_0 \sin\phi) = f^2(R_0)\,. \tag{16.74}$$

A necessary condition for (16.71) and (16.72) to be satisfied is that (16.74) be satisfied for some r_0, R_0, and ϕ. Observe that if r_0 in assumption (B2) and ρ_1 in assumption (B5) satisfy $r_0 = \rho_1 = r_0$, and we have $R + r_0 = R_0$ and $\phi = 0$, then (16.74) is satisfied with both sides of the equation equal to zero. (However, this is not the only possible solution.)

Remark 16.5. A similar approach has been followed in [18] to derive three-dimensional curve tracking control laws for a particle to track one of the lines of curvature on a smooth surface viewed as a level surface of a scalar-valued function on $\mathbb{R}^3$, which is motivated by cooperative sensing applications. The control laws in [18] have not been designed to avoid collisions since there is no need for this functionality. In addition, these control laws would have difficulty tracking a specific curve on a spherical or planar surface, since any curve would be a line of curvature. The control law (16.54) is more general than the ones used in [18] since it achieves both collision avoidance and tracking for a specific curve at the same time.

Remark 16.6. A result analogous to **Proposition 16.2** can be proved for a planar rather than spherical obstacle surface (i.e., the limiting situation as $R \to \infty$). In (16.41), $k_1 = -1/R \equiv 0$, and $\mathbf{M}_1$ remains normal to the planar obstacle surface for all time. The speed of the guide point becomes

$$\nu = \frac{\mathbf{T} \cdot \boldsymbol{\tau}}{1 - k_2(\mathbf{r} \cdot \mathbf{M}_2)}, \tag{16.75}$$

which is well defined for $z_2 = \mathbf{r} \cdot \mathbf{M}_2$ sufficiently small. The shadow point is now the closest point on the planar obstacle surface to the free particle, and we replace $|\boldsymbol{\sigma} - \mathbf{p}|$ in the argument of $h(\cdot)$ in (16.43) by $|\boldsymbol{\sigma} - \boldsymbol{\lambda}|$, where $\boldsymbol{\lambda}$ denotes the shadow point. We then note that

$$|\boldsymbol{\sigma} - \boldsymbol{\lambda}| = |\mathbf{r} \cdot \mathbf{M}_1|, \tag{16.76}$$

and we may assume without loss of generality that initially $\mathbf{r} \cdot \mathbf{M}_1 > 0$. Thus, the function $h_1(\cdot)$ in (16.43) may be set to zero (i.e., h_1 may be subsumed into h, where $\lim_{\rho \to 0} h(\rho) = \infty$ and $\lim_{\rho \to \infty} h(\rho) = \infty$). To summarize, for a planar obstacle surface, we have

$$\overline{V}_{3D} = -\ln(\mathbf{T} \cdot \boldsymbol{\tau}) + h(\mathbf{r} \cdot \mathbf{M}_1) + h_2(\mathbf{r} \cdot \mathbf{M}_2), \tag{16.77}$$

and

$$\begin{aligned} \bar{\mathbf{a}} = {} & \mu\mathbf{T} - f(\mathbf{r} \cdot \mathbf{M}_1)\mathbf{M}_1 - f_2(\mathbf{r} \cdot \mathbf{M}_2)\mathbf{M}_2 \\ & + \frac{\nu k_2}{\mathbf{T} \cdot \boldsymbol{\tau}} \left[\mathbf{M}_2 + (\mathbf{M}_1 \cdot \boldsymbol{\tau}) \Big(\mathbf{m}_2(\mathbf{T} \cdot \mathbf{m}_1) - \mathbf{m}_1(\mathbf{T} \cdot \mathbf{m}_2) \Big) \right], \end{aligned} \tag{16.78}$$

which lead to the shape dynamics

$$\begin{aligned}
\dot{z}_1 &= z_4, \\
\dot{z}_2 &= z_5, \\
\dot{z}_3 &= \mu\left(1 - z_3^2\right) + z_3\overline{\chi}, \\
\dot{z}_4 &= -f(z_1) + z_4(-\mu z_3 + \overline{\chi}), \\
\dot{z}_5 &= -f_2(z_2) + z_5(-\mu z_3 + \overline{\chi}),
\end{aligned} \tag{16.79}$$

where

$$\overline{\chi} = f(z_1)z_4 + f_2(z_2)z_5. \tag{16.80}$$

Note that the dimension of the shape space is now 5 (rather than 7, as in **Proposition 16.2**).

16.4.9 Simulation Example

To illustrate the three-dimensional spherical-curve tracking law, Fig. 16.7 shows a free particle tracking a circular curve lying on the surface of a sphere. After an initial transient, the free particle trajectory is observed to converge to a circular orbit on a sphere concentric to the one shown. The trajectory of the guide point on the circular curve being tracked is also shown. Observe that during the initial transient, the free particle approaches, and then feels the influence of, the spherical surface—causing it to steer away from the sphere to avoid colliding with it.

For the simulation shown in Fig. 16.7, the functions $f(\cdot)$, $f_1(\cdot)$, and $f_2(\cdot)$ are given by

$$\begin{aligned}
f(\rho) &= \alpha\left[1 - \frac{(R + r_0)^2 - R^2}{\rho^2 - R^2}\right], \\
f_1(\rho) &= \alpha\tanh(\rho - r_0), \\
f_2(\rho) &= \alpha\tanh(\rho),
\end{aligned} \tag{16.81}$$

where r_0 is the target separation between the free particle and the guiding curve, and α is a positive constant. Observe that if the guide point on the guiding curve and the shadow point on the sphere coincide, and $|\mathbf{r}| = r_0$, we have

$$\begin{aligned}
f(|\boldsymbol{\sigma} - \mathbf{p}|) &= f(R + r_0) = 0, \\
f_1(\mathbf{r} \cdot \mathbf{M}_1) &= f_1(r_0) = 0.
\end{aligned} \tag{16.82}$$

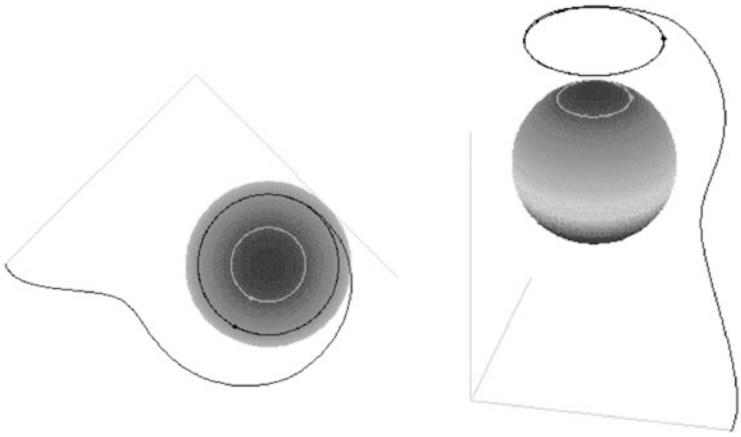

Fig. 16.7 Simulation of a free particle interacting with a circular curve lying on the surface of a sphere (from two different vantage points)

Furthermore, as the distance between the free particle and the surface of the sphere goes to zero, $f(|\boldsymbol{\sigma} - \mathbf{p}|) \to \infty$.

Taking the spherical curve to be a circular curve guarantees that (almost everywhere) the guide point is unique and given by a simple formula. The natural curvatures for the guide curve are also constant (recall that $k_1 = -1/R$ is constant for any spherical curve, so what is special about the circular curve is that k_2 is also an easily computed constant).

16.5 Conclusions

In the plane, collision avoidance and boundary tracking are defined with respect to the same planar boundary curve and are handled together by our choice of feedback law. Furthermore, if f has a unique zero at r_0, then the shape space equilibrium is unique, with the separation between the free particle and boundary curve equal to r_0, and the corresponding "relative equilibrium" is a Bertrand mate to the boundary curve.

In three dimensions, collision avoidance with a sphere corresponds to maintaining positive separation from the sphere, distinct from tracking a curve on the sphere. In particular, the parameters R_0 and r_0 associated with these two goals are not uniquely determined (selected) by the shape space equilibrium conditions. But the confinement to a sublevel set of V_{3D} on the shape space ($\mathbf{z}$-space) implies upper and lower bounds for R_0 and r_0, provided some natural specifications are made for $f = dh/d\rho$ and $f_i = dh/d\rho_i$, $i = 1, 2$. These bounds are dependent on initial conditions in $\mathbf{z}$-space.

Modeling high-speed vehicles (or animals) as self-propelled particles subject to gyroscopic (steering) feedback control has several key advantages. As we have shown here in the context of boundary tracking, this level of modeling is amenable to analysis, even in three dimensions with the simultaneous objectives of curve tracking and obstacle avoidance. Furthermore, the steering laws we propose have straightforward physical interpretations, and the implications for sensing follow naturally. This provides a basis for hypothesizing the biological plausibility of (approximations to) these steering laws.

We speculate that boundary tracking with obstacle avoidance may play an essential role in various biological schooling/flocking behaviors, where instead of a fixed, stationary obstacle surface (as we have assumed above for purposes of analysis), the "obstacle" is actually a perceptual construct encompassing the rest of the school/flock from the viewpoint of an individual. Just as pursuit at the level of individuals can lead to cohesion at the level of collective behavior, so too can boundary tracking (with the correct interpretation assigned to the "boundary"). In the collective behavior setting, our guiding curve would play the role of a further perceptual construct for an individual's desired motion relative to the "boundary"—for example, tracking the path of a particular neighbor. Indeed, both physical and perceptual boundaries could drive individuals' steering, and a basic question is whether generalizations of the steering laws presented here can explain observed schooling/flocking patterns in nature.

Acknowledgements This research was supported in part by the Naval Research Laboratory under Grant Nos. N00173-02-1G002, N00173-03-1G001, N00173-03-1G019, and N00173-04-1G014; by the Air Force Office of Scientific Research under AFOSR Grant Nos. F49620-01-0415, FA95500410130, and FA95501010250; by the Army Research Office under ODDR&E MURI01 Program Grant No. DAAD19-01-1-0465 to the Center for Communicating Networked Control Systems (through Boston University); and by ONR MURI Grant No. N000140710734. P.S. Krishnaprasad also gratefully acknowledges the support of the Bernoulli Center at EPFL, Lausanne during the early stages of this body of work. E.W. Justh was supported in part by the Office of Naval Research.

Appendix

These are the calculations required to obtain (16.62). First, note that

$$\begin{aligned}\frac{d}{dt}(\mathbf{T}\cdot\boldsymbol{\tau}) &= \dot{\mathbf{T}}\cdot\boldsymbol{\tau}+\mathbf{T}\cdot\dot{\boldsymbol{\tau}} = \dot{\mathbf{T}}\cdot\boldsymbol{\tau}+\mathbf{T}\cdot(\mathbf{a}-(\mathbf{a}\cdot\boldsymbol{\tau})\boldsymbol{\tau})\\ &= \mu\left[1-(\mathbf{T}\cdot\boldsymbol{\tau})^2\right]\\ &\quad+(\mathbf{T}\cdot\boldsymbol{\tau})\left[f(|\boldsymbol{\sigma}-\mathbf{p}|)\frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}+f_1(\mathbf{r}\cdot\mathbf{M}_1)\mathbf{M}_1+f_2(\mathbf{r}\cdot\mathbf{M}_2)\mathbf{M}_2\right]\cdot\boldsymbol{\tau},\end{aligned} \tag{16.83}$$

and

$$\frac{d}{dt}(\mathbf{r}\cdot\mathbf{M}_1) = \dot{\mathbf{r}}\cdot\mathbf{M}_1+\mathbf{r}\cdot\dot{\mathbf{M}}_1=(\boldsymbol{\tau}-\nu\mathbf{T})\cdot\mathbf{M}_1-\nu k_1(\mathbf{r}\cdot\mathbf{T})=\boldsymbol{\tau}\cdot\mathbf{M}_1, \quad (16.84)$$

$$\frac{d}{dt}(\mathbf{r}\cdot\mathbf{M}_2) = \boldsymbol{\tau}\cdot\mathbf{M}_2, \quad (16.85)$$

$$\frac{d}{dt}\,|\boldsymbol{\sigma}-\mathbf{p}| = \frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}\cdot\boldsymbol{\tau}. \quad (16.86)$$

Furthermore,

$$\begin{aligned}\frac{d}{dt}(\mathbf{M}_1\cdot\boldsymbol{\tau}) &= \dot{\mathbf{M}}_1\cdot\boldsymbol{\tau}+\mathbf{M}_1\cdot\dot{\boldsymbol{\tau}} = -\nu k_1(\mathbf{T}\cdot\boldsymbol{\tau})+(\mathbf{a}\cdot\mathbf{M}_1)-(\mathbf{a}\cdot\boldsymbol{\tau})(\mathbf{M}_1\cdot\boldsymbol{\tau})\\ &= -\nu k_1(\mathbf{T}\cdot\boldsymbol{\tau})-f(|\boldsymbol{\sigma}-\mathbf{p}|)\frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}\cdot\mathbf{M}_1-f_1(\mathbf{r}\cdot\mathbf{M}_1)+\frac{\nu k_1}{\mathbf{T}\cdot\boldsymbol{\tau}}\\ &\quad+(\mathbf{M}_1\cdot\boldsymbol{\tau})\Big[-\mu\mathbf{T}+f(|\boldsymbol{\sigma}-\mathbf{p}|)\frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}\\ &\qquad+f_1(\mathbf{r}\cdot\mathbf{M}_1)\mathbf{M}_1+f_2(\mathbf{r}\cdot\mathbf{M}_2)\mathbf{M}_2\Big]\cdot\boldsymbol{\tau}\\ &\quad-\frac{\nu k_1(\mathbf{M}_1\cdot\boldsymbol{\tau})^2+\nu k_2(\mathbf{M}_1\cdot\boldsymbol{\tau})(\mathbf{M}_2\cdot\boldsymbol{\tau})}{\mathbf{T}\cdot\boldsymbol{\tau}}\\ &\quad+\eta(\mathbf{M}_2\cdot\boldsymbol{\tau}), \end{aligned} \quad (16.87)$$

where we have used

$$\begin{aligned}&[\mathbf{M}_1-(\mathbf{M}_1\cdot\boldsymbol{\tau})\boldsymbol{\tau}]\cdot[\mathbf{m}_2(\mathbf{T}\cdot\mathbf{m}_1)-\mathbf{m}_1(\mathbf{T}\cdot\mathbf{m}_2)]\\ &\quad=(\mathbf{M}_1\cdot\mathbf{m}_2)(\mathbf{T}\cdot\mathbf{m}_1)-(\mathbf{M}_1\cdot\mathbf{m}_1)(\mathbf{T}\cdot\mathbf{m}_2)\\ &\quad=\mathbf{M}_2\cdot\boldsymbol{\tau}. \end{aligned} \quad (16.88)$$

Substituting (16.61) for η, Eq. (16.87) becomes

$$\begin{aligned}\frac{d}{dt}(\mathbf{M}_1\cdot\boldsymbol{\tau}) &= -f(|\boldsymbol{\sigma}-\mathbf{p}|)\frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}\cdot\mathbf{M}_1-f_1(\mathbf{r}\cdot\mathbf{M}_1)\\ &\quad+(\mathbf{M}_1\cdot\boldsymbol{\tau})\Big[-\mu\mathbf{T}+f(|\boldsymbol{\sigma}-\mathbf{p}|)\frac{\boldsymbol{\sigma}-\mathbf{p}}{|\boldsymbol{\sigma}-\mathbf{p}|}\\ &\qquad+f_1(\mathbf{r}\cdot\mathbf{M}_1)\mathbf{M}_1+f_2(\mathbf{r}\cdot\mathbf{M}_2)\mathbf{M}_2\Big]\cdot\boldsymbol{\tau}. \end{aligned} \quad (16.89)$$

Similarly, we find that

$$\frac{d}{dt}(\mathbf{M}_2 \cdot \boldsymbol{\tau}) = -f(|\boldsymbol{\sigma} - \mathbf{p}|)\frac{\boldsymbol{\sigma} - \mathbf{p}}{|\boldsymbol{\sigma} - \mathbf{p}|} \cdot \mathbf{M}_2 - f_2(\mathbf{r} \cdot \mathbf{M}_2) + (\mathbf{M}_2 \cdot \boldsymbol{\tau})\Big[- \mu\mathbf{T} + f(|\boldsymbol{\sigma} - \mathbf{p}|)\frac{\boldsymbol{\sigma} - \mathbf{p}}{|\boldsymbol{\sigma} - \mathbf{p}|} + f_1(\mathbf{r} \cdot \mathbf{M}_1)\mathbf{M}_1 + f_2(\mathbf{r} \cdot \mathbf{M}_2)\mathbf{M}_2\Big] \cdot \boldsymbol{\tau}. \quad (16.90)$$

Finally, we note that $[(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{T}] \equiv 0$ implies that

$$\frac{d}{dt}[(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{T}] = 0, \quad (16.91)$$

and furthermore

$$\frac{d}{dt}[(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_1] = \boldsymbol{\tau} \cdot \mathbf{M}_1 + (\boldsymbol{\sigma} - \mathbf{p}) \cdot (-\nu k_1 \mathbf{T}) = \boldsymbol{\tau} \cdot \mathbf{M}_1, \quad (16.92)$$

$$\frac{d}{dt}[(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_2] = \boldsymbol{\tau} \cdot \mathbf{M}_2 + (\boldsymbol{\sigma} - \mathbf{p}) \cdot (-\nu k_2 \mathbf{T}) = \boldsymbol{\tau} \cdot \mathbf{M}_2. \quad (16.93)$$

Observing that

$$|\boldsymbol{\sigma} - \mathbf{p}| = \sqrt{[(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_1]^2 + [(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_2]^2} \quad (16.94)$$

and that

$$(\boldsymbol{\sigma} - \mathbf{p}) \cdot \boldsymbol{\tau} = [(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_1](\mathbf{M}_1 \cdot \boldsymbol{\tau}) + [(\boldsymbol{\sigma} - \mathbf{p}) \cdot \mathbf{M}_2](\mathbf{M}_2 \cdot \boldsymbol{\tau}), \quad (16.95)$$

we can conclude that (16.83)–(16.85), (16.89), (16.90), (16.92), and (16.93) form a self-contained, autonomous system of equations.

References

1. Bishop, R.L.: There is more than one way to frame a curve. Am. Math. Monthly **82**(3), 246–251 (1975)
2. Calini, A.: Recent developments in integrable curve dynamics. In: Geometric Approaches to Differential Equations. Lecture Notes of the Australian Math. Soc., vol. 15, pp. 56–99. Cambridge University Press, Cambridge (2000)
3. Chang, D.E., Shadden, S., Marsden, J.E., Olfati-Saber, R.: Collision avoidance for multiple agent systems. Proc. IEEE Conf. Decis. Control **1**, 539–543 (2003)
4. Fajen, B.R., Warren, W.H.: Behavioral dynamics of steering, obstacle avoidance, and route selection. J. Exp. Psychol. Hum. Percept. Perform. **29**(2), 343–362 (2003)

5. Freyman, L., Livingston, S.: Obstacle avoidance and boundary following behavior of the echolocating Bat. Final Report, MERIT BIEN Program, Univ. of Maryland (see http://scottman.net/pub/BIEN2008-Freyman-Livingston.pdf) (2008)
6. Justh, E.W., Krishnaprasad, P.S.: Equilibria and steering laws for planar formations. Syst. Control Lett. **51**, 25–38 (2004)
7. Justh, E.W., Krishnaprasad, P.S.: Natural frames and interacting particles in three dimensions. In: Proc. 44th IEEE Conf. Decision and Control, pp. 2841–2846 (see also arXiv:math.OC/0503390v1) (2005)
8. Justh, E.W., Krishnaprasad, P.S.: Steering laws for motion camouflage. Proc. R. Soc. A **462**, 3629–3643 (see also arXiv:math.OC/0508023) (2006)
9. Khalil, H.: Nonlinear Systems. Macmillan Publishing Co., New York (1992)
10. Malisoff, M., Mazenc, F., Zhang, F.: Stability and robustness analysis for curve tracking control using input-to-state stability. IEEE Trans. Autom. Control **57**(2), 1320–1326 (2012)
11. Micaelli, A., Samson, C.: Trajectory tracking for unicycle-type and two-steering-wheels mobile robots. INRIA-Sophia Antipolis Research Report No. 2097 (1993)
12. Millman, R.S., Parker, G.D.: Elements of Differential Geometry. Prentice Hall, Englewood Cliffs (1977)
13. Reddy, P.V.: Steering Laws for Pursuit. M.S. Thesis. University of Maryland (2007)
14. Reddy, P.V., Justh, E.W., Krishnaprasad, P.S.: Motion camouflage in three dimensions. In: Proc. 45th IEEE Conf. Decision and Control, pp. 3327–3332 (see also arXiv:math.OC/0603176) (2006)
15. Samson, C., Ait-Abderrahim, K.: Mobile robot control. Part 1: Feedback control of a nonholonomic wheeled cart in Cartesian space. INRIA-Sophia Antipolis Research Report No. 1288 (1990)
16. Singh, L., Stephanou, H., Wen, J.: Real-time robot motion control with circulatory fields. Proc. IEEE Int. Conf. Robot. Autom. **3**, 2737–2742 (1996)
17. Wei, E., Justh, E.W., Krishnaprasad, P.S.: Pursuit and an evolutionary game. Proc. R. Soc. A **465**, 1539–1559 (2009)
18. Wu, W., Zhang, F.: Cooperative exploration of level surfaces of three dimensional scalar fields. Automatica **47**(9), 2044–2051 (2011)
19. Zhang, F., O'Connor, A., Luebke, D., Krishnaprasad, P.S.: Experimental study of curvature-based control laws for obstacle avoidance. Proc. IEEE Int. Conf. Robot. Autom. **4**, 3849–3854 (2004)
20. Zhang, F., Justh, E.W., Krishnaprasad, P.S.: Boundary following using gyroscopic control. In: Proc. 43rd IEEE Conf. Decision and Control, pp. 5204–5209 (2004)

Chapter 17
Random Hill's Equations, Random Walks, and Products of Random Matrices

Fred C. Adams, Anthony M. Bloch, and Jeffrey C. Lagarias

Dedicated to Jürgen Scheurle on the occasion of his 60th birthday

Abstract Hill's equations arise in a wide variety of physical problems, and are specified by a natural frequency, a periodic forcing function, and a forcing strength parameter. This classic problem can be generalized by allowing the forcing strength q_k, the frequency λ_k, and the period $(\Delta\tau)_k$ of the forcing function to vary from cycle to cycle. The growth rates for the solutions are then given by the growth rates of a matrix transformation, under matrix multiplication, where the elements vary from cycle to cycle. Simplified models of such problems are given by products of 2×2 random matrices drawn from a given class.

This paper analyzes two simple classes of models of 2×2 random matrices where the growth rates (Lyapunov exponents) can be computed in an explicit form. Both models are special cases of random products involving random similarity transformations. The first of these corresponds to the random Hill's equation in a regime where the solutions are highly unstable. This model is a product of random similarity transformations of a fixed singular matrix. The second class of models is a two parameter class that studies products of 2×2 random symmetric matrices of a special form which are conjugated by random orthogonal similarities. These matrices are nonsingular in general, but in a special case they give rank one matrices,

F.C. Adams
Department of Physics, University of Michigan, Ann Arbor, MI, USA
e-mail: fca@umich.edu

A.M. Bloch (✉) · J.C. Lagarias
Department of Mathematics, University of Michigan, Ann Arbor, MI, USA
e-mail: abloch@.umich.edu; lagarias@umich.edu

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_17,

which may be compared with the first model. In the latter case the two models have different growth rate behavior, which arises from the different nature of the allowed similarity transformations in the models.

17.1 Introduction

Random matrices arise in a wide variety of applications. This work is motivated by consideration of Hill's equation, a second order periodic differential equation that describes many physical problems [23]. It takes the form

$$\frac{d^2y}{dt^2} + [\lambda + q\hat{Q}(t)]y = 0\,, \tag{17.1}$$

in which $\hat{Q}(t)$ is periodic with period π and λ and q are constants, where we have imposed the normalization $\int_0^\pi \hat{Q}(t)dt = 1$. In addition to its original application to lunar orbits [15], Hill's equation describes celestial dynamics [6, 7, 22], orbit instabilities in dark matter halos [5], particle production at the end of the inflationary epoch in the early universe [14, 20], the motion of a jogger's ponytail [18], and many other physical systems. This problem can be generalized so that its parameters vary from cycle to cycle, i.e., the differential equation takes the form

$$\frac{d^2y}{dt^2} + [\lambda_k + q_k\hat{Q}(t)]y = 0, \qquad (k-1)\pi \leq t < k\pi, \tag{17.2}$$

where the parameters (λ_k, q_k) vary from cycle to cycle and are drawn independently from well-defined distributions. The index k labels the cycle. In general, the period can vary as well; however, one can show that cycle to cycle variations in $\Delta\tau$ can be scaled out of the problem and included in the variations of the (λ_k, q_k) by changing their distributions accordingly (see Theorem 1 of [2]).

The initial motivation for considering random Hill's equations arose in studies of orbit problems in astrophysics. Briefly, when an orbit starts in the principal plane of a triaxial, extended mass distribution (such as a dark matter halo), the motion is unstable to perturbations in the perpendicular direction (out of the plane). The development of the instability is described by a random Hill's equation with the form given by Eq. (17.2).

Periodic differential equations in this class can be described by a discrete mapping of the coefficients of the principal solutions from one cycle to the next. In the special case where the periodic functions $\hat{Q}(t)$ are symmetric about the midpoint of the period, a condition we assume in this paper, the transformation matrix takes the form

$$\mathbb{M}_k = \begin{bmatrix} h_k & (h_k^2 - 1)/g_k \\ g_k & h_k \end{bmatrix}\,, \tag{17.3}$$

where the subscript denotes the cycle. The matrix elements for the kth cycle are given by

$$h_k = y_1(k\pi) \qquad \text{and} \qquad g_k = \dot{y}_1(k\pi)\,, \tag{17.4}$$

where y_1 and y_2 are the principal solutions for that cycle. That is,

$$y_1((k-1)\pi) = \dot{y}_2((k-1)\pi) = 1, \qquad \dot{y}_1((k-1)\pi) = y_2((k-1)\pi) = 0.$$

The index k indicates that the quantities (λ_k, q_k), and hence the solutions (h_k, g_k), vary from cycle to cycle. We define the product

$$\mathbb{M}^{(N)} \equiv \mathbb{M}_N \mathbb{M}_{N-1} \cdots \mathbb{M}_2 \mathbb{M}_1\,. \tag{17.5}$$

Note that the matrix in Eq. (17.3) has only two independent elements (g_k, h_k) (not four). Since the Wronskian of the original differential equation (17.1) is unity, the determinant of the matrix map (17.3) must also be unity, and this constraint eliminates one independent element. In addition, this paper specializes to the case where the periodic functions $\hat{Q}(t)$ are symmetric about the midpoint of the period. This property implies that $y_1(\pi) = \dot{y}_2(\pi) = h_k$ for the kth cycle where y_1 and y_2 are the principle solutions for that cycle, which eliminates a second independent matrix element [1, 23]. These two constraints imply that $y_2(k\pi) = (h_k^2 - 1)/g_k$ for the given cycle, resulting in the form for the matrix given by Eq. (17.3).

The growth rates for Hill's equation (17.1) are determined by the growth rates for matrix multiplication of the matrices $\mathbb{M}_k$ given by Eq. (17.3). Here we denote the product of N such matrices as $\mathbb{M}^{(N)}$, and the growth rate γ is defined by

$$\gamma = \lim_{N\to\infty} \frac{1}{N} \log ||\mathbb{M}^{(N)}||\,. \tag{17.6}$$

provided the limit exists. If it exists it is independent of the choice of the norm $||\cdot||$ (see e.g. [32]).

For a wide range of general random matrix models the limit (17.6) is known to exist almost surely, as shown in previous work [13, Theorem 2], [12, 21]. In such models growth rate formulas like (17.6) give the largest *Lyapunov exponent* of the random matrix product , as considered in [21, 26, 27]. Here the Lyapunov exponents $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_d$ for $d \times d$ random matrix products are determined by

$$\gamma_j := \lim_{N\to\infty} \frac{1}{2N} \log \left(\text{j-th largest eigenvalue of } (\mathbb{M}^{(N)})^T \mathbb{M}^{(N)}\right) \tag{17.7}$$

for $j = 1, 2$. These limits exist almost surely under wide circumstances, and it is also known that $\gamma_1 = \gamma$ as given in (17.6), by previous work, see [26].There are limited families of general examples where they are known explicitly, see [8, 9, 16, 21, 25–27, 30], and for various 2×2 matrix models,

see [10, 11, 19, 21, 24, 29]. In general Lyapunov exponents of general random matrix models are hard to approximate and sometimes impossible to compute [32].

In this paper we study two models of random matrix products of 2×2 matrices. The first model, as discussed above, was suggested by Hill's equation in the unstable regime and was previously considered in the first two authors' paper [2]. The second class of models is introduced here and is presented as a contrasting model, to shed light on features of the models that affect the size of the Lyapunov exponents. A useful property of these models is that the growth rate (top Lyapunov exponent) can be determined in closed form in many cases. These give some new cases of exact formulas for Lyapunov exponents.

The structure of the paper is as follows.

In Sect. 17.2 we present a random matrix model for solutions of the Hill's equation in a regime where the solutions are highly unstable. It is a product of random matrices of the form

$$\mathbb{A}_k \equiv \begin{bmatrix} 1 & x_k \\ 1/x_k & 1 \end{bmatrix}$$

where the x_k are independent identically distributed elements drawn from a given probability distribution on the real line. This model can be reformulated as studying products of random similarity transformations of a fixed matrix, with allowed similarities of the form

$$\mathbb{P}_k \equiv \begin{bmatrix} -x_k & x_k \\ 1 & 1 \end{bmatrix}.$$

We obtain growth rate formulas valid for a large class of such distributions, and determine both Lyapunov exponents γ_1 and γ_2 for these models.

In Sect. 17.3 we present and study a class of 2×2 random matrices, with different dynamics, which are also products of random similarity transformed matrices, where the allowed similarities are rotations. This model has two parameters at each step, a parameter x_k describing a matrix and α_k describing the rotation angle in the similarity transformation. The matrices have the form

$$\mathbb{C}_k = Q_k \begin{bmatrix} 1 & x_k \\ x_k & 1 \end{bmatrix} Q_k^{-1}, \tag{17.8}$$

where Q_k are orthogonal matrices, so $Q^T = Q^{-1}$, encoding random rotations by angles α_k. We obtain explicit formulas for the growth rates for many distributions of x_k and α_k. This model is comparable to Sect. 17.2 models in the special case that all $x_k = 1$ are nonrandom.

In Sect. 17.4 we compare Lyapunov exponents for unstable Hill equations models of Sect. 17.2, with the special subclass of the random rotation models of Sect. 17.3 which occurs when all variables $x_k = x = 1$. In this case both models consist

of random similarity transformations of the fixed rank one matrix $\left[\begin{smallmatrix}0 & 0\\ 0 & 2\end{smallmatrix}\right]$. The unperturbed model has growth rate $\gamma_1 = \log 2$. We observe that the perturbations in the two models have opposite effects on the growth rate. The models of Sect. 17.2 for nonnegative distributions x_k increase the growth rate, while the models of Sect. 17.3 decrease it.

17.2 Matrices from Hill's Equation in the Unstable Regime

This section considers matrices arising from Hill's equation; these models were developed in previous work of the first two authors [2–4]. The discrete map for the general problem can be rewritten in the form

$$\mathbb{M}_k = h_k \begin{bmatrix} 1 & x_k \\ 1/x_k & 1 \end{bmatrix} + \begin{bmatrix} 0 & -1/g_k \\ 0 & 0 \end{bmatrix} \approx h_k \begin{bmatrix} 1 & x_k \\ 1/x_k & 1 \end{bmatrix} , \tag{17.9}$$

where we have defined $x_k \equiv h_k/g_k$. The second equality corresponds to the regime where $h_k \gg 1$. The growth due to the leading factors of h_k can be found using standard methods. We thus need to find the growth rate for products of random matrices of the form

$$\mathbb{A}_k \equiv \begin{bmatrix} 1 & x_k \\ 1/x_k & 1 \end{bmatrix} . \tag{17.10}$$

The matrix (17.10) is a singular matrix with eigenvalues 0 and 2 (independent of x_k). The corresponding matrix of eigenvectors can be written as

$$\mathbb{P}_k \equiv \begin{bmatrix} -x_k & x_k \\ 1 & 1 \end{bmatrix} . \tag{17.11}$$

and the matrix $\mathbb{A}_k$ may be written

$$\mathbb{A}_k = \mathbb{P}_k \mathbb{D} \mathbb{P}_k^{-1} \qquad \text{where} \qquad \mathbb{D} \equiv \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} , \tag{17.12}$$

so that this problem is equivalent to analyzing products of a randomly conjugated fixed (singular) matrix.

The problem of finding growth rates for Hill's equation in the regime $h_k \gg 1$ thus motivates study of the problem of finding growth rates for matrices $\mathbb{A}_k$ of the form given by Eq. (17.10). These matrices are specified by only one random variable x_k. For applications to Hill's equation, $x_k = h_k/g_k$, and hence the values of x_k depend on the parameters (q_k, λ_k) through the original differential equation.

We now consider the random matrix model in which the variables x_k are modeled by independent identically distributed variables on the real line, from some specified distribution. Note that this condition would hold for $x_k = h_k/g_k$, if the g_k (resp. the h_j) were independent and identically distributed, possibly with different distributions, and the g_k were independent of the h_j's. Under this assumption on the distribution of x_k, we can calculate growth rates as shown below. In this case $\gamma_1 = \gamma$ as given by (17.6) and furthermore, since all matrices in the product are singular, we will have $\gamma_2 = -\infty$.

We note that the large h_k assumption that we make here leads to a singular model that has an analysis given below. The more general case can also be handled, in the sense that one can give recursions for the growth rates of finite matrix products that will converge to the growth rate, see Theorem 1 of the first two authors' paper [4].

17.2.1 Growth Rates for Positive Matrix Elements

We first consider the case where the ratios x_k that define the discrete map $\mathbb{A}_k$ all have the same sign.

Theorem 17.1. *Consider matrices of the form* $\mathbb{A}_k \equiv \begin{bmatrix} 1 & x_k \\ 1/x_k & 1 \end{bmatrix}$ *with the* x_k *drawn independently and identically distributed and with the distribution supported on the positive real line, satisfying*

$$\mathbb{E}_x[\log(1+x)] < \infty \qquad \textit{and} \quad \mathbb{E}_x[\log(1+\frac{1}{x})] < \infty\,. \tag{17.13}$$

Then with probability one the growth rate γ_1 *exists and is given by*

$$\gamma_1 = \mathbb{E}_{x_1,x_2}[\log(1+\frac{x_1}{x_2})]\,. \tag{17.14}$$

where the expectation is taken over two independent draws x_1, x_2 *from the distribution of* x*. Here* $\gamma_2 = -\infty$*.*

Proof. The hypothesis (17.13) can be shown to be equivalent to $\mathbb{E}_x[\log^+ ||\mathbb{A}_1||] < \infty$, where $||\mathbb{A}|| = \max_i \sum_j |A_{i,j}|$. Now Theorem 2 of [13] implies the existence almost everywhere of a limit γ_1. The value of γ_2 follows from the matrices being of rank one.

To obtain a formula for γ_1, after N cycles, one can show (by an induction argument) that the product matrix $\mathbb{A}^{(N)}$ takes the form

$$\mathbb{A}^{(N)} = \prod_{k=1}^{N} \mathbb{A}_k = \begin{bmatrix} \Sigma_T^{(N)} & x_1 \Sigma_T^{(N)} \\ \Sigma_B^{(N)}/x_1 & \Sigma_B^{(N)} \end{bmatrix}\,, \tag{17.15}$$

where x_1 is the value of the variable for the first cycle and where the sums $\Sigma_T^{(N)}$ and $\Sigma_B^{(N)}$ are given by

$$\Sigma_T^{(N)} = \sum_{j=1}^{2^{N-1}} r_j \qquad \text{and} \qquad \Sigma_B^{(N)} = \sum_{j=1}^{2^{N-1}} \frac{1}{r_j} ,$$

where the variables r_j are ratios of the form

$$r_j = \frac{x_{a_1} x_{a_2} \dots x_{a_n}}{x_{b_1} x_{b_2} \dots x_{b_n}} . \tag{17.16}$$

The ratios r_j arise from repeated multiplication of the matrices $\mathbb{A}_k$, and hence the indices lie in the range $1 \le a_i, b_i \le N$. The r_j always have the same number of factors in the numerator and the denominator, but the number of factors (n) varies from 0 (where $r_j = 1$) up to $N/2$. This upper limit arises because each composite ratio r_j has $2n$ values of x_j, which must all be different, and because the total number of possible values is N.

Carrying the induction argument one step further, one finds that from one cycle to the next the sums $\Sigma_T^{(N)}$ and $\Sigma_B^{(N)}$ vary according to

$$\Sigma_T^{(N+1)} = \Sigma_T^{(N)} + \frac{x}{x_1} \Sigma_B^{(N)} ,$$

and

$$\Sigma_B^{(N+1)} = \Sigma_B^{(N)} + \frac{x_1}{x} \Sigma_T^{(N)} .$$

In this notation, the variable x (no subscript) represents the value of the x variable at the current cycle, whereas x_1 represents the value at the initial cycle. Next we note that the ratio of these two factors reduces to the simple form

$$\frac{\Sigma_T^{(N+1)}}{\Sigma_B^{(N+1)}} = \frac{\Sigma_T^{(N)} + (x/x_1)\Sigma_B^{(N)}}{\Sigma_B^{(N)} + (x_1/x)\Sigma_T^{(N)}} = \frac{x}{x_1} . \tag{17.17}$$

The growing eigenvalue of the product matrix of Eq. (17.15) is given by

$$\Lambda = \Sigma_T^{(N)} + \Sigma_B^{(N)} . \tag{17.18}$$

As a result, the eigenvalue (growth factor) varies from cycle to cycle according to

$$\Lambda^{(N+1)} = \Lambda^{(N)} + \frac{x}{x_1} \Sigma_B^{(N)} + \frac{x_1}{x} \Sigma_T^{(N)} = \Lambda^{(N)} \left[1 + \frac{(x/x_1)\Sigma_B^{(N)} + (x_1/x)\Sigma_T^{(N)}}{\Sigma_B^{(N)} + \Sigma_T^{(N)}} \right] .$$

The overall growth factor is then determined by the product

$$\Lambda^{(N)} = \prod_{j=1}^{N} \left[1 + \frac{(x/x_1)\Sigma_B^{(j)} + (x_1/x)\Sigma_T^{(j)}}{\Sigma_B^{(j)} + \Sigma_T^{(j)}} \right] .$$

Using Eq. (17.17) to eliminate $\Sigma_T^{(j)}$ and $\Sigma_B^{(j)}$, this factor can be rewritten in the form

$$\Lambda^{(N)} = \prod_{j=1}^{N} \left[\left(\frac{x_1 + x}{x_1 + x_j} \right) \left(1 + \frac{x_j}{x} \right) \right] = \prod_{j=1}^{N} \left(\frac{x_1 + x_{j1}}{x_1 + x_{j2}} \right) \prod_{j=1}^{N} \left(1 + \frac{x_{j2}}{x_{j1}} \right) . \tag{17.19}$$

In the second equality, we have replaced the random elements x and x_j (which are two elements chosen in succession) with x_{j1} and x_{j2} (which are two elements chosen independently). Equality holds because the matrices, and matrix elements, are independent and identically distributed. The growth rate γ for matrix multiplication is determined by setting the above product equal to $\exp[N\gamma]$, which implies that

$$\gamma = \lim_{N\to\infty} \frac{1}{N} \left\{ \log \left[\prod_{j=1}^{N} \left(\frac{x_1 + x_{j1}}{x_1 + x_{j2}} \right) \right] + \log \left[\prod_{j=1}^{N} \left(1 + \frac{x_{j2}}{x_{j1}} \right) \right] \right\} ,$$

which can be rewritten in the form

$$\gamma = \lim_{N\to\infty} \left\{ \frac{1}{N} \sum_{j=1}^{N} \log \left[x_1 + x_{j1} \right] - \frac{1}{N} \sum_{j=1}^{N} \log \left[x_1 + x_{j2} \right] + \frac{1}{N} \sum_{j=1}^{N} \log \left[1 + \frac{x_{j2}}{x_{j1}} \right] \right\} . \tag{17.20}$$

Since the x_{j1} and x_{j2} are chosen from the same distribution, under the assumption $\mathbb{E}_x(\log(1+x)] < \infty$ and $\mathbb{E}_x(\log(1+\frac{1}{x})] < \infty$, the first two terms will cancel in the limit (with probability one), and the growth rate reduces to the form

$$\gamma = \lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{N} \log \left[1 + \frac{x_{j1}}{x_{j2}} \right] , \tag{17.21}$$

where x_{j1} and x_{j2} are any two independent realizations of the x_k. As a result, we obtain with probability one, the limit exists and is given by the expected value $\mathbb{E}_{x_1,x_2}[\log(1 + \frac{x_1}{x_2})]$. □

17.2.2 Matrix Elements with Varying Signs

We now generalize to consider the case in which the signs of the variables x_k can be either positive or negative.

Theorem 17.2. *Consider matrices of the form* $\mathbb{A}_k \equiv \begin{bmatrix} 1 & x_k \\ 1/x_k & 1 \end{bmatrix}$ *where the* x_k *can be either positive or negative. Suppose that the values* x_k *are independently and identically drawn from a distribution satisfying*

$$\mathbb{E}_x[\log(1+|x|)] < \infty \qquad \textit{and} \quad \mathbb{E}_x[\log(1+\frac{1}{|x|})] < \infty. \tag{17.22}$$

Let positive signs occur with probability p *and negative signs occur with probability* $1-p$*. Then, with probability one, the growth rate exists and is given by*

$$\gamma_1 = \left[p^2+(1-p)^2\right]\mathbb{E}_{x_1,x_2}\left[\log\left(1+\left|\frac{x_1}{x_2}\right|\right)\right] + 2p(1-p)\,\mathbb{E}_{x_1,x_2}\left[\log\left|1-\left|\frac{x_1}{x_2}\right|\right|\right]. \tag{17.23}$$

where x_1, x_2 *represent two independent draws from the distribution of* x_k*. Here* $\gamma_2 = -\infty$*.*

Proof. The condition (17.22) guarantees existence of γ_1 similarly to Theorem 17.1. Also, the same arguments leading to Eq. (17.19) in the proof of Theorem 17.1 can be used, where the signs of the ratios x_{j1}/x_{j2} must be taken into account. If p is the probability of the x_j variables being positive, the probability of the ratio of two variables being positive will be given by $p^2+(1-p)^2$, i.e., the probability of getting either two positive signs or two negative signs. The probability of the ratio being negative is then $2p(1-p)$. With this consideration of signs, the intermediate form of Eq. (17.19) is modified to take the form

$$\Lambda^{(N)} = \prod_{j=1}^{N_P}\left[\left(\frac{x_1+x_{j1}}{x_1+x_{j2}}\right)\left(1+\left|\frac{x_{j2}}{x_{j1}}\right|\right)\right]\prod_{k=1}^{N_Q}\left[\left(\frac{x_1+x_{k1}}{x_1+x_{k2}}\right)\left(1-\left|\frac{x_{k2}}{x_{k1}}\right|\right)\right], \tag{17.24}$$

where N_P is the number of terms where the ratios have positive signs and N_Q is the number of terms where the ratios have negative signs ($N_P+N_Q=N$). To find the growth rate, we set this product equal to $\exp[N\gamma]$ and take the limit $N\to\infty$, which results in six terms, i.e.,

$$\gamma = \lim_{N\to\infty}\frac{1}{N}\left\{\sum_{j=1}^{N_P}\log\left(1+\left|\frac{x_{j2}}{x_{j1}}\right|\right)+\sum_{k=1}^{N_Q}\log\left|1-\left|\frac{x_{k2}}{x_{k1}}\right|\right|\right\}$$

$$+ \lim_{N\to\infty} \frac{1}{N} \left\{ \sum_{j=1}^{N_P} \log|x_1 + x_{j1}| - \sum_{j=1}^{N_P} \log|x_1 + x_{j2}| \right\} \tag{17.25}$$

$$+ \lim_{N\to\infty} \frac{1}{N} \left\{ \sum_{k=1}^{N_Q} \log|x_1 + x_{k1}| - \sum_{k=1}^{N_Q} \log|x_1 + x_{k2}| \right\} .$$

In the limit $N \to \infty$, under the assumption $\mathbb{E}_x[\log(1 + |x|)] < \infty$ and $\mathbb{E}_x[\log(1 + \frac{1}{|x|})] < \infty$, with probability one the second and third limits will be 0, and the first term will have a limit. Then the growth rate is given by only the first two terms in Eq. (17.25). By definition, $N_P/N = p^2 + (1-p)^2$ and $N_Q/N = 2p(1-p)$, so the two remaining sums can be converted to sums to N by including these weight factors. As result, we obtain

$$\gamma = \lim_{N\to\infty} \frac{1}{N} \left\{ [p^2 + (1-p)^2] \sum_{j=1}^{N} \log\left(1 + \left|\frac{x_{j1}}{x_{j2}}\right|\right) + 2p(1-p) \sum_{k=1}^{N} \log\left|1 - \left|\frac{x_{k1}}{x_{k2}}\right|\right| \right\} . \tag{17.26}$$

with existence of the limit with probability one. The limit is then given by Eq. (17.23). □

17.3 Products of Randomly Rotated Matrices

In this section we study a different two-parameter random matrix model with parameters x_k and α_k, which can be viewed as modeling the time 1 solutions of a different sort of randomly perturbed periodic differential equation, definitely not of Hill's equation type. The model considers products of 2×2 random matrices of the form

$$\mathbb{C}_k = Q_f \mathbb{X}_k Q_k^T := Q_k \begin{bmatrix} 1 & x_k \\ x_k & 1 \end{bmatrix} Q_k^T , \tag{17.27}$$

where the real entry x_k may be random and where

$$Q_k = Q(\alpha_k) = \begin{bmatrix} \cos\alpha_k & \sin\alpha_k \\ -\sin\alpha_k & \cos\alpha_k \end{bmatrix} \tag{17.28}$$

are rotation matrices with a possibly random angle α_k. In this product are two parameters x_k and α_k. In the special case where all $x_k \equiv x$ are constant, with

$x = 1$, the matrix

$$\mathbb{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = Q \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} Q^T, \quad Q = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix},$$

where Q is a rotation with angle $\alpha = \pi/4$. This gives a model that has the same form as (17.12) in the previous model, as similarities $\mathbb{P}_k = Q_k Q$. Thus, restricting to the case $x_k = 1$, we have a one-parameter model whose freedom is to choose a distribution of angles in the random rotations. All such models have a similar form to the Unstable Hill's equation model in Sect. 17.2, which differ only in the allowed form for the similarity transformations. Thus one can compare the effects on the growth rates of making random rotations versus that of the similarity transformations $\mathbb{P}_k$ [given by (17.11) and (17.12)] considered in Sect. 17.2. We make such a comparison in Sect. 17.4.

Here we will study this model and show that one can obtain explicit formulas for the Lyapunov exponents in various cases. Regarding the variables x_k, one can check that matrices of the form

$$\mathbb{B}_k = \begin{bmatrix} 1 & x_k \\ x_k & 1 \end{bmatrix}, \tag{17.29}$$

pairwise commute. This holds since

$$\begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix} \begin{bmatrix} 1 & y \\ y & 1 \end{bmatrix} = \begin{bmatrix} 1+xy & x+y \\ x+y & 1+xy \end{bmatrix} \tag{17.30}$$

where the right side is manifestly invariant under interchange of x and y. Thus the model with no rotations is simple to analyze, and we do this in the Appendix. The contribution of random rotations to the model makes the products of the $\mathbb{C}_k$ noncommutative in general.

Since we are considering probability models with independent draws of $\mathbb{C}_k$ at each step, the random matrix product

$$\mathbb{C}^{(N)} := \mathbb{C}_N \mathbb{C}_{N-1} \cdots \mathbb{C}_2 \mathbb{C}_1. \tag{17.31}$$

represents a random walk on the semigroup $M_{2\times 2}(\mathbf{R})$ of real 2×2 matrices under matrix multiplication, using the terminology of [17, Chapter 3]. Geometrically it can be viewed as a (multiplicative) random walk on the plane, acting on a fixed initial (column) vector. The symmetric matrix (17.29) acts on a vector in $\mathbb{R}^2$ by reflecting it around the line $x = y$, then scaling it by a random factor x_k, and adding it to itself. The matrices $\mathbb{C}_k$ add a preliminary rotation through an angle $-\alpha_k$, then application of (17.29), then rotation back by α_k. We then repeat.

17.3.1 Deterministic Formulas for Product Matrices

It is possible to obtain an explicit expression for the product of N matrices of the form above. First write out the form of the matrix

$$\mathbb{C}_k = \begin{bmatrix} 1 - x_k \sin 2\alpha_k & x_k \cos 2\alpha_k \\ x_k \cos 2\alpha_k & 1 + x_k \sin 2\alpha_k \end{bmatrix} . \tag{17.32}$$

The eigenvalues of $\mathbb{C}_k$ are given by

$$\lambda_\pm = (1 \pm x_k) .$$

The corresponding eigenvectors are given by

$$\mathbf{v}_k^+ = \frac{1}{\sqrt{2}} \begin{bmatrix} \cos\alpha_k - \sin\alpha_k \\ \cos\alpha_k + \sin\alpha_k \end{bmatrix} \qquad \text{and} \qquad \mathbf{v}_k^- = \frac{1}{\sqrt{2}} \begin{bmatrix} -\cos\alpha_k - \sin\alpha_k \\ \cos\alpha_k - \sin\alpha_k \end{bmatrix} , \tag{17.33}$$

where the vectors have been normalized. Now suppose we have a arbitrary vector $\mathbf{a}_k$ of unit length, written in the form

$$\mathbf{a_k} = \begin{bmatrix} \cos\phi_k \\ \sin\phi_k \end{bmatrix} .$$

The vector can also be written in terms of the eigenvectors

$$\mathbf{a_k} = C_+ \mathbf{v}_k^+ + C_- \mathbf{v}_k^- . \tag{17.34}$$

Letting the angle

$$\theta_k := \phi_k - \alpha_k . \tag{17.35}$$

we find that the coefficients are given by

$$C_+ = \frac{1}{\sqrt{2}} (\cos\theta_k + \sin\theta_k) \qquad \text{and} \qquad C_- = \frac{1}{\sqrt{2}} (-\cos\theta_k + \sin\theta_k) ,$$

If we multiply the vector $\mathbf{a}$ by the matrix $\mathbb{C}_k$ we obtain

$$L_{k+1}\mathbf{a}_{k+1} := \mathbb{C}_k \mathbf{a_k} = (1 + x_k) C_+ \mathbf{v}_k^+ + (1 - x_k) C_- \mathbf{v}_k^- ,$$

which has length L_{k+1} given by

$$L_{k+1}^2 = (1 + x_k)^2 C_+^2 + (1 - x_k)^2 C_-^2 = 1 + x_k^2 + 2x_k \sin 2\theta_k .$$

and new unit vector $\mathbf{a}_{k+1} = \begin{bmatrix} \cos\phi_{k+1} \\ \sin\phi_{k+1} \end{bmatrix}$ with angle ϕ_{k+1} determined by ϕ_k, α_k and x_k. We obtain the vector norm

$$||\mathbb{C}^{(N)}(\mathbf{a}_0)||^2 = \prod_{k=1}^{N}(1 + x_k^2 + 2x_k \sin 2\theta_k). \tag{17.36}$$

This expression is deterministic.

We can use this result to analyze a model of random matrix products.

Theorem 17.3. *Consider random matrices $\mathbb{C}_k$ of the form $\mathbb{C}_k = Q_k \begin{bmatrix} 1 & x_k \\ x_k & 1 \end{bmatrix} Q_k^T$ where the x_k are real. Assume that the x_k are drawn independently and identically from a distribution having*

$$\mathbb{E}_x[\log(1 + |x|)] < \infty,$$

and that the α_k are independently and identically distributed angles in $[0, 2\pi]$, and that the α_k are drawn independent of the x_j. Then the corresponding growth rate γ for matrix multiplication exists with probability one and can be written in the form

$$\gamma_1 = \lim_{N\to\infty} \frac{1}{2N} \sum_{k=1}^{N} \log\left[1 + x_k^2 + 2x_k \sin 2\theta_k\right] . \tag{17.37}$$

where $\theta_k = \phi_k - \alpha_k$, where the α_k are the rotation angles, and where the ϕ_k determine the direction of the two-dimensional vector at the kth step of the iteration, for a given ϕ_0. In addition, the second Lyapunov exponent is given by

$$\gamma_2 = -\gamma_1 + \lim_{N\to\infty} \sum_{k=1}^{N} \log|1 - x_k^2|, \tag{17.38}$$

where with probability one the second limit exists, on the x_k-distribution.

Proof. From the definition, the limit $\gamma = \gamma_1$ exists with probability one and is constant, by the result of Furstenberg and Kesten [13], Theorem 2.

Now the formula (17.37) follows by inserting the product norm formula (17.36) after N steps. To obtain the second formula, we observe that we have , independently of the choice of rotations,

$$\det((\mathbb{C}^{(N)})^T \mathbb{C}^{(N)}) = \prod_{k=1}^{N} d_k^2 \equiv \prod_{k=1}^{N} (1 - x_k^2)^2. \tag{17.39}$$

This yields

$$\gamma_1 + \gamma_2 = \lim_{N\to\infty} \sum_{k=1}^{N} \log|1 - x_k^2|, \tag{17.40}$$

yielding the result. □

Remark. An feature of this result is that the existence of the limit (17.37) is not apparent from its general form, due to lack of direct information on the distribution of θ_k. Instead we infer the existence of the limit from general results on random matrix products, which only then implies that the distribution of the θ_k must be such that the limit exists with probability one.

17.3.2 Uniformly Distributed Rotations Case

We now treat the special case where the angles α_k are uniformly distributed. We first present a lemma concerning random angles:

Lemma 17.1. *Let x be a random variable that is uniformly distributed on $[0, 1]$, and let y be a random variable on $[0, 1]$ with an arbitrary distribution with cumulative distribution f_y. Consider the composite random variable $z = x + y$ (mod 1). Then the distribution of z is uniform on $[0, 1]$.*

Proof. We construct the cumulative distribution function of z, denoted here as $P_z(t) := Prob[z : z \leq t]$ for the variable z. The distribution of interest is then given by $p = dP_z(t)/dt$. To prove the result in question, we must show that $p=$ constant.

The cumulative probability is given by integrating over the portion of the unit square where $x + y < z$ (mod 1). The integral has three parts, which take the form

$$P_z(z) = \int_0^z f_y dy \int_0^{z-y} dx + \int_0^z f_y dy \int_{1-y}^1 dx + \int_z^1 f_y dy \int_{1-y}^{1+z-y} dx\,, \tag{17.41}$$

where we integrate over the variable x first.

Because the distribution of x is uniform, the x-integrals can be evaluated to find

$$P(z) = \int_0^z f_y dy(z-y) + \int_0^z f_y dy(y) + \int_z^1 f_y dy(z) = z \int_0^1 f_y dy = z\,. \tag{17.42}$$

Thus, $P_z(z) = z$, $p = dP_z(t)/dt = 1$, and hence the distribution of the random variable Z z is uniform on $[0, 1]$, as claimed. □

Using the lemma we obtain a formula for the Lyapunov exponents in the uniformly distributed case. The answers are expressed using Stieltjes integrals as in [31, 33].

Theorem 17.4. *Consider random matrices* $\mathbb{C}_k$ *of the form* $\mathbb{C}_k = Q_k \begin{bmatrix} 1 & x_k \\ x_k & 1 \end{bmatrix} Q_k^T$, *where the* x_k *are real. Suppose that the* x_k *are independently and identically distributed, with distribution which satisfies*

$$E_x[\log(1 + |x|)] < \infty.$$

Suppose that the rotation angles α_k *are independent of all the* x_k *and are uniformly distributed. Then the top Lyapunov exponent is given by the Stieltjes integral*

$$\gamma_1 = \int_{(|x|>1)} \log|x| \, dP_x(x) \,, \tag{17.43}$$

where the integral is taken only over the region where $|x| > 1$. *The second Lyapunov exponent is given by*

$$\gamma_2 = -\gamma_1 + \int_{-\infty}^{\infty} \log|1 - x^2| dP_x(x). \tag{17.44}$$

The value $\gamma_2 = -\infty$ *is permitted.*

Proof. An immediate consequence of Lemma 17.1 is: *Let the angle* α *be uniformly distributed on* $[0, 2\pi]$ *and let the angle* ϕ *have an arbitrary distribution. Then the composite variable* $\theta = \phi - \alpha$ *is uniformly distributed on* $[0, 2\pi]$.

This fact together with our hypotheses allow us to directly use the formula (17.37) for γ given in Theorem 17.3. Under the assumption that the angles α_k are independent and uniformly distributed over $[0, 2\pi]$, with no assumption on the x_k, the angles θ_k defined by Eq. (17.35) are also independent of each other and are uniformly distributed. If we furthermore assume that the distribution of each α_k is independent of all x_j with $j \le k$, then it follows that each variable θ_k is independent of variable x_k (they depend on α_j with $j \le k$ and on x_j with $j < k$). In that case the expression (17.37) for the growth rate converges with probability one to the Stieltjes integral

$$\gamma_1 := \frac{1}{2} \int_{-\infty}^{\infty} dP_x \Big(\frac{1}{2\pi} \int_0^{2\pi} \log\left[1 + x^2 + 2x \, \sin 2\theta\right] d\theta \Big) \,. \tag{17.45}$$

Since the α_k are uniformly distributed, the angular integral can be evaluated, using the formula

$$\frac{1}{2\pi} \int_0^{2\pi} \log\left[1 + (2\sin 2\theta)x + x^2\right] d\theta = \begin{cases} \log|x| & \text{if } \ |x| > 1, \\ 0 & \text{if } \ |x| \le 1. \end{cases}$$

The growth rate becomes the Stieltjes integral

$$\gamma_1 = \int_{|x|\geq 1} \log|x| dP_x(x), \tag{17.46}$$

The expression for the second Lyapunov exponent then follows using the fact that with probability one the integral in (17.38) converges to $\int_{-\infty}^{\infty} \log|1-x^2| dP_x(x)$. □

Observe that for an absolutely continuous density the formula (17.46) becomes

$$\gamma_1 = \int_{(|x|>1)} \frac{dP_x}{dx} \log|x| dx. \tag{17.47}$$

17.3.3 Uniformly Distributed Case with Constant x_k

We consider the special case where the values $x_k = x$ are constant. A complete analysis can be given for uniformly distributed rotations.

Theorem 17.5. *Suppose that all $x_k = x$ are real and constant, with $|x| \neq 1$, and that the rotation angles α_k are drawn independently from the uniform distribution on $[0, 2\pi]$. Then the Lyapunov exponents of the product of random $\mathbb{C}_k := Q_k \begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix} Q_k^T$ take the form, for $|x| > 1$,*

$$\gamma_1 = \log(|x|) \qquad \text{and} \qquad \gamma_2 = \log(|x| - \frac{1}{|x|}) \quad (|x| > 1). \tag{17.48}$$

The Lyapunov exponents take the form, for $0 \leq |x| < 1$,

$$\gamma_1 = 0 \qquad \text{and} \qquad \gamma_2 = \log(1-x^2) \qquad (|x| < 1)\,. \tag{17.49}$$

The Lyapunov exponents for $|x| = 1$ take the form,

$$\gamma_1 = 0 \qquad \text{and} \qquad \gamma_2 = -\infty\,. \tag{17.50}$$

Remark. This result may be compared with the purely deterministic model with no rotations present. This case considers products of the matrices $\begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix}$, and has $\gamma_1 = \log(1+|x|)$ and $\gamma_2 = \log|1-|x||$ (see Appendix).

Proof. We first suppose $|x| \neq 1$. By Theorem 17.3, the two Lyapunov exponents exist with probability one. To compute them we apply the recursion given in Theorem 17.3. One first obtains the intermediate form

$$\gamma_1 = \frac{1}{2}\log(1+x^2) + \frac{1}{2}\frac{1}{2\pi}\int_0^{2\pi} \log\left[1 + \frac{2x}{1+x^2}\sin 2\theta\right] d\theta\,.$$

The second integral can be evaluated to obtain

$$\gamma_1 = \frac{1}{2}\log(1+x^2) + \frac{1}{2}\log\left[\frac{1+(1-a^2)^{1/2}}{2}\right], \tag{17.51}$$

where we have defined $a \equiv 2x/(1+x^2)$. Since we need to take the positive root of $(1-a^2)^{1/2}$, we get different expressions for $|x| > 1$ and $|x| < 1$. For $x > 1$, one obtains $1 + (1-a^2)^{1/2} = 2x^2/(x^2+1)$, whereas for $|x| < 1$ one obtains $1 + (1-a^2)^{1/2} = 2/(x^2+1)$. As a result, the growth rate is given by Eq. (17.48).

We obtain the second Lyapunov exponent using (17.38). It yields

$$\gamma_2 = \begin{cases} -\log(|x|) + \log|1-x^2| & \text{if } |x| > 1, \\ \log|1-x^2| & \text{if } |x| < 1, \end{cases} \tag{17.52}$$

from which (17.49) follows.

For the case $|x| = 1$, where the matrix $\begin{bmatrix} 1 & x \\ x & 1 \end{bmatrix}$ is singular, the formula of Theorem 17.3 still applies to give $\gamma_1 = 0$. In this case $\gamma_2 = -\infty$ since the matrices in the product have rank at most one. □

In this example the top Lyapunov exponent is a piecewise analytic function of the parameter x in several separate regimes. See [28] for a general discussion of this phenomenon.

As noted earlier $x = 1$ is the case that parallels the model considered in Sect. 17.2. Here we have

$$\mathbb{C}_k = Q_k \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} Q_k^T = Q_k' \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} (Q_k')^T$$

where Q_k' is a new rotation matrix which adds $\frac{\pi}{4}$ to the rotation angle Q_k. For the case of uniformly distributed rotations, the distribution of Q_k' is the same as Q_k. Up to scaling, the matrix $\begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$ is a one-dimensional projection, so the growth rate will be $\gamma = 2\gamma^*$ where γ^* is the growth rate for a random product of uniformly distributed one-dimensional projections. This latter problem was solved in [16], and the result is as follows.

Theorem 17.6. *Consider matrix products of independent uniformly distributed random orthogonal projections $\mathbb{D}_k$ onto a one-dimensional subspace of $\mathbf{R}^2$, drawn with uniformly distributed angle. The top Lyapunov exponent of such products is given by*

$$\gamma_1 = -\log 2.$$

One also has $\gamma_2 = -\infty$.

Proof. The result follows from Theorem 17.5 for the case $x = 1$, since rescaling of each matrix by a factor $\frac{1}{2}$ to get a projection shifts the top Lyapunov exponent by $-\log 2$.

A second approach is to directly calculate the Lyapunov exponent. We use the fact that it is given by the expected value of the logarithm of the length of a random projection of the unit vector $(0, 1)^T$ representing the initial eigenvector, taken over one step. This is given by

$$\gamma_1 := \frac{1}{2\pi} \int_0^{2\pi} \log |\cos\theta| \, d\theta = \frac{2}{\pi} \int_0^{\pi/2} \log \cos\theta \, d\theta.$$

One may verify that the integral on the right is $-\log 2$. □

17.4 Comparison of the Unstable Regime Hill Equation Model and Random Rotation Model with all $x_k = 1$

The dynamics of the random rotation model of Sect. 17.3 is directly comparable with that of the model of Sect. 17.2 in one special case, which is that where all $x_k = x = 1$. In that case the model is equivalent to drawing independent random matrices

$$\mathbb{C}_k := \mathbb{Q}_k \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \mathbb{Q}_k^{-1}. \tag{17.53}$$

where $\mathbb{Q}_k := Q_k Q_{\pi/4}$ are random rotations, with Q_k (and hence $\mathbb{Q}_k$) being independent draws from a given distribution of allowed rotations, and $\mathbb{Q}_k^T = \mathbb{Q}_k^{-1}$. Recall that the model of Sect. 17.2 has

$$\mathbb{A}_k := \mathbb{P}_k \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix} \mathbb{P}_k^{-1}. \tag{17.54}$$

with $\mathbb{P}_k \equiv \begin{bmatrix} -x_k & x_k \\ 1 & 1 \end{bmatrix}$, with each x_k drawn independently from a given distribution on the real line. The unperturbed model in both cases is simply iteration of the singular matrix $\mathbb{D} = \begin{bmatrix} 0 & 0 \\ 0 & 2 \end{bmatrix}$. We now observe that the addition of random perturbations has opposite effects on the top Lyapunov exponents on these models. For the perturbed Hill's equation model of Sect. 17.2, in the case of nonnegative random variables it

always increases the Lyapunov exponent relative to the unperturbed model, while for the random rotation model it always decreases the Lyapunov exponent relative to the unperturbed model.

For the unstable Hill's equation random model of Sect. 17.2 we have the following easy observation, in the case of nonnegative random variables.

Theorem 17.7. *For the unstable Hill's equation random matrix model (17.54), where the distribution of x_k is on the positive real line, the growth rate satisfies*

$$\gamma_1 \geq \log 2.$$

Equality holds if and only if the model has no randomness, i.e., a fixed matrix $\begin{bmatrix} 1 & x \\ 1/x & 1 \end{bmatrix}$ *is drawn with probability one at each step.*

Proof. We assert that

$$\mathbb{E}_{x_1,x_2}[\log(1+\frac{x_1}{x_2})] = \frac{1}{2}\mathbb{E}_{x_1,x_2}[\log(1+\frac{x_1}{x_2}) + \log(1+\frac{x_2}{x_1})] \geq \log 2.$$

This holds using the identity

$$\log(1+\frac{x_1}{x_2}) + \log(1+\frac{x_2}{x_1}) = \log(1+\frac{x_1}{x_2})(1+\frac{x_2}{x_1}) = \log(2+\frac{x_1}{x_2}+\frac{x_2}{x_1}) \geq 2\log 2,$$

together with $y + \frac{1}{y} \geq 2$ when $y > 0$. To obtain equality one must have $x_1 = x_2$ almost everywhere. □

Note that for the model of Theorem 17.2, allowing x_k of both signs, the exponent γ_1 can be larger than or smaller than $\log 2$, depending on the distribution.

Turning to the random rotation model of Sect. 17.3, we show that the addition of any randomness in this model always decreases the Lyapunov exponent relative to unperturbed model.

Theorem 17.8. *For the random rotation model (17.53), for any fixed distribution of random rotations the top Lyapunov exponent satisfies*

$$\gamma_1 \leq \log 2.$$

Equality holds if and only if the model has no randomness, i.e., if with probability one a fixed rotation $Q := \begin{bmatrix} \cos\alpha_0 & -\sin\alpha_0 \\ \sin\alpha_0 & \cos\alpha_0 \end{bmatrix}$ *is drawn at each step.*

Proof. For this model we observe that each matrix $\frac{1}{2}\mathbb{C}_k$ is a rank one orthogonal projection. The product of N matrices has norm 2^N times the product of N rank one orthogonal projections. But the product of rank one orthogonal projections is a constant times a rank one orthogonal projection, and this constant is always ≤ 1.

The inequality above on the Lyapunov exponent follows. To get strict inequality if the distribution is nontrivial it suffices to observe that there will be some $\delta > 0$ such that with positive probability the product of projections shrinks by a factor $1 + \delta$. This will give a strict decrease in γ_1. □

These models (17.54) and (17.53) illustrate how the form of the allowed random similarities can affect the long-term dynamics of the model. Rotational similarities always smooth the instability while the form of the shears $\mathbb{P}_k$ in (17.54), in models with all $x_k > 0$, increases the instability.

17.5 Concluding Remarks

The starting point of this paper was the observation made by the first two authors in [2] that Hill's equations can be generalized so that the parameters of the differential equation vary from cycle to cycle; the time development of the solutions is then given by the product of matrices with random elements. This motivates the study of various random matrix models of 2×2 matrices. This paper presented analytic results for the growth rates for two such random matrix models. The first model describes Hill's equation in the regime where solutions are highly unstable (Theorems 17.1 and 17.2). The second, entirely different model, is a two-parameter model of random matrices and models entirely different dynamics. The models have the same unperturbed model in one special case, and in Sect. 17.4 we demonstrated a contrast of the growth rates for the unperturbed and perturbed models in the two models. The unstable Hill's equation model increases the growth rate while the other model decreases it. Finally the Appendix presents a simple special case of the two-parameter model where the matrices commute.

The paper addresses at two goals. The first is to gain insight into the mechanism underlying the unstable regime Hill's equation model. The second is to present simple random matrix models with explicitly determinable Lyapunov exponents, which add to the small list of models where such formulas are known.

Appendix: Commuting Matrix Family

We consider the special case of the model in Sect. 17.3 with no rotations present, i.e., all $\alpha_k = 0$. That is, we consider products of matrices $\mathbb{B}_k$ that have the form

$$\mathbb{B}_k = \begin{bmatrix} 1 & x_k \\ x_k & 1 \end{bmatrix}, \tag{17.55}$$

where the variables x_k have an arbitrary distribution. These matrices form a commuting family, which permits an easy analysis of their growth rates.

Theorem 17.9. *Consider matrices of the form* $\mathbb{B}_k = \begin{bmatrix} 1 & x_k \\ x_k & 1 \end{bmatrix}$, *where the random matrix elements are independently and identically drawn from a distribution supported in* $(0,1)$ $0 \le x_k < 1$. *The growth rate for matrix products* $\mathbb{B}^{(N)}$ *takes the form*

$$\gamma = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \log(1+x_k) = \mathbb{E}_x[\log(1+x)]\,. \tag{17.56}$$

so that $\gamma \ge 0$. *The second Lyapunov exponent is given by*

$$\gamma_2 = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \log(1-x_k) = \mathbb{E}_x[\log(1-x)]\,. \tag{17.57}$$

The value $\gamma_2 = -\infty$ *is allowed.*

Proof. We first define the determinant of the matrix $\mathbb{B}_k$, $d_k \equiv 1 - x_k^2$, and rewrite the matrix in the form

$$\mathbb{B}_k = \sqrt{d_k} \begin{bmatrix} \cosh\theta_k & \sinh\theta_k \\ \sinh\theta_k & \cosh\theta_k \end{bmatrix},$$

where

$$\theta_k = \frac{1}{2} \log\left(\frac{1+x_k}{1-x_k}\right).$$

The product of N such matrices takes the form

$$\mathbb{B}^{(N)} = \left\{ \prod_{k=1}^{N} (1-x_k^2)^{1/2} \right\} \begin{bmatrix} \cosh\theta_N & \sinh\theta_N \\ \sinh\theta_N & \cosh\theta_N \end{bmatrix}, \tag{17.58}$$

where the new angle (argument) θ_N is given by

$$\theta_N = \sum_{k=1}^{N} \theta_k = \sum_{k=1}^{N} \frac{1}{2} \log\left(\frac{1+x_k}{1-x_k}\right),$$

so that

$$\mathrm{e}^{\theta_N} = \prod_{k=1}^{N} \left(\frac{1+x_k}{1-x_k}\right)^{1/2}.$$

For simplicity of notation, we define

$$P_N = \prod_{k=1}^{N} d_k^{1/2} = \prod_{k=1}^{N} (1 - x_k^2)^{1/2} .$$

The largest eigenvalue of the matrix at the Nth iteration in then given by

$$\lambda_N = P_N e^{\theta_N} = \prod_{k=1}^{N} (1 + x_k) . \tag{17.59}$$

The growth rate for matrix multiplication is then given by

$$\gamma = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \log(1 + x_k) , \tag{17.60}$$

as claimed in Eq. (17.56). By the same token, the second eigenvalue is given by

$$\frac{\det(\mathbb{B}^{(N)})}{\lambda_N} = P_N e^{-\theta_N} = \prod_{k=1}^{N} (1 - x_k) . \tag{17.61}$$

This yields

$$\gamma = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \log(1 - x_k) , \tag{17.62}$$

which is the smaller exponent since $0 \leq x_k < 1$. □

Remark. Alternatively, Eq. (17.60) provides the growth rate as follows: The matrices have eigenvalues $(1 \pm x_k)$ and eigenvectors $[1, 1]$ and $[1, -1]$. Since the eigenvectors are the same for all k, after repeated matrix multiplications only the vector component with the first eigenvector $[1, 1]$ survives (since this eigenvector always corresponds to the larger eigenvalue). Further, the leading coefficient will be $\prod(1 + x_k)$. The growth rate of Eq. (17.60) then follows from the definitions.

Acknowledgements We thank Jake Ketchum for useful discussions. This work was supported in part by the NSF and NASA. The first author received support from NSF grant DMS-0806795 and NASA grant NNX11AK87G9. The second author received support form NSF grants DMS-0806756, DMS-0907949 and DMS-1207693. The third author received support from NSF grant DMS-1101373.

References

1. Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions. Dover, New York (1970)
2. Adams, F.C., Bloch, A.M.: Hill's equation with random forcing terms. SIAM J. Appl. Math. **68**, 947–980 (2008)

3. Adams, F.C., Bloch, A.M.: Hill's equation with random forcing terms: the limit of delta function barriers. J. Math. Phys. **50**, 073501 (2009)
4. Adams, F.C., Bloch, A.M.: Hill's equation with random forcing parameters: determination of growth rates through random matrices. J. Stat. Phys. **139**, 139–158 (2010)
5. Adams, F.C., Bloch, A.M., Butler, S.C., Druce, J.M., Ketchum, J.A.: Orbits and instabilities in a triaxial cusp potential. Astrophys. J. **670**, 1027–1047 (2007)
6. Binney, J.: Resonant excitation of motion perpendicular to galactic planes. Mon. Not. R. Astron. Soc. **196**, 455–467 (1981)
7. Binney, J., Tremaine, S.: Galactic Dynamics. Princeton Univ. Press, Princeton (1987)
8. Cohen, J.E., Newman, C.M.: The stability of large random matrices and their products. Ann. Prob. **12**, 283–310 (1984)
9. Carmona, R., Lacroix, J.: Spectral Theory of Random Schrödinger Operators. Birkhauser, Boston (1990)
10. Comtet, A., Texler, C., Tourigny, Y.: Products of random matrices and generalized quantum point scatterers. J. Stat. Phys. **140**, 427–466 (2010)
11. Cook, J., Derrida, B.: Lyapunov exponents of large, sparse random matrices and the problem of directed polymers with complex random weights. J. Stat. Phys. **61**, 961–986 (1990)
12. Furstenberg, H.: Noncommuting random products. Trans. Am. Math. Soc. **108**, 377–428 (1963)
13. Furstenberg, H., Kesten, H.: Products of random matrices. Ann. Math. Stat. **31**, 457–469 (1960)
14. Guth, A.H.: Inflationary universe: a possible solution to the horizon and flatness problems. Phys. Rev. D **23**, 347–356 (1981)
15. Hill, G.W.: On the part of the motion of the lunar perigee which is a function of the mean motions of the Sun and Moon. Acta Math. **8**, 1–36 (1886)
16. Högnäs, G.: On products of random projections. Acta Acad. Aboensis Ser. B **44**(5), 18 pp. (1984)
17. Högnäs, G., Mukherjea, A.: Probability Measures on Semigroups. Convolution Products, Random Walks and Random Matrices, 2nd edn. Springer, New York (2011)
18. Keller, J.B.: Ponytail motion. SIAM J. Appl. Math. **70**, 2667–2672 (2010)
19. Key, E.: Computable examples of the maximal Lyapunov exponent. Probab. Theory Related Fields **75**, 97–107 (1987)
20. Kofman, L., Linde, A., Starobinsky, A.A.: Reheating after inflation. Phys. Rev. Lett. **73**, 3195–3198 (1994)
21. Lima, R., Rahibe, M.: Exact Lyapunov exponent for infinite products of random matrices. J. Phys. A Math. Gen. **27**, 3427–3437 (1994)
22. Lubow, S.H.: Tidally driven inclination instability in Keplerian disks. Astrophys. J.**398**, 525–530 (1992)
23. Magnus, W., Winkler, S.: Hill's Equation. Wiley, New York (1966)
24. Marklof, J.: Explicit invariant measures for products of random matrices. Trans. Am. Math. Soc. **360**, 3391–3427 (2008)
25. Moshe, Y.: Random matrix produces and applications to cellular automata. J. Anal. Math. **99**, 267–294 (2006)
26. Newman, C.M.: The distribution of Lyapunov exponents: exact results for random matrices. Commun. Math. Phys. **103**, 121–126 (1986)
27. Newman, C.M.: Lyapunov exponents for some products of random matrices: exact expressions and asymptotic distributions. In: Cohen, J.E., Kesten, H., Newman, C.M. (eds.) Random Matrices and Their Applications. Contemp., Math., vol. 50, pp. 121–141. AMS, Providence (1986)
28. Peres, Y.: Domains of analytic continuation for the top Lyapunov exponent. Ann. Inst. H. Poincaré Probab. Stat. **28**, 131–148 (1992)
29. Pincus, S.: Strong laws of large numbers for products of random matrices. Trans. Am. Math. Soc. **287**, 65–89 (1985)
30. Pollicott, M.: Maximal Lyapunov exponents for random matrix products. Invent. Math. **181**, 209–226 (2010)

31. Stroock, D.: Essentials of Integration Theory for Analysis. Graduate Texts in Math. No. 262. Springer, New York (2011)
32. Tsiitsiklis, J.N., Blondel, V.D.: The Lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate. Math. Control Signals Syst. **10**, 31–40 (1997)
33. Widder, D.H.: The Laplace Transform. Princeton Univ. Press, Princeton (1941)

Part III
Continuum Mechanics: Solids, Fluids and Other Materials

Chapter 18
The Three-Dimensional Globally Modified Navier–Stokes Equations: Recent Developments

T. Caraballo and P.E. Kloeden

Herrn Prof. Dr. Jürgen Scheurle zu seinem sechzigsten Geburtstag gewidmet

Abstract The globally modified Navier–Stokes equations (GMNSE) were introduced by Caraballo, Kloeden and Real (Adv. Nonlinear Stud. 6:411–436, 2006) in 2006 and have been investigated in a number of papers since then, both for their own sake and as a means of obtaining results about the three-dimensional Navier–Stokes equations. These results were reviewed by Kloeden et al. (Advances in Nonlinear Analysis: Theory, Methods and Applications, Cambridge Scientific Publishers, Cambridge, 2009; pp 11–22.), which was published in 2009, but there have been some important developments since then, which will be reviewed here.

18.1 Introduction

The three-dimensional Navier–Stokes equations (NSE) are an intriguing system of partial differential equations. They have been intensively investigated for many years, but some very basic issues on their solvability remain unresolved. For example, although weak solutions are known to exist for all future time for each initial condition in the function space H, it is not known if there is a unique

T. Caraballo
Dpto. Ecuaciones Diferenciales y Análisis Numérico, Universidad de Sevilla,
Apdo. de Correos 1160, ES-41080 Sevilla, Spain
e-mail: caraball@us.es

P.E. Kloeden (✉)
Institut für Mathematik, Goethe-Universität, D-60054 Frankfurt am Main, Germany
e-mail: kloeden@math.uni-frankfurt.de

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_18,

weak solution. Nor is it known if a strong solution for each initial condition in the function space V can exist for more than a short time.

Let $\Omega \subset \mathbb{R}^3$ be an open bounded set with regular boundary Γ. The system of Navier–Stokes equations (NSE) on Ω with a homogeneous Dirichlet boundary condition is given by

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \Delta u + (u \cdot \nabla)u + \nabla p = f(t) \ \text{ in } \ (\tau, +\infty) \times \Omega, \\ \nabla \cdot u \ = 0 \ \text{ in } \ (\tau, +\infty) \times \Omega, \\ u = 0 \ \text{ on } (\tau, +\infty) \times \Gamma, \\ u(\tau, x) = u_0(x), \quad x \in \Omega, \end{cases} \tag{18.1}$$

where $\nu > 0$ is the kinematic viscosity, u is the velocity field of the fluid, p the pressure, $\tau \in \mathbb{R}$ the initial time, u_0 the initial velocity field, and $f(t)$ a given external force field.

There have been many modifications of the Navier–Stokes equations, starting with Leray and mostly involving the nonlinear term, see the review paper of Constantin [6]. Another modification, called the globally modified Navier–Stokes equations (GMNSE), was introduced by Caraballo, Kloeden, and Real [1] in 2006.

Fix $N \in \mathbb{R}^+$ and define $F_N : \mathbb{R}^+ \to \mathbb{R}^+$ by

$$F_N(r) := \min\left\{1, \frac{N}{r}\right\}, \quad r \in \mathbb{R}^+.$$

The system

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \Delta u + F_N\left(\|u\|\right)\left[(u \cdot \nabla)u\right] + \nabla p = f(t) \ \text{ in } \ (\tau, +\infty) \times \Omega, \\ \nabla \cdot u \ = 0 \ \text{ in } \ (\tau, +\infty) \times \Omega, \\ u = 0 \ \text{ on } (\tau, +\infty) \times \Gamma, \\ u(\tau, x) = u_0(x), \quad x \in \Omega, \end{cases} \tag{18.2}$$

is called the *globally modified Navier–Stokes equations (GMNSE)* with parameter N.

The GMNSE (18.2) are indeed *globally* modified since the modifying factor $F_N\left(\|u\|\right)$ depends on the norm $\|u\| = \|\nabla u\|_{(L^2(\Omega))^{3\times 3}}$, which in turn depends on ∇u over the whole domain Ω and not just at or near the point $x \in \Omega$ under consideration. Essentially, it prevents large gradients dominating the dynamics and leading to explosions. It is worth mentioning that, for a different purpose, Flandoli

and Maslowski [8] used a similar global cutoff function involving the $D(A^{1/4})$ norm for the two-dimensional stochastic Navier–Stokes equations.

The GMNSE (18.2) violate the basic laws of mechanics, but mathematically they are a well-defined system of equations, just like the modified versions of the NSE of Leray and others with other mollifications of the nonlinear term. They are nevertheless interesting mathematically in their own right but are also useful for obtaining new results about the three-dimensional Navier–Stokes equations, which will be briefly discussed below.

18.1.1 Notation

The usual notation and abstract framework for the Navier–Stokes equations of Lions [16] and Temam [22] is used with H denoting the closure of

$$\mathscr{V} = \left\{ u \in (C_0^\infty(\Omega))^3 : \operatorname{div} u \ = 0 \right\},$$

in $(L^2(\Omega))^3$ with inner product $(u,v) = \sum_{j=1}^3 \int_\Omega u_j(x)v_j(x)\,\mathrm{d}x$, for $u, v \in (L^2(\Omega))^3$, with associated norm $|\cdot|$, and V denoting the closure of $\mathscr{V}$ in $(H_0^1(\Omega))^3$ with inner product $((u,v)) = \sum_{i,j=1}^3 \int_\Omega \frac{\partial u_j}{\partial x_i}\frac{\partial v_j}{\partial x_i}\,\mathrm{d}x$, for $u, v \in (H_0^1(\Omega))^3$, with associated norm $\|\cdot\|$. In addition, b_N and B_N are defined by

$$b_N(u,v,w) = F_N(\|v\|)b(u,v,w), \quad \forall\, u,v,w \in V.$$

and

$$\langle B_N(u,v), w\rangle = b_N(u,v,w), \quad \forall\, u,v,w \in V,$$

respectively, where b is the trilinear form on $V \times V \times V$ given by

$$b(u,v,w) = \sum_{i,j=1}^3 \int_\Omega u_i \frac{\partial v_j}{\partial x_i} w_j\,\mathrm{d}x, \quad \forall\, u,v,w \in V.$$

Finally, define $A : V \to V'$ by $\langle Au, v\rangle = ((u,v))$. Then $Au = -P\Delta u, \forall\, u \in D(A)$, where $D(A) = (H^2(\Omega))^3 \cap V$ and P is the orthonormal projector from $(L^2(\Omega))^3$ onto H.

18.2 Existence and Regularity of Solutions

The existence, uniqueness, and regularity theory of strong and weak solutions of the three-dimensional GMNSE is closer to that of the two-dimensional than the three-dimensional NSE due to the special properties of b_N, which is linear in u and w, but nonlinear in v, and satisfies $b_N(u,v,v) = 0$ for all $u, v \in V$ as well as the estimate

$$|b_N(u,v,w)| = F_N(\|v\|)|b(u,v,w)| \leq NC_1\|u\|\|w\| \quad \forall\, u,v,w \in V. \tag{18.3}$$

This and many other estimates, which can be found in Caraballo, Kloeden, and Real [1], are very similar to those for the two-dimensional NSE and lead to similar results. In particular, the GMNSE have a unique global strong solution for each initial condition in the function space V as well as global weak solutions for each initial condition in the function space H, which instantaneously become strong solutions. Originally, in [1] it was not known if the weak solutions were unique, but this was later established by Romito [21] and thus allowed a number of proofs that had appeared in some papers published in the period between these two to be simplified.

18.2.1 Weak Solutions

Let $u_0 \in H$ and $f \in L^2(\tau, T; (L^2(\Omega))^3)$ for all $T > \tau$ be given. A weak solution of (18.2) is any $u \in L^2(\tau, T; V)$ for all $T > \tau$ such that

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t}u(t) + \nu Au(t) + B_N(u(t),u(t)) = f(t) \text{ in } \mathscr{D}'(\tau,+\infty;V'), \\ u(\tau) = u_0, \end{cases} \tag{18.4}$$

or equivalently

$$(u(t),w) + \nu\int_\tau^t ((u(s),w))\,ds + \int_\tau^t b_N\,(u(s),u(s),w)\;ds = (u_0,w) + \int_\tau^t (f(s),w)\;ds, \tag{18.5}$$

for all $t \geq \tau$ and all $w \in V$.

Due to (18.3), unlike the three-dimensional NSE, any weak solution $u(t)$ of GMNSE belongs to $C([\tau,+\infty);H)$ and satisfies (see Remark 1 in [1]) the energy equality

$$|u(t)|^2 - |u(s)|^2 + 2\nu\int_s^t \|u(r)\|^2\,dr = 2\int_s^t (f(r),u(r))\,dr \quad \text{for all } \tau \leq s \leq t. \tag{18.6}$$

The existence of weak solutions of the GMNSE is contained in Theorem 23.2 below which also considers the existence of strong solutions. The following result is the counterpart of Serrin's classical theorem on the three-dimensional NSE which says that a strong solution, if it exists, is unique in the class of weak solutions. Strong solutions (to be defined below) for the GMNSE are examples of the weak solutions in the next theorem.

Theorem 18.1 ([1], Theorem 3). *If there exists a weak solution u of (18.2) such that $u \in L^2(\tau, T; D(A))$ for all $T > \tau$, then u is the unique weak solution of (18.2).*

This result is not as important as originally thought since the weak solutions of the GMNSE have been shown to be unique. The proof is similar to the NSE case and depends on the following result.

Lemma 18.1 ([1], Lemma 6). *For all M, N, p, $r \in \mathbb{R}^+$ it holds*

$$|F_M(p) - F_N(r)| \leq \frac{|M - N|}{r} + \frac{|p - r|}{r}.$$

18.2.2 Strong Solutions

The following theorem is the basic existence and regularity result for strong and also weak solutions of the GMNSE.

Theorem 18.2 ([2]). *Suppose $f \in L^2(\tau, T; (L^2(\Omega))^3)$ for all $T > \tau$, and let $u_0 \in H$ be given. Then, there exists a unique weak solution u of (18.2), which is, in fact, a strong solution in the sense that*

$$u \in C([\tau + \varepsilon, T]; V) \cap L^2(\tau + \varepsilon, T; D(A)), \tag{18.7}$$

for all $T > \tau + \varepsilon > \tau$.

Moreover, if $u_0 \in V$, then

$$u \in C([\tau, T]; V) \cap L^2(\tau, T; D(A)), \tag{18.8}$$

for all $T > \tau$.

The first statement in Theorem 23.1 was originally given as "there exists *at least one* weak solution u of the GMNSE" in Theorem 7 of [1], but takes its present form after Romito showed that "there exists *at most one* weak solution u of the GMNSE" in Theorem 1.1 of [21]. Romito used the estimate

$$|(w, B(u, v))| \leq \|u\|_{L^6} \|\nabla v\|_{L^2} \|w\|_{L^2}^{\frac{1}{2}} \|w\|_{L^6}^{\frac{1}{2}} \leq c_0 \|u\| \|v\| |w|^{\frac{1}{2}} \|w\|^{\frac{1}{2}}, \tag{18.9}$$

for $u, v, w \in V$, since $\|u\|_{L^6} \leq c\|u\|$. He used this to show that the nonlinear term $\mathscr{NL}(u,v) := F_N(\|u\|)B(u,u) - F_N(\|v\|)B(v,v)$ could be estimated by

$$(w, \mathscr{NL}(u,v)) \leq \nu\|w\|^2 + C(c_0,\nu)N^4|w|^2,$$

where $w = u - v$, the difference of two weak solutions.

18.2.2.1 Continuity of Strong Solutions on Data

Strong solutions $u^{(N)}(t,\tau,u_0)$ of the GMNSE (18.2) with parameter N depend continuously on the parameter N as well as on the initial value u_0.

Theorem 18.3 ([12], Theorem 8). *Suppose that $f \in L^2(\tau,T;(L^2(\Omega))^3)$ for all $T > \tau$, and let $N, M > 0$, and $u_0, v_0 \in V$ be given. Denote by $u^{(N)}(t) = u^{(N)}(t,\tau,u_0)$ (respectively, $v^{(M)}(t) = u^{(M)}(t,\tau,v_0)$) the solution of the GMNSE (18.2) corresponding to the parameter N and the initial value u_0 (respectively, to the parameter M and the initial condition v_0). Then, there exists a positive constant $C > 0$ depending only on Ω and ν such that for all $t \geq \tau$*

$$||v^{(M)}(t) - u^{(N)}(t)||^2 \leq [||v_0 - u_0||^2 + C(M-N)^2 \int_\tau^t |Au^{(N)}(s)|^2 ds] \times$$
$$\times \exp\left(C\left(M^4(t-\tau) + \int_\tau^t |Au^{(N)}(s)|^2 ds\right)\right) \tag{18.10}$$

and

$$\nu \int_\tau^t |Av^{(M)}(s) - Au^{(N)}(s)|^2 ds \leq [||v_0 - u_0||^2 + C(M-N)^2 \int_\tau^t |Au^{(N)}(s)|^2 ds] \times$$
$$\left[1 + \left(C\int_\tau^t \left(|Au^{(N)}(s)|^2 + M^4\right) ds\right) \times \exp\left(C\left(M^4(t-\tau) + \int_\tau^t |Au^{(N)}(s)|^2 ds\right)\right)\right]. \tag{18.11}$$

As a consequence of the previous theorem, we obtain

Theorem 18.4. *Suppose that $f \in L^2(\tau,T;(L^2(\Omega))^3)$ for all $T > \tau$. Then, for any $u_0 \in V$ and $N > 0$ given,*

$$u^{(M)}(\cdot,\tau,v_0) \to u^{(N)}(\cdot,\tau,u_0) \ \text{ in } \ C([\tau,T];V) \cap L^2(\tau,T;D(A))$$

as $(M,v_0) \to (N,u_0)$ in $\mathbb{R}^+ \times V$ for all $T > \tau$.

18.2.2.2 Estimates of Strong Solutions in $D(A)$

With stronger assumptions on the external forcing term f, estimates of the solution in the norm of $D(A)$ can be obtained.

Theorem 18.5 ([3], Proposition 4). *Suppose that $f \in W^{1,\infty}(\tau, +\infty; H)$, and let $u^{(N)}(t)$ be a solution of the GMNSE (18.2) with parameter N. Then*

$$u^{(N)}(t) \in D(A), \quad \forall\, t > \tau, \tag{18.12}$$

and there exist two positive constants $R_f^{(N)}$ and $M_f^{(N)}$, independent of ε, τ, u_0, and t, and increasing with $|f|_\infty$ and $|f'|_\infty$, such that

a) if $u(\tau) = u_0 \in V$, then

$$|Au^{(N)}(t)|^2 \le (1+\varepsilon^{-1})\left[R_f^{(N)} + M_f^{(N)}(1+t-\tau)\|u_0\|^2 e^{-\nu\lambda_1(t-\tau)}\right], \tag{18.13}$$

for all $t \ge \tau + \varepsilon$, $\varepsilon \in (0,1]$;

b) in general, if $u(\tau) = u_0 \in H$, then

$$|Au^{(N)}(t)|^2 \le (1+\varepsilon^{-1})R_f^{(N)} + \varepsilon^{-1}(1+\varepsilon^{-1})M_f^{(N)}(1+t-\tau)(1+|u_0|^2)e^{-\nu\lambda_1(t-\tau)}, \tag{18.14}$$

for all $t \ge \tau + 2\varepsilon$, $0 < \varepsilon \le 1$. In particular, there exists a $T_0 = T_0(|u_0|)$ depending only on $|u_0|$, $R_f^{(N)}$ and $M_f^{(N)}$, such that

$$|Au^{(N)}(t)|^2 \le 2R_f^{(N)}, \quad \forall\, t \ge \tau + T_0(|u_0|). \tag{18.15}$$

Remark 18.1. Observe that (18.14) implies that if $f \in W^{1,\infty}(\tau, +\infty; H)$, then every solution of GMNSE belongs to $L^\infty(\tau+\varepsilon, +\infty; D(A))$ for all $\varepsilon > 0$. If, moreover, the initial datum $u_0 \in D(A)$, then it can be proved that the corresponding solution $u = u^{(N)}(t)$ of GMNSE belongs to $L^\infty(\tau, +\infty; D(A))$, and more exactly,

$$\sup_{t\ge\tau} |Au(t)| < +\infty.$$

18.3 Global Attractor in V: Existence and Dimension Estimate

18.3.1 Autonomous Case

Assume now that the forcing term f does not depend on time and for each $u_0 \in V$ define $S^{(N)}(t)u_0 := u^{(N)}(t, u_0)$, where $u^{(N)}(t, u_0)$ is the unique strong solution

$u^{(N)}(t)$ of (18.2) with initial time $\tau = 0$. From Theorems 23.1 and 18.4, it follows that $\{S^{(N)}(t)\}_{t\geq 0}$ is a C^0 semigroup in V. Let $u^{(N)}(t) = S^{(N)}(t)u_0$ with $u_0 \in V$. The same arguments as for the NSE give the inequality

$$\frac{d}{dt}|u^{(N)}|^2 + \nu\lambda_1|u^{(N)}|^2 \leq \frac{1}{\nu\lambda_1}|f|^2, \tag{18.16}$$

where λ_1 is the first eigenvalue of A, and hence the estimate

$$|u^{(N)}(t)|^2 \leq |u_0|^2 e^{-\nu\lambda_1 t} + \frac{1}{\nu^2\lambda_1^2}|f|^2\left(1 - e^{-\nu\lambda_1 t}\right),$$

from which it follows that $S^{(N)}(t)$ possesses a set $\mathscr{B}_H$ in H which absorbs bounded sets of V, and which is given by $\mathscr{B}_H := \{u \in H \ : |u|^2 \leq 1 + \frac{1}{\nu^2\lambda_1^2}|f|^2\}$. Similarly, but more complicatedly, $S^{(N)}(t)$ has an absorbing set $\mathscr{B}_V^{(N)}$ in V (i.e., which absorbs bounded sets of V) given by

$$\mathscr{B}_V^{(N)} := \left\{u \in V \ : \|u\|^2 \leq 1 + \frac{|f|^2}{\nu^2\lambda_1}\left(2 + \frac{C^{(N)}}{\nu\lambda_1^2}\right)\right\}. \tag{18.17}$$

Note that $\mathscr{B}_V^{(N)} \subset \mathscr{B}_V^{(N^*)}$ for $N \leq N^*$ in view of the definition of the constant $C^{(N)}$ (see [1] for details).

Moreover, the semigroup $S^{(N)}(t)$ in V is asymptotically compact since it satisfies the flattening property ([11], see also [20]): "*For any bounded set B of V and for any $\varepsilon > 0$, there exists $T_\varepsilon(B) > 0$ and a finite dimensional subspace V_ε of V, such that $\left\{P_\varepsilon S^{(N)}(t)B, t \geq T_\varepsilon(B)\right\}$ is bounded and*

$$\left\|(I - P_\varepsilon)S^{(N)}(t)u_0\right\| < \varepsilon \quad \textit{for} \quad t \geq T_\varepsilon(B), u_0 \in B, \tag{18.18}$$

where $P_\varepsilon : V \to V_\varepsilon$ is the projection operator." It thus follows that the GMNSE (18.2) has a global attractor $\mathscr{A}_N$ in V for each N. In particular, $\mathscr{A}_N \subset \mathscr{B}_V^{(N)}$ for each N and

Theorem 18.6 ([1], Theorem 10). *If $f \in (L^2(\Omega))^3$, then the GMNSE (18.2) has a global attractor $\mathscr{A}_N$ in V for each $N > 0$. Moreover the set-valued mapping $N \mapsto \mathscr{A}_N$ is upper semi continuous, i.e.*

$$dist_V\left(\mathscr{A}_M, \mathscr{A}_N\right) \to 0 \quad as \quad M \to N, \tag{18.19}$$

where $dist_V$ is the Hausdorff semi distance on V.

The upper semi continuous dependence of the global attractors $\mathscr{A}_N$ in N follows by standard theorems in dynamical systems theory in view of the continuity of the semigroups $S^{(N)}$ in N established in Theorem 18.4.

18.3.1.1 Global Attractor in $D(A)$

With time-independent forcing (so that the stronger assumption of Theorem 18.5 is satisfied) it is possible to obtain an absorbing set $\mathscr{B}_N := \{v \in D(A) : |Av|^2 \leq 2R_f^{(N)}\}$ in $D(A)$ for the semigroup $\{S^{(N)}(t)\}_{t\geq 0}$ and hence the above global attractor $\mathscr{A}_N$ actually belongs to $D(A)$. In fact

Corollary 18.1 ([3], Corollary 7). *The global attractor $\mathscr{A}_N$ of the GMNSE is a bounded subset of $D(A)$.*

18.3.2 Nonautonomous Case

In the nonautonomous case, when f depends on time, the counterpart of a semigroup is a 2-parameter semigroup of operators $U^{(N)}(t,\tau)$, with $U^{(N)}(t,\tau)u_0 = u^{(N)}(t,\tau,u_0)$ the solution of (18.2) for $u_0 \in V$. In addition, the counterpart of an attractor is a pullback attractor, i.e., a family of nonempty compact subsets $\{\mathscr{A}_N(t), t \in \mathbb{R}\})$ in V, which is invariant in the sense that $S^{(N)}(t,\tau)\mathscr{A}_N(\tau) = \mathscr{A}_N(t)$ for all $t \geq \tau$ and is pullback attracting in V, see [12]. Supposing that f belongs to $L^2_{loc}(\mathbb{R};(L^2(\Omega))^3)$ and satisfies

$$\int_{-\infty}^{t} e^{\nu\lambda_1 s}|f(s)|^2\,ds < +\infty \quad \text{for all } t \in \mathbb{R}, \tag{18.20}$$

the existence of a pullback attractor in V for the GMNSE was established in [12, Theorem 13]. Among other properties for the pullback attractor in V, a finite bound on the fractal dimension, which could increase with increasing time, was also obtained in [12].

Theorem 18.7 ([12], Theorem 22). *Suppose that $f \in W^{1,2}_{loc}(\mathbb{R};L^2(\Omega)^3)$ satisfies*

$$f \in L^\infty(-\infty,t_0;L^2(\Omega)^3), \text{ and } \sup_{r\leq t_0}\int_r^{r+1}|f'(s)|^2\,ds < +\infty, \quad \text{for all } t_0 \in \mathbb{R}. \tag{18.21}$$

Then, for each $N > 0$ and each $t_0 \in \mathbb{R}$ there exists a $d^{(N)}(t_0) \in [0,+\infty)$ such that the fractal dimension of the pullback attractor $\{\mathscr{A}_N(t), t \in \mathbb{R}\})$ of the GMNSE (18.2) satisfies the bound

$$d_F^V(\mathscr{A}_N(t)) \leq d^{(N)}(t_0) \quad \text{for all } t \leq t_0. \tag{18.22}$$

Recall that the fractal dimension of a nonempty subset C of a metric space (X, d_X) is given by

$$\mathrm{d}_F^X(C) := \limsup_{\varepsilon \downarrow 0} \frac{\log(N_\varepsilon(C))}{\log(1/\varepsilon)}, \tag{18.23}$$

where $N_\varepsilon(C)$ denotes the minimum number of balls in X with radius ε which are required to cover C.

18.4 Globally Modified NSE with Delays

There are many real situations in which one can consider that a model is better described if we allow some delay in the equations. These situations may appear, for instance, when we want to control the system by applying a force which takes into account not only the present state of the system but also the history of the solutions. Therefore, it is interesting to consider the following version of GMNSE (we will refer to it as GMNSED):

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \Delta u + F_N\left(\|u\|\right)\left[(u \cdot \nabla)u\right] + \nabla p = g(t, u_t) \ \text{ in } (\tau, +\infty) \times \Omega, \\ \nabla \cdot u \ = 0 \ \text{ in } \ (\tau, +\infty) \times \Omega, \\ u = 0 \ \text{ on } (\tau, +\infty) \times \Gamma, \\ u(\tau, x) = u^0(x), \quad x \in \Omega, \\ u(\tau + s, x) = \phi(s, x), \ s \leq 0, x \in \Omega, \end{cases} \tag{18.24}$$

where $\tau \in \mathbb{R}$ is an initial time, the term $g(t, u_t)$ is an external force depending eventually on the history of the solution, where u_t denotes the segment of solution up to time t (in other words, $u_t : s \in (-\infty, 0] \mapsto u_t(s) := u(t + s)$) and ϕ is a given velocity field defined for $s \leq 0$.

This is a general formulation when the delay is allowed to be infinite. But on some occasions it can be finite or bounded. In these cases, we consider the initial vector field ϕ defined in a bounded interval $[-h, 0]$ and the segment solution u_t is also defined in the same interval.

Some examples for the delay external force will be given below, but first, it is important to note that the function g is not defined directly on the phase space but on some class of continuous functions: either in $\mathbb{R} \times C([-h, 0]; H)$ with the sup norm (for finite delays), or $\mathbb{R} \times C_\gamma((-\infty, 0]; H)$ (in the infinite delay case) where the space $C_\gamma(H) := C_\gamma((-\infty, 0]; H)$, defined as

$$C_\gamma((-\infty, 0]; H) := \left\{ \varphi \in C((-\infty, 0]; H) : \exists \lim_{s \to -\infty} e^{\gamma s} \varphi(s) \in H \right\},$$

is a Banach space for the norm

$$\|\varphi\|_\gamma := \sup_{s\in(-\infty,0]} e^{\gamma s}|\varphi(s)|.$$

1. *Constant delay*. Consider $g(t,u_t) := G_1(u(t-h))$ where $G_1 : H \to H$ is a suitable function and $h > 0$ is the constant delay. Here the function $g : \mathbb{R} \times C([-h,0];H) \to H$ is defined as:

$$g(t,\xi) = G_1(\xi(-h)),\ \ \xi \in C([-h,0];H).$$

   Notice that the time variable t does not play any role, so we are in an autonomous situation.
2. *Variable delay*. In this case, the delay term is given by $g(t,u_t) := G_2(t,u(t-\rho(t)))$, where $\rho(t) \in [-h,0]$ is a delay function. Now, the function g is given by

$$g(t,\xi) = G_2(\xi(-\rho(t))),\ \ \xi \in C([-h,0];H),$$

   where it is clear that the time variable t is necessary for this case. So, we are in a nonautonomous model.
3. *Distributed infinite delay*. (cf. [18]) Let us consider the operator $g : \mathbb{R} \times C_\gamma(H) \to (L^2(\Omega))^3$ defined as

$$g(t,\xi) = \int_{-\infty}^{0} G_3(t,s,\xi(s))ds,\ \ t \in \mathbb{R},\ \xi \in C_\gamma(H),$$

   where the function $G_3 : \mathbb{R} \times (-\infty,0) \times \mathbb{R}^3 \to \mathbb{R}^3$ satisfies suitable assumptions. This situation corresponds to the case

$$g(t,u_t) = \int_{-\infty}^{0} G_3(t,s,u(t+s))ds,$$

   which is also nonautonomous.

On the one hand, the two first cases (constant and variable delay) have been analyzed in [5], where the authors proved existence and uniqueness of weak solutions, existence, and asymptotic behavior of stationary solutions, and the existence of pullback attractor (which becomes the global attractor in the autonomous case). On the other hand, the infinite delay case is studied in [18], where the existence and uniqueness of solutions, and the existence and asymptotic behavior of stationary solutions is proved.

We will only include below some representative results from the paper [5], so we consider g to be defined as in case (2).

Assume $G_2 : \mathbb{R} \times H \longrightarrow H$ is such that

c1) $G_2(\cdot, u) : \mathbb{R} \longrightarrow H$ is measurable, $\forall u \in H$,
c2) there exists nonnegative function $m \in L^p_{loc}(\mathbb{R})$ for some $1 \leq p \leq +\infty$, and a nondecreasing function $L : (0, \infty) \to (0, \infty)$, such that for all $R > 0$ if $|u|, |v| \leq R$, then

$$|G_2(t, u) - G_2(t, v)| \leq L(R) m^{1/2}(t) |u - v|,$$

for all $t \in \mathbb{R}$, and
c3) there exists a nonnegative function $f \in L^1_{loc}(\mathbb{R})$, such that for any $u \in H$,

$$|G_2(t, u)|^2 \leq m(t) |u|^2 + f(t), \quad \forall\, t \in \mathbb{R}.$$

Finally, we suppose $\phi \in L^{2p'}(-h, 0; H)$ and $u^0 \in H$, where $\frac{1}{p} + \frac{1}{p'} = 1$.
In this situation, we consider a delay function $\rho \in C^1(\mathbb{R})$ such that $0 \leq \rho(t) \leq h$ for all $t \in \mathbb{R}$, and there exists a constant ρ_* satisfying

$$\rho'(t) \leq \rho_* < 1 \quad \forall\, t \in \mathbb{R}. \tag{18.25}$$

Definition 18.1. Let $\tau \in \mathbb{R}$, $u^0 \in H$ and $\phi \in L^{2p'}(-h, 0; H)$ be given. A weak solution of (18.24) is a function

$$u \in L^{2p'}(\tau - h, T; H) \cap L^2(\tau, T; V) \cap L^\infty(\tau, T; H)$$

for all $T > \tau$, such that

$$\begin{cases} \dfrac{\mathrm{d}}{\mathrm{d}t} u(t) + \nu A u(t) + B_N(u(t), u(t)) = G_2(t, u(t - \rho(t))) \text{ in } \mathscr{D}'(\tau, +\infty; V'), \\ u(\tau) = u^0, \\ u(t) = \phi(t - \tau) \quad t \in (\tau - h, \tau), \end{cases}$$

or equivalently

$$(u(t), w) + \nu \int_\tau^t ((u(s), w))\, ds + \int_\tau^t b_N\left(u(s), u(s), w\right)\, ds = \left(u^0, w\right) + \int_\tau^t \left(G_2(s, u(s - \rho(s))), w\right)\, ds, \tag{18.26}$$

for all $t \geq \tau$ and all $w \in V$, and coincides with $\phi(t)$ in $(\tau - h, \tau)$.

The existence and uniqueness of weak (and strong) solutions of our problem is established in a similar way as we did in the non-delay case, but with necessary changes due to the delay term.

Theorem 18.8 ([5], Theorem 3.1). *Under the conditions c1)–c3), assume that* $\tau \in \mathbb{R}$, $u^0 \in H$ *and* $\phi \in L^{2p'}(-h, 0; H)$ *are given. Then, there exists a unique weak solution* u *of (18.24) which is, in fact, a strong solution in the sense that*

$$u \in C([\tau + \varepsilon, T]; V) \cap L^2(\tau + \varepsilon, T; D(A)), \tag{18.27}$$

for all $T - \tau > \varepsilon > 0$.

Moreover, if $u^0 \in V$, *then* $u \in C([\tau, T]; V) \cap L^2(\tau, T; D(A))$, *for all* $T > \tau$.

Next, we state a result about the asymptotic behavior of the solutions of problem (18.24) when t goes to $+\infty$.

Let us suppose that c1)–c3) hold with $m \in L^\infty(\mathbb{R})$, assume also that

$$\nu^2 \lambda_1^2 (1 - \rho_*) > |m|_\infty,$$

where $|m|_\infty := \|m\|_{L^\infty(\mathbb{R})}$,
and let us denote by $\varepsilon > 0$ the unique solution of

$$\varepsilon - \nu\lambda_1 + \frac{|m|_\infty e^{\varepsilon h}}{\nu\lambda_1(1 - \rho_*)} = 0. \tag{18.28}$$

We can now formulate the following result (see also [18] for a similar result in the infinite delay case).

Theorem 18.9 ([5], Theorem 4.1). *Under the previous assumptions, for any* $(u^0, \phi) \in H \times L^2(-h, 0; H)$, *and any* $\tau \in \mathbb{R}$, *the corresponding solution* $u(t; \tau, u^0, \phi)$ *of problem (18.24) satisfies*

$$\begin{aligned} &\left|u(t; \tau, u^0, \phi)\right|^2 \\ &\leq \left(\left|u^0\right|^2 + \frac{|m|_\infty e^{\varepsilon h}}{\nu\lambda_1(1 - \rho_*)} \int_{-h}^{0} e^{\varepsilon s} \left|\phi(s)\right|^2 ds\right) e^{\varepsilon(\tau - t)} \\ &+ \frac{e^{-\varepsilon t}}{\nu\lambda_1} \int_\tau^t e^{\varepsilon s} f(s)\, ds, \end{aligned} \tag{18.29}$$

for all $t \geq \tau$.

In particular, if $\int_\tau^\infty e^{\varepsilon s} f(s)\, ds < \infty$, *then every solution* $u(t; \tau, u^0, \phi)$ *of (18.24) converges exponentially to* 0 *as* $t \to +\infty$.

Finally, the existence of pullback attractor is also proved in [5] by following a similar scheme to the one used in [4] for the two-dimensional Navier–Stokes equations with delay.

18.5 Statistical Solutions of GMNSE

The autonomous GMNSE with $\tau = 0$ and $f \in H$, i.e. f is independent of time t, are considered in this section and N is held fixed here. Let $S^{(N)}$ be the semigroup in V generated by the autonomous GMNSE and let $\mathscr{A}_N$ be its global attractor in V. Probability measures on H here are with respect to the σ-algebra of Borel subsets of H.

Definition 18.2. A probability measure on H is said to be $S^{(N)}$-invariant if

$$\mu(V) = 1 \quad \textit{and} \quad \mu(E) = \mu\left(S^{(N)}(t)^{-1}E\right), \quad \forall\, t \geq 0, \tag{18.30}$$

for every Borel subset E of V (recall that a Borel set in V is a Borel set in H).

Theorem 18.10 ([3], Theorem 10). *The support of any $S^{(N)}$-invariant measure on H is included in the global attractor $\mathscr{A}_N$.*

The existence of such measures is obtained by time averaging. The results below generalize those of Foias et al. [9] (see also Lukaszewicz [17]) for the two-dimensional NSE to the GMNSE.

18.5.1 Time-Averages Solutions in the Autonomous Case

Classical ergodic results yield the equivalence of the limit of time averages with a spatial average with respect to an invariant measure. An analogous results holds for the NSE and GMNSE, but, following Foias et al. [9], requires the use of a generalized limit in a Banach space (see [3] for the definition).

Let LIM denote a generalized limit on $\mathscr{B}([0,\infty))$, the space of all bounded real-valued functions on $[0,\infty)$.

Definition 18.3. A time-average measure of the solution $u(t)$ of the autonomous GMNSE is a probability measure μ on H such that $C(H) \subset L^1(H,\mu)$ and

$$\mathrm{LIM}_{T\to\infty}\frac{1}{T}\int_0^T \varphi(u(t))\,dt = \int_H \varphi(v)\,d\mu(v), \quad \forall\,\varphi \in C(H). \tag{18.31}$$

The following results from [3] show first that a time-average measure, if it exists, has the domain of the operator A as its domain and then that such measures do in fact exist and are invariant w.r.t. the semigroup $S^{(N)}$.

Proposition 18.1 ([3], Proposition 13). *Any time-average measure μ of a solution $u(t)$ of the autonomous GMNSE is carried by $D(A)$, i.e., $\mu(D(A)) = 1$.*

Proposition 18.2 ([3], Proposition 14). *For any solution $u(t)$ of the autonomous GMNSE such that $u(0) \in V$ there exists a time-average measure μ of this solution such that moreover $C(V) \subset L^1(H, \mu)$ and*

$$\mathrm{LIM}_{T\to\infty} \frac{1}{T} \int_0^T \varphi(u(t))\, dt = \int_H \varphi(v)\, d\mu(v) \quad \forall\, \varphi \in C(V). \tag{18.32}$$

Proposition 18.3 ([3], Proposition 16). *Let $u(t)$ be the solution of the autonomous GMNSE corresponding to $u_0 \in V$ and let μ be a time-average measure of $u(t)$ such that $C(V) \subset L^1(H, \mu)$ and (18.32) is satisfied for all $\varphi \in C(V)$. Then μ is an $S^{(N)}$-invariant measure.*

18.5.2 Stationary Statistical Solutions of the Autonomous GMNSE

Foias et al. [9] also investigated stationary statistical solutions of the autonomous NSE. Analogous definitions also hold for the GMSE with the operator B replaced by B_N.

Define

$$G_N(v) = -\nu A v - B_N(v, v) + f, \quad \forall\, v \in V, \tag{18.33}$$

and let $\mathscr{T}$ be the set of real valued functionals $\Phi = \Phi(v)$ on H such that

(i) $c_r := \sup_{|v|\le r} |\Phi(v)| < +\infty$ for all $r > 0$;

(ii) for any $v \in V$ there exists $\Phi'(v) \in V$ such that

$$\frac{|\Phi(v + w) - \Phi(v) - (\Phi'(v), w)|}{|w|} \to 0 \quad \text{as } |w| \to 0 \text{ with } w \in V; \tag{18.34}$$

(iii) the mapping $v \mapsto \Phi'(v)$ is continuous and bounded as function from V into V.

Definition 18.4. A stationary statistical solution of the GMNSE is a probability measure μ on H such that

(i) $\displaystyle\int_H \|v\|^2\, d\mu(v) < +\infty$;

(ii) $\displaystyle\int_H \langle G_N(v), \Phi'(v)\rangle\, d\mu(v) = 0$ for any $\Phi \in \mathscr{T}$;

(iii) $\displaystyle\int_{\{a\le|v|^2<b\}} \{\nu\|v\|^2 - (f, v)\}\, d\mu(v) \le 0$ for any $0 \le a < b \le +\infty$.

The following results were proved in [3] correspond to those in [9] for the autonomous NSE, namely that $S^{(N)}$-invariant probability measure and time-average measures are stationary statistical solutions of the autonomous GMNSE.

Theorem 18.11 ([3], Theorem 19). *Any $S^{(N)}$-invariant probability measure on H is a stationary statistical solution of the autonomous GMNSE.*

Corollary 18.2 ([3], Corollary 20). *Let μ be a time-average measure of a solution $u(t)$ of the GMNSE such that $C(V) \subset L^1(H,\mu)$ holds and (18.32) is satisfied for all $\varphi \in C(V)$. Then μ is a stationary statistical solution of the autonomous GMNSE.*

As partial counterpart of Theorem 18.11 was given in [18].

Theorem 18.12 ([13], Theorem 15). *Let μ be a stationary statistical solution of GMNSE such that there exists a bounded and measurable subset $\mathscr{B}_N$ of $D(A)$ satisfying $\mu(H \setminus \mathscr{B}_N) = 0$. Then μ is an S_N-invariant probability measure on H.*

18.6 Numerical Solution of the Globally Modified NSE

There is an extensive literature on the numerical analysis of the three-dimensional Navier–Stokes equations, much of which is based on the pioneering ideas of Temam [22]. In this spirit Deugoue and Djoko [7] investigated the implicit Euler scheme applied to the GMNSE, specifically

$$\frac{u^{m+1}-u^m}{k} + \nu A u^{m+1} + B_N(u^{m+1}, u^{m+1}) = f^{m+1} \tag{18.35}$$

with time stepsize k, where

$$f^{m+1} = \frac{1}{k}\int_{mk}^{(m+1)k} f(t)\,dt.$$

They establish uniform bounds on u^m (with respect to m) and its temporal difference quotient in different function spaces and find conditions under which u^m is continuous in N and u^0 and for which (18.35) is uniquely solvable. They also establish the existence of absorbing sets in both H and V spaces, which is the first step in showing the existence of attractors. Finally they consider the limit as $N \to \infty$ and prove the following theorem.

Theorem 18.13 ([7], Theorem 6.1). *Let $f, f' \in L_\infty(\mathbb{R}^+, H)$ and $v^0 \in D(A)$ with k sufficiently small. Then the sequence $\{u^{m,N}\}_N$ of solutions of the implicit Euler scheme (18.35) converges to a weak solution of the following time discrete three-dimensional Navier–Stokes equations*

$$\frac{1}{k}(u^{m+1}-u^m, w) + \nu(\nabla u^m, \nabla w) + b(u^m, u^m, w) = (f^m, w), \quad \text{for all } w \in V, \tag{18.36}$$

as $N \to \infty$.

18.7 Weak Solutions of the Three-Dimensional Navier–Stokes Equations

Useful results about the three-dimensional Navier–Stokes equations can be obtained from the GMNSE.

18.7.1 Weak Kneser Property of the Attainability Set of Weak Solutions

The Kneser property for ordinary differential equations says that the attainability set of the solutions emanating from a given initial value is compact and connected. This property was shown by Kloeden and Valero [14] in a combination of Corollary 3.2 and Theorem 3.3 to hold for the weak solutions of the GMNSE in the strong topology of space H before it was known that the weak solutions of the GMNSE for a given initial value were unique, which makes the result trivial. This result was then used in [14] to show that the attainability set of the weak solutions of the three-dimensional Navier–Stokes equations satisfying an energy inequality are weakly compact and weakly connected. A simplified proof, also using properties of the GMNSE, was later given in [15].

More precisely, for every initial datum $u_0 \in H$ it is well known that at least one weak solution of (18.1) exists such that

$$V_\tau(u(t)) \leq V_\tau(u(s)) \quad \text{for all } t \geq s, \text{ a.a. } s > \tau \text{ and } s = \tau, \tag{18.37}$$

where $V_\tau(u(t)) := \frac{1}{2}|u(t)|^2 + \nu \int_\tau^t \|u(r)\|^2 \, dr - \int_\tau^t (f(r), u(r)) \, dr$. Denote the corresponding attainability set for $t \geq \tau$ by

$$K_t(u_0) = \{u(t) \ : \ u(\cdot) \text{ is a weak solution of (18.1) satisfying (18.37)}\}.$$

We have:

Theorem 18.14 ([14], Theorem 2.1). *Let $f \in L^\infty(\tau, T; H)$ for all $T > \tau$. Then, for all $t \geq \tau$ and $u_0 \in H$, the attainability set $K_t(u_0)$ is compact and connected with respect to the weak topology on H.*

18.7.2 Convergence to Weak Solutions of the Three-Dimensional NSE

Theorem 18.15 ([1], Theorem 13). *Suppose that $f \in L^2(\tau, T; (L^2(\Omega))^3)$ for each $T > \tau$ and let $u^{(N)}(t)$ be a weak solution of the GMNSE (18.2) with the initial value $u_0^{(N)} \in H$, where $u_0^{(N)} \rightharpoonup u_0$ weakly in H as $N \to \infty$.*

Then, there exists a subsequence $\left\{u^{(N_j)}(t)\right\}$ *which converges as* $N_j \to \infty$, *weak-star in* $L^\infty(\tau, T; H)$, *weakly in* $L^2(\tau, T; V)$ *and strongly in* $L^2(\tau, T; H)$, *to a weak solution* $u(t)$ *on the interval* $[\tau, T]$ *of the NSE (18.1) with initial condition* u_0, *for every* $T > \tau$.

The proof is based on the fact that a weak solution of the GMNSE (18.2) with the initial value $u_0^{(N)} \in H$, where $u_0^{(N)} \rightharpoonup u_0$ weakly in H as $N \to \infty$, satisfies the energy inequality

$$\frac{d}{dt}|u^{(N)}|^2 + \nu\|u^{(N)}\|^2 \leq \frac{1}{\nu\lambda_1}|f|^2 \tag{18.38}$$

uniformly in $N > 0$. One easily obtains a convergent subsequence. The main difficulty is to show that limiting function is a weak solution of the NSE (18.1) for the given initial condition u_0, i.e. satisfies the variational equation (18.4) with b_N replaced by b. The following lemma is required here.

Lemma 18.2 ([1], Lemma 12). *For each* $p \geq 1$, *it follows that*

$$F_N\left(\|u^{(N)}(s)\|\right) \to 1 \qquad in \quad L^p(\tau, T; \mathbb{R}), \qquad as \quad N \to \infty.$$

18.7.3 Existence of Bounded Entire Weak Solutions of Three-Dimensional NSE

When the forcing term $f \in (L^2(\Omega))^3$ is independent of time, Theorem 18.15 and the existence of a global attractor $\mathscr{A}_N$ of the GMNSE (18.2) for each N can be used to show that the NSE (18.1) have bounded entire weak solutions, that is, weak solutions which exist and are bounded for all $t \in \mathbb{R}$. Such solutions are interesting as they would belong to a global attractor of the three-dimensional NSE, if such an attractor were to exist.

Theorem 18.16 ([1], Theorem 11). *Suppose that* $f \in (L^2(\Omega))^3$. *Then there exists a bounded entire weak solution of the NSE (18.1). More exactly, there exists a bounded entire weak solution of the NSE (18.1) with initial value* u_0 *for each* $u_0 \in \mathscr{U}_0$, *where* $\mathscr{U}_0$ *is the subset in* H *consisting of the weak* H*-cluster points of sequences* $u_0^{(N)} \in \mathscr{A}_N$ *for* $N \to \infty$.

The set $\mathscr{U}_0$ here is obviously a nonempty subset of the closed and bounded subset $\mathscr{B}_H$ of H. A similar result holds with essentially the same proof in the nonautonomous case, as well as for the GMNED analyzed in Sect. 18.4 (see [19] for more details).

Acknowledgments This work was partially supported by the Spanish Ministerio de Ciencia e Innovación project MTM2011-22411, the Consejería de Innovación, Ciencia y Empresa (Junta de Andalucía) under the Ayuda 2009/FQM314 and the Proyecto de Excelencia P07-FQM-02468.

References

1. Caraballo, T., Kloeden, P.E., Real, J.: Unique strong solutions and V-attractors of a three-dimensional system of globally modified Navier–Stokes equations. Adv. Nonlinear Stud. **6**, 411–436 (2006)
2. Caraballo, T., Kloeden, P.E., Real, J.: Addendum to the paper "Unique strong solutions and V-attractors of a three-dimensional system of globally modified Navier–Stokes equations". Adv. Nonlinear Stud. **6**, 411–436 (2006). Adv. Nonlinear Stud. **10**, 245–247 (2006)
3. Caraballo, T., Kloeden, P.E., Real, J.: Invariant measures and statistical solutions of the globally modified Navier–Stokes equations. Discrete Conts. Dynam. Syst. Ser. B **10**, 761–781 (2008)
4. Caraballo, T., Real, J.: Attractors for 2D-Navier–Stokes models with delays. J. Differ. Eqns. **205**, 270–296 (2004)
5. Caraballo, T., Real, J., Márquez, A.M.: Three-dimensional system of globally modified Navier–Stokes equations with delay. Int. J. Bifur. Chaos Appl. Sci. Eng. **20**, 2869–2883 (2010)
6. Constantin, P.: Near identity transformations for the Navier–Stokes equations. In: Handbook of Mathematical Fluid Dynamics, vol. II, pp. 117–141. North-Holland, Amsterdam (2003)
7. Deugoue, G., Djoko, J.K.: On the time discretization for the globally modified three-dimensional Navier–Stokes equations. J. Comput. Appl. Math. *235*(8), 2015–2029 (2011)
8. Flandoli, F., Maslowski, B.: Ergodicity of the 2-D Navier–Stokes Equation under random perturbations. Commun. Math. Phys. **171**, 119–141 (1995)
9. Foias, C., Manley, O., Rosa, R., Temam, R.: Navier–Stokes Equations and turbulence. Encyclopedia of Mathematics and Its Applications, vol. 83. Cambridge University Press, Cambridge (2001)
10. Kloeden, P.E., Caraballo, T., Langa, A., Real, J., Valero, J.: The 3-dimensional globally modified Navier–Stokes equations. In: Vasundhara Devi, J., Sivasundaram, S., Drici, Z., Mcrae, F. (eds.) Legacy of the Legend, Professor V. Lakshmikantham, vol. 3 Advances in Nonlinear Analysis: Theory, Methods and Applications, pp. 11–22. Cambridge Scientific Publishers, Cambridge (2009)
11. Kloeden, P.E., Langa, J.A.: Flattening, squeezing and the existence of random attractors. Proc. Roy. Soc. Lond. A **463**, 163–181 (2007)
12. Kloeden, P.E., Langa, J.A., Real, J.: Pullback V-attractors of a 3-dimensional system of nonautonomous globally modified Navier–Stokes equations: Existence and finite fractal dimension. Commun. Pure Appl. Anal. **6**, 937–955 (2007)
13. Kloeden, P.E., Marín-Rubio, P., Real, J.: Equivalence of invariant measures and stationary statistical solutions for the autonomous globally modified Navier–Stokes equations. Commun. Pure Appl. Anal. **8**, 785–802 (2009)
14. Kloeden, P.E., Valero, J.: The weak connectedness of the attainability set of weak solutions of the $3D$ Navier–Stokes equations. Proc. Roy. Soc. Lond. A **463**, 1491–1508 (2007)
15. Kloeden, P.E., Valero, J.: The Kneser property of the weak solutions of the three dimensional Navier–Stokes equations. Discrete Conts. Dyn. Syst. Ser. S **28**, 161–179 (2010)
16. Lions, J.L.: Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires. Dunod, Paris (1969)
17. Lukaszewicz, G.: Pullback attractors and statistical solutions for 2D Navier–Stokes equations. Discrete Contin. Dyn. Syst. Ser. B **9**, 643–659 (2008)
18. Marín-Rubio, P., Márquez-Durán, A.M., Real, J.: Three dimensional system of globally modified Navier–Stokes equations with infinite delays. Discrete Contin. Dyn. Syst. Ser. B **14**(2), 655–673 (2010)

19. Marín-Rubio, P., Márquez-Durán, A.M., Real, J.: On the convergence of solutions of the globally modified Navier–Stokes equations with delays to solutions of the Navier–Stokes equations with delays. Adv. Nonlinear Stud. **11**, 917–928 (2011)
20. Ma, Q., Wang, S., Zhong, C.: Necessary and sufficient conditions for the existence of global attractors for semigroups and applications. Indiana Univ. Math. J. **51**, 1541–1559 (2002)
21. Romito, M.: The uniqueness of weak solutions of the globally modified Navier–Stokes equations. Adv. Nonlinear Stud. **9**, 425–427 (2009)
22. Temam, R.: The Theory and Numerical Analysis of the Navier–Stokes Equations. North-Holland, Amsterdam (1977).

Chapter 19
Simulation of Hard Contacts with Friction: An Iterative Projection Method

Christoph Glocker

Abstract An algorithm for the simulation of mechanical systems with hard contacts and Coulomb friction is presented. The contact laws together with the switching rules between the different states of the contacts are formulated as normal cone inclusions and then generalized to contact-impact laws. By using proximal point methods, the combined contact-impact laws are rewritten as nonlinear equations and then iteratively solved within an implicit time integration algorithm. As illustrative examples, a funnel-deflector arrangement with 3,500 rigid balls and a bobsled simulator software is briefly presented. The theoretical setting is in accordance with classical Lagrangian mechanics and can be used to model various set-valued interaction laws in arbitrary finite-dimensional multibody systems.

19.1 Introduction

To understand the dynamics of real systems, models are needed in which the main effects responsible for a certain dynamic behavior are taken into account. These models have to be designed specifically, according to the needs of the associated real life problem. In dynamics, the models contain the so-called bodies that represent the inertia properties of the real system, but they also have to take into account the various force interactions between the bodies. These force interactions are classically represented by springs, dampers, and bilateral constraints. For many applications, however, the classical force elements are too limited to represent in a proper way hard stops, sprag clutches, and dry friction, as they occur in nearly every mechanism or machine. The reason for this deficiency is the ambivalent character of the said elements, which may act either as a force law allowing for displacements

C. Glocker (✉)
IMES - Center of Mechanics, ETH Zurich, CLA J23.1, Tannenstrasse 3, CH-8092 Zurich, Switzerland
e-mail: glocker@imes.mavt.ethz.ch

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_19,

or as a constraint preventing them. Despite the practical relevance to have proper modeling tools and numerical algorithms for the evaluation of such dry friction and impact processes, no commercial multibody simulation code exists up to date which would satisfyingly meet these demands. The constraint states of the above force elements are normally regularized by stiff springs or dampers, which may fundamentally alter the desired properties of perfect constraints and may even be combined with poor numerical schemes. For these reasons, an alternative approach is imperative, which allows for state switching and which deals with the constraints as they are.

Friction and impact phenomena have to be attributed to non-smooth dynamics. Research in this area has reached a status, where all the above-mentioned force elements are formulated as set-valued maximal monotone operators, and the evolution equations as measure differential inclusions. In addition, robust and accurate algorithms have been developed to solve the underlying inequality problems and to combine them with implicit time discretization schemes. By the methods as described in this paper, friction and impact elements can now be used in arbitrary number and any order to model much sharper such effects in dynamics.

The paper at hand is organized as follows: Normal cones to convex sets and their relation to proximal points are introduced in Sect. 19.2, and illustrated in Sect. 19.3 by the example of the exact regularization of the set-valued sign function. The exact regularization is the key to the iterative numerical treatment of *spatial* friction elements, as it allows to represent the friction laws by nonlinear equations, including all the various states as stick and slip and the transition rules between them. Based on the exact regularization, the iterative projection method as used for the dynamic simulation of Lagrangian systems with hard contacts and dry friction is explained in Sect. 19.4–19.9.

With the help of the principle of virtual power, it is shown in Sect. 19.4 and Sect. 19.5 how the contact forces are included in Lagrange's equations of second kind, and how the relative contact velocities have to be calculated. The associated contact laws are introduced in Sect. 19.6 and formulated as normal cone inclusions. In particular, models for bilateral and unilateral constraints are discussed as used in the normal contact direction, as well as force elements for planar and spatial Coulomb friction in the isotropic and orthotropic case.

In Sect. 19.7 we present how the equations of motion and the contact laws can be consistently extended to impacts by including, in addition to the Lebesgue integrable parts, the impulsive forces as Dirac point measures. The resulting measure differential inclusions are discretized in Sect. 19.8 by a midpoint rule for the positions and an Euler backward step for the velocities. The discretization of the velocities has to be performed implicitly to allow for switching actions in the considered time interval and requires the discretized inclusion problem to be solved, which is then described in Sect. 19.9. Two application problems, one with many frictional contacts and the other one with just a few contacts but real time capability, will finally be presented in Sect. 19.10.

19.2 The Normal Cone and Proximal Points

The *normal cone* $\mathscr{N}_{\mathscr{C}}(\mathbf{x})$ to a convex subset $\mathscr{C}$ of $\mathbb{R}^n$ at a point $\mathbf{x} \in \mathscr{C}$ consists by definition [23] of all vectors $\mathbf{y}$, which do not form an acute "angle" with any line segment emanating from $\mathbf{x}$ and arbitrary endpoint $\mathbf{x}^\star \in \mathscr{C}$,

$$\mathscr{N}_{\mathscr{C}}(\mathbf{x}) = \{\mathbf{y} \mid \mathbf{y}^{\mathsf{T}}(\mathbf{x}^\star - \mathbf{x}) \leq 0, \quad \forall \mathbf{x}^\star \in \mathscr{C}\}\,. \tag{19.1}$$

For example, if $\mathscr{C}$ is a subset of $\mathbb{R}^2$ as in Fig. 19.1 and $\mathbf{x}_1$ is a point on a smooth section of the boundary of $\mathscr{C}$, then the normal cone $\mathscr{N}_{\mathscr{C}}(\mathbf{x}_1)$ is formed by a half-line "orthogonal" to the boundary of $\mathscr{C}$ at $\mathbf{x}_1$, with its null element located at $\mathbf{x}_1$. In the case of a corner point as for $\mathbf{x}_2$, the associated normal cone $\mathscr{N}_{\mathscr{C}}(\mathbf{x}_2)$ is nonnegatively generated by the two half-lines orthogonal to the adjoining smooth sections of the boundary of $\mathscr{C}$. If a point lies in the interior of $\mathscr{C}$ as e.g. $\mathbf{x}_3$, then the normal cone consists of one element only which is null element, $\mathscr{N}_{\mathscr{C}}(\mathbf{x}_3) = \{0\}$. This important special case evidently follows from (19.1), as there are no restrictions on the various directions $\mathbf{x}^\star - \mathbf{x}$. Finally, note that $\mathscr{N}_{\mathscr{C}}(\mathbf{x})$ as defined in (19.1) is indeed a cone, because $r\,\mathbf{y} \in \mathscr{N}_{\mathscr{C}}(\mathbf{x})$ whenever $\mathbf{y} \in \mathscr{N}_{\mathscr{C}}(\mathbf{x})$ and $r \geq 0$.

As above, let $\mathscr{C} \subset \mathbb{R}^n$ be a convex set, and let $\mathbf{z}$ be an arbitrary point of $\mathbb{R}^n$. We denote by

$$\mathbf{x} = \mathrm{prox}_{\mathscr{C}}(\mathbf{z}) \tag{19.2}$$

the *proximal point* to $\mathbf{z}$ in the set $\mathscr{C}$, i.e. the point $\mathbf{x}$ that minimizes for given $\mathbf{z}$ the Euclidean distance $f(\mathbf{x}) = \|\mathbf{z} - \mathbf{x}\|$ under the constraint $\mathbf{x} \in \mathscr{C}$. Figure 19.1 shows three different points $\mathbf{z}_i$ together with their proximal points $\mathbf{x}_i \in \mathscr{C}$ according to (19.2).

One observes in the figure that the directed line element $\mathbf{z}_i - \mathbf{x}_i$ is always contained in the normal cone $\mathscr{N}_{\mathscr{C}}(\mathbf{x}_i)$. This property has been proven in [12] to apply for general situations and leads, after multiplication of the cone $\mathscr{N}_{\mathscr{C}}(\mathbf{x})$ by an arbitrary positive number $r > 0$, to the equivalence

$$\mathbf{x} = \mathrm{prox}_{\mathscr{C}}(\mathbf{z}) \quad \Leftrightarrow \quad (\mathbf{z} - \mathbf{x}) \in r\mathscr{N}_{\mathscr{C}}(\mathbf{x})\,. \tag{19.3}$$

The latter allows us to replace normal cone inclusions by *nonlinear equations*: By setting $\mathbf{y} := \frac{1}{r}(\mathbf{z} - \mathbf{x})$ and eliminating $\mathbf{z}$ from (19.3), one obtains

$$\mathbf{y} \in \mathscr{N}_{\mathscr{C}}(\mathbf{x}) \quad \Leftrightarrow \quad \mathbf{x} = \mathrm{prox}_{\mathscr{C}}(r\,\mathbf{y} + \mathbf{x})\,. \tag{19.4}$$

The nonlinear map $\mathrm{prox}_{\mathscr{C}} : \mathbb{R}^n \mapsto \mathscr{C} \subset \mathbb{R}^n$ is continuous, weakly contractive, and idempotent, and therefore well suited for numerical purposes. Note also that $\mathrm{prox}_{\mathscr{C}}(\mathbf{z})$ becomes the identity map when $\mathbf{z} \in \mathscr{C}$, as for the case $\mathbf{z}_3 = \mathbf{x}_3$ in Fig. 19.1.

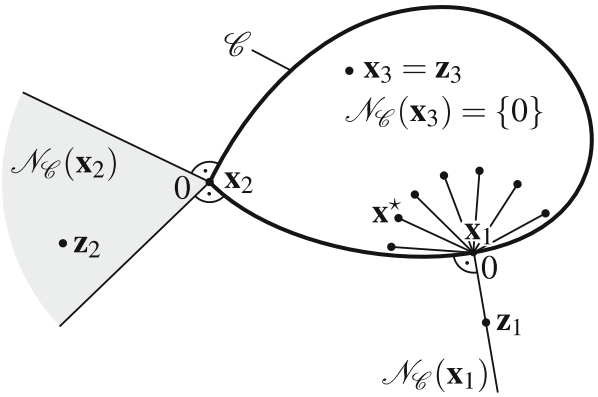

Fig. 19.1 Shown is the normal cone $\mathscr{N}_{\mathscr{C}}(\mathbf{x}_i)$ to a convex set $\mathscr{C} \subset \mathbb{R}^n$ at various points $\mathbf{x}_i \in \mathscr{C}$. These points determine the distance to the set $\mathscr{C}$ of some given points $\mathbf{z}_i$, i.e. $\mathbf{x}_i = \mathrm{prox}_{\mathscr{C}}(\mathbf{z}_i)$

19.3 Exact Regularization of the Set-Valued Sign Function

As an example, we consider the closed interval $\mathscr{C} := [-1, 1]$ as a convex subset of $\mathbb{R}$. The normal cone (19.1) at a chosen $x \in [-1, 1]$ is in this case determined by the variational inequality $\mathscr{N}_{[-1,1]}(x) = \{y \mid y\,(x^\star - x) \leq 0\}$, which has to hold for any and all $x^\star \in [-1, 1]$. For the evaluation of the normal cone, one has to distinguish between the three structurally different cases that x is in the interior or at the left or right boundary of the interval, see upper diagram of Fig. 19.2. The normal cone inclusion (19.4) therefore results in

$$y \in \mathscr{N}_{[-1,1]}(x) = \begin{cases} (-\infty, 0] & \text{for } x = -1 \\ \{0\} & \text{for } -1 < x < 1 \\ [0, +\infty) & \text{for } x = 1 \end{cases} . \tag{19.5}$$

The graph of the associated set-valued map $x \mapsto \mathscr{N}_{[-1,1]}(x)$ is depicted in the lower diagram of Fig. 19.2 and leads, after inversion, to the *set-valued sign function*

$$x \in \mathrm{Sgn}(y) = \begin{cases} -1 & \text{for } y < 0 \\ [-1, 1] & \text{for } y = 0 \\ 1 & \text{for } y > 0 \end{cases} , \tag{19.6}$$

which is shown in the upper diagram of Fig. 19.3. In order to determine now the proximal point function $\mathrm{prox}_{[-1,1]}(z)$, which assigns to each point z the point with minimal distance from z in the interval $[-1, 1]$, one has again to proceed in three steps: For $-1 \leq z \leq 1$, we have $z \in [-1, 1]$, and the closest point to z in $[-1, 1]$ is z itself. For $z > 1$, the closest point in the interval is $+1$, and for $z < -1$, the closest point is -1. Together, one obtains

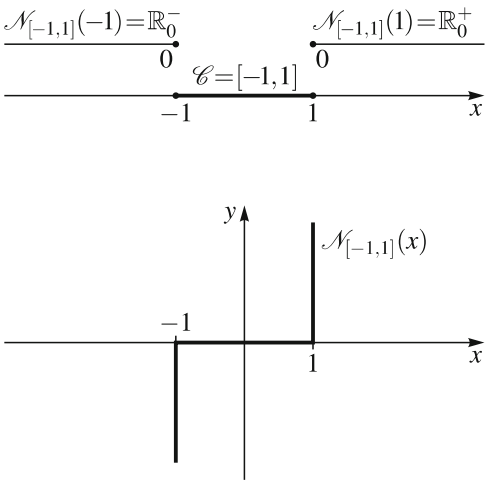

Fig. 19.2 The upper diagram shows the normal cone at the two boundary points of the interval $\mathscr{C} = [-1, 1]$. With $\mathcal{N}_{[-1,1]}(x) = \{0\}$ for the interior points of $\mathscr{C}$, the (maximal monotone) graph of the associated set-valued map $x \mapsto \mathcal{N}_{[-1,1]}(x)$ is depicted in the lower diagram

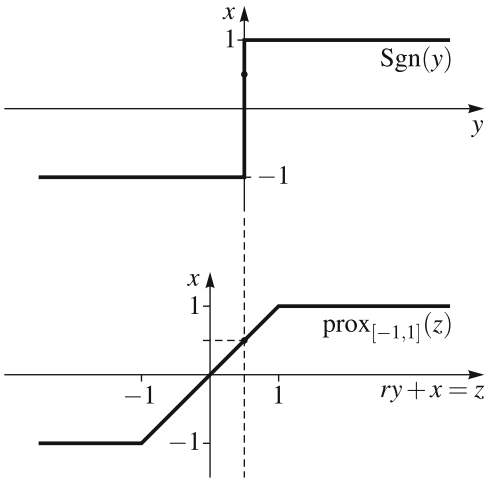

Fig. 19.3 The set-valued sign function $y \mapsto \text{Sgn}(y)$ as the inverse of the normal cone map $x \mapsto \mathcal{N}_{[-1,1]}(x)$ is depicted in the upper diagram. The lower diagram shows its exact regularization $z \mapsto \text{prox}_{[-1,1]}(z)$, where $z = ry + x$. The point $(y, x) = (0, \frac{1}{2})$ on the set-valued branch of the sign function corresponds to the point $(z, x) = (\frac{1}{2}, \frac{1}{2})$ of the prox function: As an interior point of $[-1, 1]$, one obtains for $z = \frac{1}{2}$ the identity map, i.e. $\text{prox}_{[-1,1]}(\frac{1}{2}) = \frac{1}{2}$

$$\text{prox}_{[-1,1]}(z) = \begin{cases} -1 \text{ for } z < -1 \\ z \quad \text{for } -1 \le z \le 1 \\ 1 \quad \text{for } z > 1 \end{cases} . \tag{19.7}$$

The graph of this function is depicted in the lower diagram of Fig. 19.3 and reminds of a regularized version of the set-valued sign function. However, note the different entities on the abscissae which is y for the sign function, and z for the prox function with $z = ry + x$ according to (19.3), (19.4). As a function of y, the graph of the prox function is horizontally stretched by the value of r, and in addition horizontally shifted by the value of x, such that a point (y, x) on the graph of the sign function

precisely ends up on the graph of the prox function. By applying the proximal point function as described above, the same values are obtained as if the original sign function would have been used. This method is therefore called an *exact regularization*. It is based on the equivalence (19.4), which in the case of the sign function becomes

$$y \in \mathscr{N}_{[-1,1]}(x) \quad \Leftrightarrow \quad x \in \mathrm{Sgn}(y) \quad \Leftrightarrow \quad x = \mathrm{prox}_{[-1,1]}(ry + x)\,. \tag{19.8}$$

The strategy to formulate and numerically evaluate normal cone inclusions via proximal point functions allows us even to skip the step (19.5) in which the normal cone is formulated. In most cases, already a geometric imagination of the normal cone is sufficient to directly write down the proximal point function (19.7), which is the only thing needed at the end.

19.4 Equations of Motion in Lagrangian Mechanics

In this section, we briefly review how non-potential forces have to be taken into account in the equations of motion in Lagrangian dynamics. We will need this approach in the next section when we introduce the contact forces, because they do in general not derive from a potential. We consider a mechanical system with n hard unilateral contacts under the influence of Coulomb friction. In the case that all contacts are open, the system is assumed to have f degrees of freedom, parameterized by a tuple of f local coordinates $\mathbf{q}$. The generalized velocities of the system are denoted by $\mathbf{u}$. As functions of time, $\mathbf{u}(t)$ are assumed to be of special bounded variations, leading to absolutely continuous positions $\mathbf{q}(t)$ with $\mathbf{u} = \dot{\mathbf{q}}$ almost everywhere. To derive the equations of motion for the impact-free case, the concept of virtual power δP together with Lagrange equations of second kind is used, which yields the variational expression

$$\forall \delta\mathbf{u}: \quad 0 = \delta P = \delta\mathbf{u}^{\mathsf{T}}\left[\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\partial T}{\partial \mathbf{u}}\right)^{\mathsf{T}} - \left(\frac{\partial T}{\partial \mathbf{q}}\right)^{\mathsf{T}} + \left(\frac{\partial V}{\partial \mathbf{q}}\right)^{\mathsf{T}} - \mathbf{f}_{NP}\right] \tag{19.9}$$

that is required to hold for all virtual velocities $\delta\mathbf{u}$. In (19.9), $T(\mathbf{q}, \mathbf{u}, t)$ denotes the kinetic energy and $V(\mathbf{q}, t)$ the potential energy of the system. All generalized forces which do not admit for a classical potential are collected in the vector $\mathbf{f}_{NP}$. After having carried out the differentiation process in (19.9), one obtains the equations of motion of the system, for which we impose the structure

$$\forall \delta\mathbf{u}: \quad 0 = \delta P = \delta\mathbf{u}^{\mathsf{T}}\left[\,\mathbf{M}(\mathbf{q}, t)\,\dot{\mathbf{u}} - \mathbf{h}(\mathbf{q}, \mathbf{u}, t) - \mathbf{f}\,\right]. \tag{19.10}$$

Here, $\mathbf{M}(\mathbf{q}, t)$ denotes the symmetric and positive definite mass matrix, which may be identified from $(\frac{\partial T}{\partial \mathbf{u}})^{\mathsf{T}} = \mathbf{M}(\mathbf{q}, t)\,\mathbf{u}$. All gyroscopic accelerations, i.e. the

Christoffel symbols, all potential forces, and all classical non-potential forces are collected in the term $\mathbf{h}(\mathbf{q}, \mathbf{u}, t)$, such that only the contact forces remain in the generalized force vector $\mathbf{f}$. These contact forces are introduced in the next section by setting up a suitable contact model, together with all the kinematic relations needed to finally evaluate the associated virtual power expression $\delta\mathbf{u}^{\mathsf{T}}\,\mathbf{f}$.

19.5 The Contact Model

Figure 19.4 shows the model that will be used to formulate hard contacts with Coulomb friction in the planar case. In the left part, some geometric entities are displayed as needed to set up the kinematics of the contact. We first discuss the normal direction. The *contact points* P and Q are by definition the points on the (convex) contours of the bodies that share a common normal. As soon as the contact points are known, one may specify an outward normal vector $\mathbf{n}(\mathbf{q}, t)$ of unit length at one of the bodies. The *relative velocity* γ_N of the contact points in the normal direction is then given by the difference of the absolute velocities $\mathbf{v}_Q$ and $\mathbf{v}_P$ of the rigid body points placed at Q and P, projected on the normal $\mathbf{n}$. This gives

$$\gamma_N = \mathbf{n}^{\mathsf{T}}(\mathbf{v}_Q - \mathbf{v}_P) = \mathbf{w}_N^{\mathsf{T}}\,\mathbf{u} + \chi_N\,, \tag{19.11}$$

where the two terms $\mathbf{w}_N(\mathbf{q}, t)$ and $\chi_N(\mathbf{q}, t)$ may directly be identified when generalized coordinates and velocities are used. Physically, $\gamma_N < 0$ corresponds with bodies that are approaching each other, whereas $\gamma_N > 0$ indicates that the bodies are moving apart from each other. In analogy with (19.11), one obtains for the virtual velocities

$$\delta\gamma_N = \mathbf{n}^{\mathsf{T}}\delta(\mathbf{v}_Q - \mathbf{v}_P) = \mathbf{w}_N^{\mathsf{T}}\,\delta\mathbf{u}\,. \tag{19.12}$$

Note that positions and time have not to be varied when building the virtual velocities, which leads directly to $\delta\mathbf{n} = 0$, because the normal $\mathbf{n}$ depends on $\mathbf{q}$ and t only, but not on any velocity.

The right part of Fig. 19.4 shows the free body diagram of the contact, with the contact forces decomposed into a normal and a tangential component each. By using the normal $\mathbf{n}$, the normal contact forces may be written as

$$\mathbf{F}_{NQ} = -\mathbf{F}_{NP} = \mathbf{n}\,\lambda_N\,, \tag{19.13}$$

where λ_N is the signed scalar value of these normal forces. The contribution δP_N of these forces to the overall virtual power (19.10) may now easily be calculated by the invariance of the virtual power under coordinate transformation. With the help of (19.12) and (19.13), one obtains

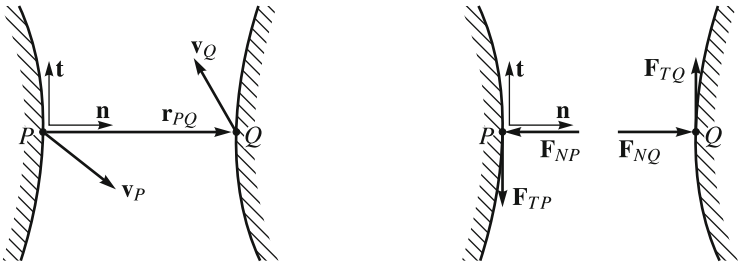

Fig. 19.4 Contact model for a planar situation. *Left diagram*: Contact points P and Q together with their absolute velocities $\mathbf{v}_p$ and $\mathbf{v}_Q$. *Right diagram*: Free body diagram, showing the contact forces in the normal and the tangential direction

$$\delta P_N = \mathbf{F}_{NQ}^{\mathsf{T}}\,\delta\mathbf{v}_Q + \mathbf{F}_{NP}^{\mathsf{T}}\,\delta\mathbf{v}_P = \lambda_N\,\mathbf{n}^{\mathsf{T}}(\delta\mathbf{v}_Q - \delta\mathbf{v}_P) = \lambda_N\,\mathbf{w}_N^{\mathsf{T}}\,\delta\mathbf{u} =: \mathbf{f}_N^{\mathsf{T}}\,\delta\mathbf{u}\,, \tag{19.14}$$

from which one identifies the generalized force associated with the normal contact direction as

$$\mathbf{f}_N = \mathbf{w}_N\,\lambda_N\,. \tag{19.15}$$

This force is composed of the *scalar normal force* λ_N and the *generalized force direction* $\mathbf{w}_N(\mathbf{q}, t)$, which we have already met in the expression for the normal relative velocity (19.11).

For the tangential contact direction, which is defined by the tangent unit vector $\mathbf{t}$ in Fig. 19.4, one proceeds in the same way as in (19.11) and (19.15). The tangential relative velocity is obtained as $\gamma_T = \mathbf{t}^{\mathsf{T}}(\mathbf{v}_Q - \mathbf{v}_P) = \mathbf{w}_T^{\mathsf{T}}\,\mathbf{u} + \chi_T$ and describes for a closed contact precisely the velocity with which the bodies are sliding against each other. The generalized tangential force results in $\mathbf{f}_T = \mathbf{w}_T\,\lambda_T$ and is in the case of sliding ($\gamma_T \neq 0$) the friction force. In the case of sticking, it is the constraint force that prevents any tangential motion of the boundaries of the bodies relative to each other.

For spatial friction, another tangent unit vector $\mathbf{b} = \mathbf{n} \times \mathbf{t}$ is needed, to project the velocities in the same way as above, and to set up another generalized friction force component by $\gamma_B = \mathbf{b}^{\mathsf{T}}(\mathbf{v}_Q - \mathbf{v}_P) = \mathbf{w}_B^{\mathsf{T}}\,\mathbf{u} + \chi_B$ and $\mathbf{f}_B = \mathbf{w}_B\,\lambda_B$, respectively. How all these terms have to be built is described in detail in [4, 22] for the planar case, and in [5] for spatial configurations. Even pivoting friction about the $\mathbf{n}$-axis can be included in quite the same way. In order to determine the generalized friction torque $\mathbf{f}_S = \mathbf{w}_S\,\lambda_S$, one just has to project the difference of the absolute angular velocities $\boldsymbol{\Omega}_Q, \boldsymbol{\Omega}_P$ of the two bodies on the common normal $\mathbf{n}$, i.e. $\gamma_S = \mathbf{n}^{\mathsf{T}}(\boldsymbol{\Omega}_Q - \boldsymbol{\Omega}_P) = \mathbf{w}_S^{\mathsf{T}}\,\mathbf{u} + \chi_S$.

In the most involved case of a spatial contact with pivoting friction, the equations of motion (19.10) together with the associated relative velocities (19.11) result in

$$\begin{gathered}\mathbf{M}\dot{\mathbf{u}} - \mathbf{h} - \mathbf{w}_N\,\lambda_N - (\mathbf{w}_T\,\lambda_T + \mathbf{w}_B\,\lambda_B + \mathbf{w}_S\,\lambda_S) = 0\,,\\ \gamma_N = \mathbf{w}_N^{\mathsf{T}}\,\mathbf{u} + \chi_N\,, \quad \gamma_T = \mathbf{w}_T^{\mathsf{T}}\,\mathbf{u} + \chi_T\,, \quad \gamma_B = \mathbf{w}_B^{\mathsf{T}}\,\mathbf{u} + \chi_B\,, \quad \gamma_S = \mathbf{w}_S^{\mathsf{T}}\,\mathbf{u} + \chi_S\,,\end{gathered} \tag{19.16}$$

where we have taken into account the special structure of the generalized contact forces according to (19.15).

For contact problems, the individual force components $\mathbf{w}_K\lambda_K$ in (19.16) only need to be taken into account if the contact is closed. This can be supervised by the so-called gap function $g_N(\mathbf{q}, t)$, which measures the signed distance of the contact points P and Q via

$$g_N = \mathbf{n}^{\mathsf{T}}\mathbf{r}_{PQ}\,, \tag{19.17}$$

see Fig. 19.4. Note that the displacement vector $\mathbf{r}_{PQ}$ and the always outward normal $\mathbf{n}$ are antiparallel to each other in the case of forbidden overlap, which is then indicated by values $g_N < 0$. An open contact is represented by $g_N > 0$, and a closed contact by $g_N = 0$, for which the two bodies just touch each other. Furthermore, it can be shown [4] that $\dot{g}_N = \gamma_N$. In other words, the changes in time of the signed distance agree with the normal relative velocity.

For situations with n contacts, we address the individual normal components N_j in increasing order by odd numbers j, and the associated tangential components $(T, B, S)_{j+1}$ by the subsequent even numbers $j + 1$. With the help of the *index set*

$$\mathscr{H}(\mathbf{q}, t) := \{\, j,\, j+1 \mid g_{Nj}(\mathbf{q}, t) = 0, \quad j = 1, 3, 5, \ldots, 2n-1 \,\}\,, \tag{19.18}$$

we may then easily address the normal and tangential components of the *closed* contacts, and (19.16) may be written for the multi-contact case as

$$\begin{gathered}\mathbf{M}\dot{\mathbf{u}} - \mathbf{h} - \sum_{i\in\mathscr{H}} \mathbf{W}_i\,\boldsymbol{\lambda}_i = 0\,,\\ \boldsymbol{\gamma}_i = \mathbf{W}_i^{\mathsf{T}}\mathbf{u} + \boldsymbol{\chi}_i\,.\end{gathered} \tag{19.19}$$

For an i which is odd and therefore related to a normal component, the matrix $\mathbf{W}_i$ consists of only one column $\mathbf{W}_i = (\mathbf{w}_N)_i \in \mathbb{R}^{f,1}$, and all the related vectors have only one entry, i.e. $\boldsymbol{\lambda}_i = (\lambda_N)_i \in \mathbb{R}$, $\boldsymbol{\gamma}_i = (\gamma_N)_i \in \mathbb{R}$, $\boldsymbol{\chi}_i = (\chi_N)_i \in \mathbb{R}$. For an i which is even and therefore related to the tangential components, one has for the most advanced case of a spatial contact with pivoting friction $\mathbf{W}_i = (\mathbf{w}_T\,\mathbf{w}_B\,\mathbf{w}_S)_i \in \mathbb{R}^{f,3}$ and $\boldsymbol{\lambda}_i = (\lambda_T\,\lambda_B\,\lambda_S)_i^{\mathsf{T}} \in \mathbb{R}^3$, $\boldsymbol{\gamma}_i = (\gamma_T\,\gamma_B\,\gamma_S)_i^{\mathsf{T}} \in \mathbb{R}^3$, $\boldsymbol{\chi}_i = (\chi_T\,\chi_B\,\chi_S)_i^{\mathsf{T}} \in \mathbb{R}^3$. If a spatial contact is modeled without pivoting friction, one just has to cancel all entries with index S to obtain $\mathbf{W}_i \in \mathbb{R}^{f,2}$ and $\boldsymbol{\lambda}_i \in \mathbb{R}^2$, $\boldsymbol{\gamma}_i \in \mathbb{R}^2$, $\boldsymbol{\chi}_i \in \mathbb{R}^2$. In the same way, a planar contact is obtained by additionally dropping the terms with index B, which gives $\mathbf{W}_i \in \mathbb{R}^{f,1}$ and $\boldsymbol{\lambda}_i \in \mathbb{R}$, $\boldsymbol{\gamma}_i \in \mathbb{R}$, $\boldsymbol{\chi}_i \in \mathbb{R}$.

19.6 Formulation of the Contact Laws by Normal Cone Inclusions

By equation (19.19), the contact forces of the closed contacts are included in the equations of motion, and all relative velocities required to describe the kinematics of the contacts are formulated. Equation (19.19), however, provides not yet a complete description of the dynamics of the system, because the *contact laws* that determine the values of the contact forces $\boldsymbol{\lambda}$ have not yet been specified. In this article, we discuss as possible contact laws the class of kinematic force laws of normal cone type. They connect the contact forces $\boldsymbol{\lambda}$ and the relative velocities $\boldsymbol{\gamma}$ by a *normal cone inclusion* of the form $\boldsymbol{\gamma} \in \mathscr{N}_{\mathscr{C}}(-\boldsymbol{\lambda})$. Depending on the particular choice of the sets $\mathscr{C}$, one obtains via (19.1) various normal cones, and hence force laws of different physical meaning. Six cases which are important for the modeling of contact problems are collected in Fig. 19.5. For additional cases, we refer to [5, 12], in which also normal cone inclusions on displacement level are discussed.

Figure 19.5a shows the case that the real numbers are chosen as the set $\mathscr{C}$. The associated normal cone inclusion results in arbitrary values for the force by an always vanishing relative velocity, which reflects the characteristic of a *bilateral kinematic constraint*. Applied on the normal direction of a currently closed contact $g_N = 0$, the condition of an always vanishing relative velocity $\gamma_N = 0$ means by $\dot{g}_N = \gamma_N \equiv 0$ that the contact will stay closed for all succeeding times, which is mechanically assured by an unbounded reservoir for the associated constraint force $\lambda_N \in \mathbb{R}$.

When taking the non-positive real numbers as the set $\mathscr{C}$, one obtains the force–velocity graph displayed in Fig. 19.5b. This graph is equivalently represented by the inequality-complementarity conditions $\gamma_N \geq 0$, $\lambda_N \geq 0$, $\gamma_N \lambda_N = 0$, see, e.g. [4], and constitutes a *unilateral kinematic constraint*. This means for the normal component of a currently closed contact $g_N = 0$ that it stays closed ($\dot{g}_N = \gamma_N \equiv 0$) as long as its normal force acts as a compressive magnitude ($\lambda_N > 0$). A lift-off ($\gamma_N > 0$), which shortly after leads to an open contact ($g_N > 0$) and therefore by (19.18) to a deletion from the equation (19.19), requires the normal force to vanish ($\lambda_N = 0$). This force law may therefore be used to describe a hard contact without adhesion when it stays closed or starts to open.

Figure 19.5c–f shows different types of Coulomb friction as used to model effects from dry friction in the contact. Note that the size of the set $\mathscr{C}$ depends for all these cases on the normal force, which is *unknown* in general. For *one-dimensional Coulomb friction* as in planar contacts, the reservoir of transferable tangential forces $-\lambda_T$ is taken as $\mathscr{C} = [-\mu\,\lambda_N, +\mu\,\lambda_N]$, where μ denotes the Coulomb friction coefficient. The force–velocity characteristic shown in Fig. 19.5c corresponds with the horizontally stretched graph of the inverted set-valued sign function from Fig. 19.2. Sliding in positive direction ($\gamma_T > 0$) requires the friction force to be $\lambda_T = -\mu\lambda_N$, whereas sliding in negative direction ($\gamma_T < 0$) changes its sign to $\lambda_T = +\mu\lambda_N$.

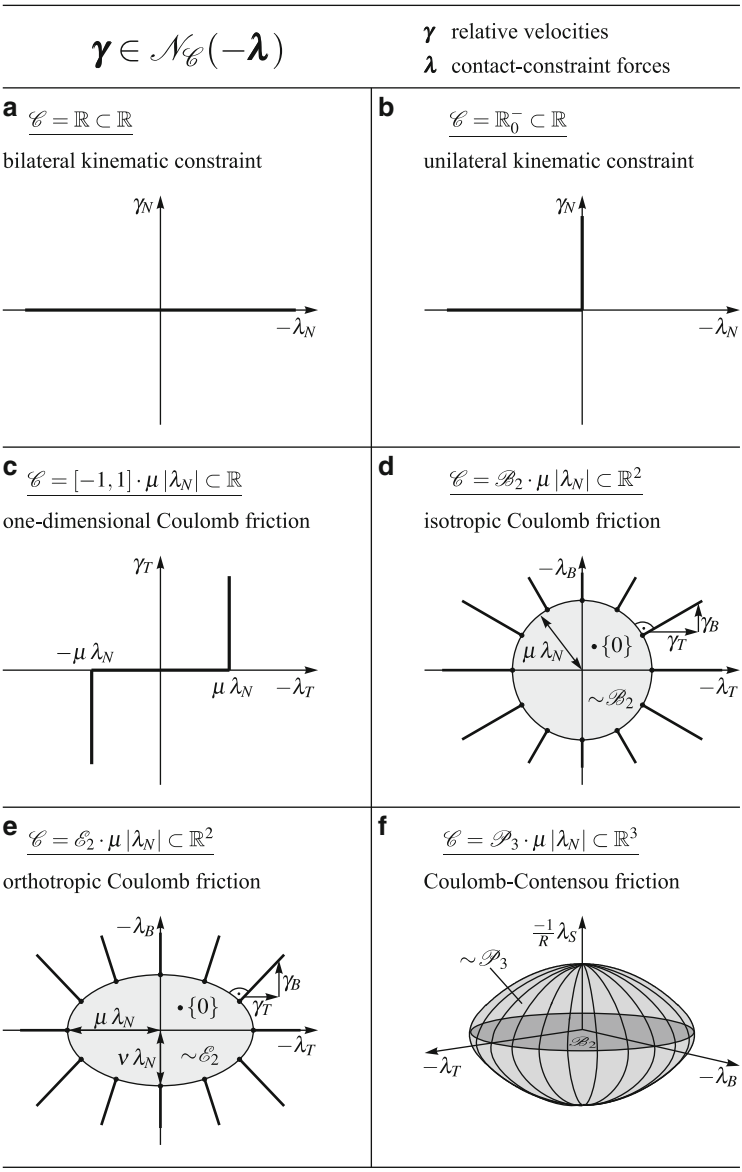

Fig. 19.5 Various kinematic force laws of normal cone type

Isotropic Coulomb friction as it occurs in spatial contacts may easily be represented when taking the set $\mathscr{C}$ of admissible tangential forces $(-\lambda_T, -\lambda_B)$ as the unit disk stretched by the value $\mu\,\lambda_N$. The normal cone law depicted in Fig. 19.5d yields stick ($\gamma_T = \gamma_B = 0$) as long as the friction force $(-\lambda_T, -\lambda_B)$

is in the interior of $\mathscr{C}$. It allows for slip in the direction opposing the friction force, if the latter reaches the boundary of $\mathscr{C}$.

A possible model for *orthotropic Coulomb friction* with a friction coefficient depending on the sliding direction is shown in Fig. 19.5e. The set $\mathscr{C}$ of admissible tangential forces is here chosen to be an ellipse with the larger principal axis being of length 1, and the ratio of the principal axes being equal the ratio of the two friction coefficients μ and ν. Note that the sliding direction as produced by the normal cone inclusion law is no longer collinear with the friction force, a fact that has been observed also in various experiments.

Pivoting friction due to differences in angular velocity of the contacting bodies must in general not be modeled separately. There is a mutual interference between linear friction and pivoting friction, known as the *Contensou effect*, that can easily be visualized on floor polishing machines: As faster the brush disk of such a machine rotates, as smoother can it be moved along the floor. The maximal transferable friction forces and friction torques influence each other and lead to a three-dimensional set $\mathscr{C}$ that is displayed in Fig. 19.5f. A detailed derivation of the normal cone law for such cases, together with an exhaustive physical explanation of this effect may be found in [6, 11].

In order to complete now equation (19.19) for the impact-free motion, we choose for each contact two normal cone inclusions, to state the desired contact laws: For each normal direction, i.e. for each odd i, one takes a characteristic according to Fig. 19.5a or 19.5b, and for each even i one chooses a friction law from Fig. 19.5c–f. This finally leads to

$$\begin{aligned} \mathbf{M}\dot{\mathbf{u}} - \mathbf{h} - \sum_{i\in\mathscr{H}} \mathbf{W}_i \boldsymbol{\lambda}_i = 0\,, \\ \boldsymbol{\gamma}_i = \mathbf{W}_i^{\mathsf{T}}\mathbf{u} + \boldsymbol{\chi}_i\,, \quad \boldsymbol{\gamma}_i \in \mathscr{N}_{\mathscr{C}_i}(-\boldsymbol{\lambda}_i)\,, \end{aligned} \tag{19.20}$$

which completes by the specific choices of the sets $\mathscr{C}_i$ according to Fig. 19.5 the formulation of the impact-free dynamics as second order differential inclusions.

19.7 Embedding Impact Dynamics and the Impact Laws

Transitions from sliding to sticking in systems with dry friction cause the accelerations to be discontinuous, and to no longer be defined at such transition points as a consequence. Strictly speaking, one therefore has to understand $\dot{\mathbf{u}}$ in (19.20) as the right derivative of the generalized velocities $\mathbf{u}$, and $\boldsymbol{\lambda}_i$ as the right limit of the contact forces, if one wants (19.20) to describe the dynamics towards future events [5]. The situation gets even more involved in the case of impacts. Impacts lead to impulsive forces and to velocity jumps, which are not at all taken into account in equation (19.20). In order to treat the impacts, one classically integrates the equations of motion over the instant of impact and processes the impacts separately. We basically proceed in the same way, but with the difference that the equations

of motion are integrated over an entire time interval of *arbitrary* length, to have in addition the impact-free motion included in the resulting representation. By carrying out this approach, one obtains from (19.20)

$$\begin{aligned} \mathbf{M}\,\mathrm{d}\mathbf{u} - \mathbf{h}\,\mathrm{d}t - \sum_{i\in\mathscr{H}} \mathbf{W}_i\,\mathrm{d}\boldsymbol{\Lambda}_i = 0\,, \\ \boldsymbol{\gamma}_i^{\pm} = \mathbf{W}_i^{\mathsf{T}}\mathbf{u}^{\pm} + \boldsymbol{\chi}_i\,, \quad (1+\varepsilon_i)\boldsymbol{\gamma}_i^{-} + \mathrm{d}\boldsymbol{\gamma}_i \in \mathscr{N}_{\mathrm{d}\mathscr{A}_i}(-\mathrm{d}\,\boldsymbol{\Lambda}_i)\,, \end{aligned} \tag{19.21}$$

where the individual terms have the following meaning: We explicitly allow for discontinuities in the generalized velocities $\mathbf{u}(t)$. For each time t, we denote by $\mathbf{u}^+(t)$ then *right limit* and by $\mathbf{u}^-(t)$ the *left limit* of $\mathbf{u}(t)$. If $\mathbf{u}^+(t) = \mathbf{u}^-(t)$, then $\mathbf{u}(t)$ is continuous at t. If, however, $\mathbf{u}^+(t) \neq \mathbf{u}^-(t)$, then t addresses a discontinuity point of $\mathbf{u}(t)$, at which $\mathbf{u}(t)$ may not even be defined, but only its right and left limit. By choosing $\mathbf{u}(t)$ as a function of *special bounded variation*, the existence of the limits $\mathbf{u}^{\pm}(t)$ is assured for every t, and a countably infinite number of discontinuities as needed in dynamics to cope with accumulation points of impacts is provided as well. The generalized coordinates $\mathbf{q}(t)$ are, as functions of time, obtained by integration of $\mathbf{u}(t)$. As a consequence, they are continuous, even *absolutely continuous*, which excludes position jumps. The discontinuities in $\mathbf{u}(t)$ are carried over to the relative velocities $\boldsymbol{\gamma}_i(t)$ via the second equation in (19.20), which is taken into account by the notation of left and right limits $(\pm)$ in the second equation of (19.21). The first equation in (19.21) has to be understood as the time integral of the equations of motion in (19.20). Mathematically, $\mathrm{d}\mathbf{u}$ denotes the *differential measure* [15] associated with $\mathbf{u}$, which contains in addition to the $\mathrm{d}t$-integrable parts also Dirac point measures, which produce the discontinuities in $\mathbf{u}(t)$ at integration, $\int_{\{t\}} \mathrm{d}\mathbf{u} = \mathbf{u}^+(t) - \mathbf{u}^-(t)$. In the same fashion, the finite forces $\boldsymbol{\lambda}_i$ in (19.20) have to be replaced by their associated *force measures* $\mathrm{d}\boldsymbol{\Lambda}_i$, which contain as Dirac measures the impulsive forces needed at the impacts to produce the velocity jumps. The maybe most challenging task is the consistent generalization of the normal cone inclusion to impacts, as displayed in the last equation of (19.21). This is certainly one of the central points in non-smooth dynamics. *Each* of the normal cone laws is equipped with a restitution coefficient ε_i $(0 \leq \varepsilon_i \leq 1)$, the physical meaning of which being discussed later. In the form as written in (19.21), it still contains the normal cone law for the impact-free motion from (19.20), but it determines by an *additionally* implemented impact law how the dynamics evolves at an impact. In the next section, a time discretization algorithm based on (19.21) will be presented which simultaneously processes the impact-free motion and the impacts. In order to show that (19.21) indeed contains both the impact-free motion according to (19.20) and the impacts, we will extract both cases from (19.21).

For the *impact-free motion*, the velocities are continuous. One therefore has $\mathbf{u}^+ = \mathbf{u}^- = \mathbf{u}$, $\boldsymbol{\gamma}_i^+ = \boldsymbol{\gamma}_i^- = \boldsymbol{\gamma}_i$, and there are no more contributions of Dirac measures in the differential measures $\mathrm{d}\mathbf{u}$, $\mathrm{d}\boldsymbol{\gamma}_i$, $\mathrm{d}\boldsymbol{\Lambda}_i$, $\mathrm{d}\mathscr{A}_i$. The latter can therefore be expressed by their $\mathrm{d}t$-integrable densities as $\mathrm{d}\mathbf{u} = \dot{\mathbf{u}}\,\mathrm{d}t$, $\mathrm{d}\boldsymbol{\gamma}_i = \dot{\boldsymbol{\gamma}}_i\,\mathrm{d}t$, $\mathrm{d}\boldsymbol{\Lambda}_i = \dot{\boldsymbol{\Lambda}}_i\,\mathrm{d}t = \boldsymbol{\lambda}_i\,\mathrm{d}t$, $\mathrm{d}\mathscr{A}_i = \dot{\mathscr{A}}_i\,\mathrm{d}t = \mathscr{C}_i\,\mathrm{d}t$, and (19.21) becomes

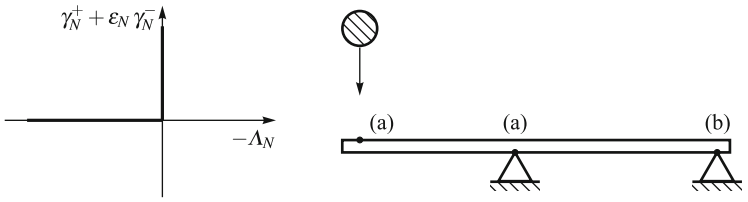

Fig. 19.6 The Newtonian impact law in inequality form for the normal direction: The left diagram shows the graph of the impact law. In the right diagram, a multi-impact configuration is depicted, in which the contacts (**a**) behave according to $\gamma_N^+ = -\varepsilon_N \gamma_N^-$, whereas contact (**b**) does not contribute to the impact by $\Lambda_N = 0$

$$\begin{aligned} \mathbf{M}\,\dot{\mathbf{u}}\,\mathrm{d}t - \mathbf{h}\,\mathrm{d}t - \sum_{i\in\mathscr{H}} \mathbf{W}_i\,\boldsymbol{\lambda}_i\,\mathrm{d}t = 0\,, \\ \boldsymbol{\gamma}_i = \mathbf{W}_i^{\mathsf{T}}\mathbf{u} + \boldsymbol{\chi}_i\,, \quad (1+\varepsilon_i)\boldsymbol{\gamma}_i + \dot{\boldsymbol{\gamma}}_i\,\mathrm{d}t \in \mathscr{N}_{\mathscr{C}_i\,\mathrm{d}t}(-\boldsymbol{\lambda}_i\,\mathrm{d}t)\,. \end{aligned} \tag{19.22}$$

The second equation in (19.21) is already verified. The first equation is obtained from (19.22) by factoring out $\mathrm{d}t$ and the fact that (19.22) has to hold for *any* time interval. We finally consider the normal cone inclusion in (19.22). For smaller and smaller integration intervals, the term $\int \dot{\boldsymbol{\gamma}}_i\,\mathrm{d}t$ tends to zero. Furthermore, the set $\mathscr{C}_i\,\mathrm{d}t$ and the force $-\boldsymbol{\lambda}_i\,\mathrm{d}t$ scale down with the very same factor, such that one obtains in the limit $(1+\varepsilon_i)\boldsymbol{\gamma}_i \in \mathscr{N}_{\mathscr{C}_i}(-\boldsymbol{\lambda}_i)$. Finally, since the right-hand side of this inclusion is a cone, one may divide by $(1+\varepsilon_i) > 0$ to get the normal cone inclusion precisely in the form as in (19.20).

In order to extract the *impact dynamics* from (19.21), we integrate (19.21) over just a singleton $\{t\}$. With $\int_{\{t\}} \mathrm{d}t = 0$, $\int_{\{t\}} \mathrm{d}\mathbf{u} = \mathbf{u}^+ - \mathbf{u}^-$, $\int_{\{t\}} \mathrm{d}\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_i^+ - \boldsymbol{\gamma}_i^-$, $\int_{\{t\}} \mathrm{d}\boldsymbol{\Lambda}_i = \boldsymbol{\Lambda}_i$, $\int_{\{t\}} \mathrm{d}\mathscr{A}_i = \mathscr{A}_i$, one obtains

$$\begin{aligned} \mathbf{M}(\mathbf{u}^+ - \mathbf{u}^-) - \sum_{i\in\mathscr{H}} \mathbf{W}_i\,\boldsymbol{\Lambda}_i = 0\,, \\ \boldsymbol{\gamma}_i^{\pm} = \mathbf{W}_i^{\mathsf{T}}\mathbf{u}^{\pm} + \boldsymbol{\chi}_i\,, \quad \boldsymbol{\gamma}_i^+ + \varepsilon_i\,\boldsymbol{\gamma}_i^- \in \mathscr{N}_{\mathscr{A}_i}(-\boldsymbol{\Lambda}_i)\,. \end{aligned} \tag{19.23}$$

The first equation, called the *impact equation*, is the equation of motion for the impact. It specifies the total impulsive force that is necessary to realize the velocity jump $\mathbf{u}^+ - \mathbf{u}^-$. The second equation gives the kinematic relation between the generalized velocities and the local contact velocities and holds in the same form as for the impact-free motion. The third relation constitutes in the form of normal cone inclusions the force laws for the impact, which are called the *impact laws* and which we want to discuss at least for the normal component of the contact.

By choosing the non-positive real numbers as the set $\mathscr{A}$ in the above normal cone inclusion, one obtains the impact law associated with the contact law of Fig. 19.5b. The graph of this impact law is depicted in the left diagram of Fig. 19.6 and consists of two branches which correspond to the following impact cases: For $-\Lambda_N < 0$, it holds that $\gamma_N^+ + \varepsilon_N\,\gamma_N^- = 0$. We call this the *regular* case, for which the value of

the normal relative velocity is inverted according to the classical Newtonian impact law under a compressive impulsive force in normal direction,

$$\Lambda_N > 0 \quad \Rightarrow \quad \gamma_N^+ = -\varepsilon_N \gamma_N^- . \tag{19.24}$$

The second branch allows for $-\Lambda_N = 0$ arbitrary values $\gamma_N^+ + \varepsilon_N \gamma_N^- > 0$ and is called the *exceptional* case,

$$\Lambda_N = 0 \quad \Rightarrow \quad \gamma_N^+ \geq -\varepsilon_N \gamma_N^- . \tag{19.25}$$

The exceptional case occurs if a closed contact does not at all participate in the impact ($\Lambda_N = 0$), and could therefore be removed from the equations (19.23) without changing the post-impact results. According to (19.25), all values for the post-impact relative velocity are admitted in the exceptional case that are larger than those prescribed by Newton's impact law for the regular case. The right part in Fig. 19.6 shows an impact configuration for which both the regular and the exceptional case occur: A rigid bar is resting on two unilateral obstacles and is hit by a rigid sphere. We expect an impulsive reaction not only between the sphere and the bar but also between the bar and the left obstacle. In contrast, the contact between the bar and the right obstacle is expected to open after the collision with a strictly positive relative velocity. This separation is solely induced by the overall impact configuration, and not supported by any local impulsive force.

As for the contact laws normal direction, restitution coefficients are also used in the friction characteristics. They are needed to model partly reversible impact processes in tangential direction, as they occur, for example, at highly elastic super balls when thrown under a table [7]. Even for bilateral kinematic constraints (Fig. 19.5a), it might be of advantage to equip them with restitution coefficients. Although useless and superfluous from the theoretical point of view, it allows in the numerical schemes to switch nifty from a projection algorithm with $\varepsilon_N = 0$ to a much less dissipative and drift-sensitive reflection method by taking $\varepsilon_N = 1$.

In the same way as the force laws, the impact laws in (19.23) are of local nature and only describe the interaction of the bodies connected by them. Impact events that are based on distant effects as, e.g., Newton's cradle with five balls [7], cannot be modeled by the presented approach. The concept of local restitution coefficients as introduced above is too limited to parameterize the entire the set of kinematically and kinetically admissible post-impact velocities. An extension to nonlocal interactions for the frictionless case can be found in [1, 7, 8].

19.8 Time Discretization

The development of reliable and accurate numerical integration routines for measure differential inclusions is today one of the central research topics in non-smooth dynamics. The so-called *time stepping methods*, the first of which having been the

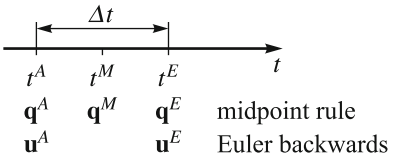

Fig. 19.7 Discretization scheme in Moreau's time-stepping algorithm: The positions $\mathbf{q}$ are processed by a midpoint rule, the velocities $\mathbf{u}$ by an Euler backwards step

midpoint rule introduced by Moreau [14], are difference schemes, which allow for a combined evaluation of the impacts and the impact-free motion, and which are designed such that all required inequality laws are taken into account. In contrast to the event-driven integration, at which a potential switching action is placed at the end of the time increments, the time stepping schemes collect them in the interior of the time intervals, such that multiple switching actions and even infinite switching sequences can be processed within one single increment.

In this section, we briefly review the discretization scheme as originally introduced in [14], and how to apply it to numerically approximate the *measure differential inclusions* (19.21). Within this scheme, each time step is evaluated by a midpoint rule for the positions, and an Euler backwards step for the velocities as indicated in Fig. 19.7. In order to process one time step, the following problem has to be solved: For a given start time t^A and given initial positions $\mathbf{q}^A := \mathbf{q}(t^A)$ and velocities $\mathbf{u}^A := \mathbf{u}(t^A)$, determine approximations of the positions $\mathbf{q}^E := \mathbf{q}(t^E)$ and velocities $\mathbf{u}^E := \mathbf{u}(t^E)$ at the end t^E of a chosen time interval $[t^A, t^E]$. For the evaluation of the midpoint rule, one has to proceed as follows, see [16] for additional comments:

1. Choose a time step Δt and calculate the midpoint $t^M := t^A + \frac{1}{2}\Delta t$ and the endpoint $t^E := t^A + \Delta t$ of the time interval.
2. Calculate the midpoint positions as $\mathbf{q}^M := \mathbf{q}^A + \frac{1}{2}\Delta t \cdot \mathbf{u}^A$.
3. Matrix calculation: Calculate $\mathbf{M}(\mathbf{q}^M, t^M)$ and $\mathbf{h}(\mathbf{q}^M, \mathbf{u}^A, t^M)$. According to (19.18), calculate the index set $\mathscr{H}(\mathbf{q}^M, t^M) := \{\, j,\, j+1 \mid g_{Nj}(\mathbf{q}^M, t^M) \leq 0 \,\}$, consisting of all closed and (numerically) interpenetrated contacts. For all $i \in \mathscr{H}(\mathbf{q}^M, t^M)$, calculate $\mathbf{W}_i(\mathbf{q}^M, t^M)$ and $\boldsymbol{\chi}_i(\mathbf{q}^M, t^M)$.
4. Calculate $\mathbf{u}^E$ from the discretized version of the inclusion problem (19.21),

$$\begin{gathered} \mathbf{M}(\mathbf{u}^E - \mathbf{u}^A) - \mathbf{h}\,\Delta t - \sum_{i \in \mathscr{H}} \mathbf{W}_i \boldsymbol{\Lambda}_i = 0\,, \\ \boldsymbol{\gamma}_i^A = \mathbf{W}_i^{\mathsf{T}} \mathbf{u}^A + \boldsymbol{\chi}_i\,, \quad \boldsymbol{\gamma}_i^E = \mathbf{W}_i^{\mathsf{T}} \mathbf{u}^E + \boldsymbol{\chi}_i\,, \\ \boldsymbol{\gamma}_i^E + \varepsilon_i \boldsymbol{\gamma}_i^A \in \mathscr{N}_{\mathscr{A}_i}(-\boldsymbol{\Lambda}_i)\,. \end{gathered} \tag{19.26}$$

   The terms $\mathrm{d}\mathbf{u}$, $\mathrm{d}\boldsymbol{\gamma}_i$, $\mathrm{d}\boldsymbol{\Lambda}_i$, $\mathrm{d}\mathscr{A}_i$, $\mathrm{d}t$ in the original problem (19.21) have been approximated by $\mathbf{u}^E - \mathbf{u}^A, \boldsymbol{\gamma}_i^E - \boldsymbol{\gamma}_i^A, \boldsymbol{\Lambda}_i, \mathscr{A}_i, \Delta t$ in the discretized version (19.26). Furthermore, the two upper indices $+$ and $-$ have been replaced by the end time E and start time A. One inequality solver that can be used to determine $\mathbf{u}^E$ in (19.26) is described in Sect. 19.9.
5. Calculate the endpoint positions as $\mathbf{q}^E := \mathbf{q}^M + \frac{1}{2}\Delta t \cdot \mathbf{u}^E$.

The above discretization scheme is of first order, and therefore requires relatively small time steps to achieve a reasonable accuracy. Integrators based on an extrapolation method with variable step size and order adjustment have recently been developed [28], as well as energy preserving codes [13]. Further discretization schemes are the Θ-method [10] which is a refined midpoint rule, an algorithm on position level with proven convergence [20, 21], as well as some other well-developed variants of the midpoint rule [3, 24, 25].

19.9 Numerical Solution of the Inclusion Problem

The inclusion problem (19.26), from which $\mathbf{u}^E$ is finally determined, can numerically be solved in various ways. In [16], for example, the forces in each individual contact are updated sequentially by cycling through all contacts until convergence is obtained. Another method, originally introduced in [18], splits the entire contact problem into an overall normal subproblem and another overall tangential subproblem, and solves both problems in turn by methods from optimization theory until convergence. The variational expressions and minimization problems associated with the inclusion problem (19.26) can be found for dynamic systems in [4, 5]. Another powerful approach, again from optimization theory, is the *augmented Lagrangian method*, which has been introduced in mechanics by [2] and successfully applied to inequality problems in contact mechanics. The augmented Lagrangian method yields at the end the same set of equations (19.33) as the proximal point method presented here. The latter, however, is even more general, because no underlying optimization problem is required. A further possibility is to rewrite (19.26) as a complementarity problem in standard form and to process it with an adequate solver. Such formulations are very involved and expensive for the spatial case [5], but they are still used for planar situations [9], for which the complementarity problem becomes linear. Several formulations are available as the ones in [19, 25], together with proofs on the existence and uniqueness of solutions.

Before we start to formulate the inclusion problem (19.26) as a system of proximal point equations, we want not only to comment a bit on its structure in view of its solvability but also to reveal the combinatorial problem hidden in such coupled contact problems. In a first step, we solve the first equation in (19.26) for $\mathbf{u}^E$

$$\mathbf{u}^E = \mathbf{u}^A + \mathbf{M}^{-1}\mathbf{h}\,\Delta t + \sum_{i\in\mathscr{H}} \mathbf{M}^{-1}\mathbf{W}_i\,\boldsymbol{\Lambda}_i\,, \tag{19.27}$$

and put it in the term $\boldsymbol{\gamma}_i^E + \varepsilon_i\,\boldsymbol{\gamma}_i^A$ built from the second line of (19.26), which gives

$$\boldsymbol{\gamma}_i^E + \varepsilon_i\,\boldsymbol{\gamma}_i^A = \mathbf{W}_i^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{W}_i\,\boldsymbol{\Lambda}_i + \sum_{k\neq i} \mathbf{W}_i^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{W}_k\,\boldsymbol{\Lambda}_k + \mathbf{W}_i^{\mathsf{T}}\mathbf{M}^{-1}\mathbf{h}\,\Delta t + (1+\varepsilon_i)\,\boldsymbol{\gamma}_i^A\,. \tag{19.28}$$

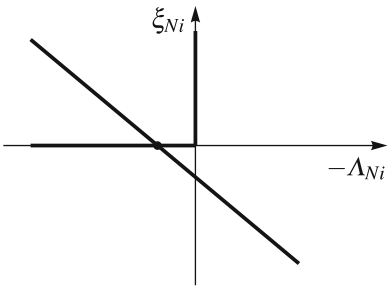

Fig. 19.8 Unique intersection of the straight line $\xi_{Ni} = g_{ii}\,\Lambda_{Ni} + \hat{c}_i$ and the inclusion $\xi_{Ni} \in \mathscr{N}_{\mathbb{R}_0^-}(-\Lambda_{Ni})$

By using the abbreviations

$$\boldsymbol{\xi}_i := \boldsymbol{\gamma}_i^E + \varepsilon_i \boldsymbol{\gamma}_i^A\,, \quad \mathbf{G}_{ij} := \mathbf{W}_i^{\mathsf{T}} \mathbf{M}^{-1} \mathbf{W}_j\,, \quad \mathbf{c}_i := \mathbf{W}_i^{\mathsf{T}} \mathbf{M}^{-1} \mathbf{h}\,\Delta t + (1+\varepsilon_i)\,\boldsymbol{\gamma}_i^A\,, \tag{19.29}$$

equation (19.28) can be compactly stated together with the normal cone inclusion from the third line of (19.26) as

$$\boldsymbol{\xi}_i = \mathbf{G}_{ii}\,\boldsymbol{\Lambda}_i + \sum_{k \neq i} \mathbf{G}_{ik}\,\boldsymbol{\Lambda}_k + \mathbf{c}_i\,, \quad \boldsymbol{\xi}_i \in \mathscr{N}_{\mathscr{A}_i}(-\boldsymbol{\Lambda}_i)\,. \tag{19.30}$$

Elimination of $\boldsymbol{\xi}_i$ from (19.30) yields an implicit inclusion for the unknown impulsive forces $\boldsymbol{\Lambda}_i$,

$$\mathbf{G}_{ii}\,\boldsymbol{\Lambda}_i + \sum_{k \neq i} \mathbf{G}_{ik}\,\boldsymbol{\Lambda}_k + \mathbf{c}_i \in \mathscr{N}_{\mathscr{A}_i}(-\boldsymbol{\Lambda}_i)\,. \tag{19.31}$$

After having determined the solution $\boldsymbol{\Lambda}_i$ from (19.31), the end point velocities $\mathbf{u}^E$ are finally calculated from (19.27).

In order to show that (19.31) has indeed very nice solution properties, one should look at the two conditions in (19.30). Unknown in these two expressions are the pairs $(\boldsymbol{\xi}_i, \boldsymbol{\Lambda}_i)$. For each pair i, (19.30) provides one set of linear equations and one normal cone inclusion. Both together should be enough to finally determine the values of $(\boldsymbol{\xi}_i, \boldsymbol{\Lambda}_i)$, which is again best seen from the normal direction of a closed contact, see Fig. 19.5b and Fig. 19.6: For the normal direction, we have $\mathscr{A}_i = \mathbb{R}_0^-$, and the two vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\Lambda}_i$ consist of one element only, which are ξ_{Ni} and Λ_{Ni}. Under the assumption of known impulsive forces $\boldsymbol{\Lambda}_k$ in all force elements different from i, the first expression in (19.30) constitutes a linear equation with slope $\mathbf{G}_{ii} = (g_{ii}) > 0$ and axis intercept $\sum_{k \neq i} \mathbf{G}_{ik}\,\boldsymbol{\Lambda}_k + \mathbf{c}_i$. The graph of this linear equation in the $(-\Lambda_{Ni}, \xi_{Ni})$-plane is depicted in Fig. 19.8 and is a straight line. Because of $g_{ii} > 0$, this line is strictly decreasing and intersects in one and only one point the graph of the normal cone inclusion $\xi_{Ni} \in \mathscr{N}_{\mathbb{R}_0^-}(-\Lambda_{Ni})$. The latter is monotonically increasing and continuous and is also depicted in Fig. 19.8. As a result, one therefore obtains a unique intersection point $(-\Lambda_{Ni}, \xi_{Ni})$, which

determines the solution. The branch on which the graph of the normal cone inclusion is intersected by the line depends on the axis intercept $\sum_{k\neq i} \mathbf{G}_{ik}\,\mathbf{\Lambda}_k + \mathbf{c}_i$, and therefore on the impulsive forces $\mathbf{\Lambda}_k$ of the other force elements. By this behavior, one immediately recognizes the combinatorial problem that is concealed inside non-smooth dynamics. Unique intersection points $(\boldsymbol{\xi}_i, \mathbf{\Lambda}_i)$ for the individual force elements are also obtained in the higher-dimensional case if the symmetric matrix $\mathbf{G}_{ii}$ is positive definite, because normal cone maps are always maximal monotone. If a system consists of only frictionless contacts as in the case of the normal directions just discussed, the conditions (19.30) reduce to a *linear complementarity problem*, consisting of the linear equations $\boldsymbol{\xi} = \mathbf{G}\mathbf{\Lambda} + \mathbf{c}$ and the inequality-complementarity conditions $\boldsymbol{\xi} \geq 0$, $\mathbf{\Lambda} \geq 0$, $\boldsymbol{\xi}^{\mathsf{T}}\mathbf{\Lambda} = 0$. For one-dimensional friction, (19.30) may still be formulated as a linear complementarity problem by using additional slack variables [9]. For two-dimensional friction with force reservoirs as in Fig. 19.5d or e, however, the concept of linear complementarity is too restrictive and therefore fails. Global statements on the existence and uniqueness of solutions of (19.31) can be found in the literature but are beyond of the scope of this paper.

The inclusion (19.31) is important for analytical purposes but cannot be used for a numerical implementation. For the latter, one has to go back to (19.30) and to replace with the help of (19.4) each normal cone inclusion by its associated proximal point equation,

$$\boldsymbol{\xi}_i = \mathbf{G}_{ii}\,\mathbf{\Lambda}_i + \sum_{k\neq i} \mathbf{G}_{ik}\,\mathbf{\Lambda}_k + \mathbf{c}_i\,, \quad -\mathbf{\Lambda}_i = \mathrm{prox}_{\mathscr{A}_i}(r_i\,\boldsymbol{\xi}_i - \mathbf{\Lambda}_i)\,. \tag{19.32}$$

After having eliminated $\boldsymbol{\xi}_i$ from (19.32), one obtains a system of nonlinear equations which are implicit in $\mathbf{\Lambda}_i$,

$$-\mathbf{\Lambda}_i = \mathrm{prox}_{\mathscr{A}_i}\left[(r_i\,\mathbf{G}_{ii} - \mathbf{E})\,\mathbf{\Lambda}_i + r_i \sum_{k\neq i} \mathbf{G}_{ik}\,\mathbf{\Lambda}_k + r_i\,\mathbf{c}_i\right], \tag{19.33}$$

and which can be solved by a fixed point iteration, i.e. by taking the left-hand side as the iterated value of the right-hand side. The values $r_i > 0$ from (19.3) may here directly be used as relaxation parameters. An extensive treatise on possible iteration methods, such as Jacobi- and Gauß-Seidel iterations with and without relaxation, together with some strategies for finding appropriate values for the relaxation parameters r_i may be found in [27]. For contact problems, the Gauß-Seidel iteration with relaxation has turned out to be the most efficient one.

Note that (19.33) differs from the classical iteration methods for systems of linear equations only in the additional prox functions, which either project on the sets $\mathscr{A}_i$ or become the identity maps. The latter case always applies for bilateral frictionless constraints, see Fig. 19.5a, for which one has $\mathscr{A}_i = \mathbb{R}$. For n bilateral constraints and $\mathscr{A}_1 \times \ldots \times \mathscr{A}_n = \mathbb{R}^n$, equation (19.33) therefore becomes

$$-\mathbf{\Lambda}_i = (r_i\,\mathbf{G}_{ii} - \mathbf{E})\,\mathbf{\Lambda}_i + r_i \sum_{k\neq i} \mathbf{G}_{ik}\,\mathbf{\Lambda}_k + r_i\,\mathbf{c}_i\,. \tag{19.34}$$

When setting all relaxation parameters $r_i = 1$ in (19.34), one obtains the classical system of linear equations

$$0 = \mathbf{G}_{ii}\,\mathbf{\Lambda}_i + \sum_{k \neq i} \mathbf{G}_{ik}\,\mathbf{\Lambda}_k + \mathbf{c}_i\,, \tag{19.35}$$

from which the constraint forces, here in impulsive form, are calculated. The efficiency and robustness of the proposed projection method has been proven by many authors in various applications. Two examples that have been processed at the author's institute are briefly presented in the next section.

19.10 Applications

In this section, we present two application problems of very different nature. The first is the funnel-deflector arrangement of Fig. 19.9 that has been taken from [17] (see also [26]) and that serves as an example with many contacts to demonstrate the robustness of the algorithm. The system consists of 3,500 rigid balls that may contact each other and the surroundings. Each contact is modeled in the normal direction by a hard unilateral constraint with an impact coefficient of $\varepsilon_N = 0.5$, and in the two tangential directions by an isotropic Coulomb friction law with a friction coefficient of $\mu = 0.3$ and a tangential restitution coefficient of $\varepsilon_T = 0.5$. During the evolution of the system, up to about 4,000 simultaneously closed contacts have been observed, which yields the dimension of (19.33) to be around 12,000. In order to take advance of parallel computing, the prox iteration (19.33) has been performed on the graphics card Tesla C2050 of the computer by applying the Jacobi iteration method. The system has been processed by a time step of $\Delta t = 0.001$ s, which causes the computation time to be about 9 h for one second of real time. The number of balls was limited by the GPU memory, but not by the convergence properties of the algorithm.

The second example is a bobsled simulator software with just a few contacts, but with required real time capability. Figure 19.10 shows a scene from a simulator run between turns 13 and 14 of the Whistler Sliding Center, at which the 2010 Winter Olympics in bobsled, luge, and skeleton took place. For the simulator, the entire track has been remodeled by NURBS and then converted to a triangle mesh, which then is used in the projection algorithm. The overall track length is 1,450 m, and the average side length of the triangles is about 20 cm. The model of the bobsled is composed of several rigid bodies, which sum up to 13 degrees of freedom. Two contacts per slider, one contact for each bumper, and another five contacts for the shell to allow for turnovers have been used, which makes 17 contacts in total. For the contact between the sliders and the ice channel, a modified orthotropic friction law has been applied, whereas the remaining contacts are standard isotropic. In the simulator, speeds up to 150 km/h are achieved, and the time for one complete run is about 50 s.

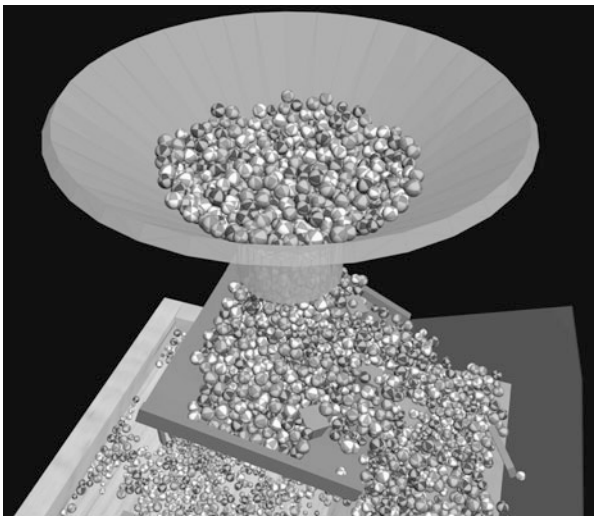

Fig. 19.9 A funnel-deflector arrangement with 3,500 rigid balls

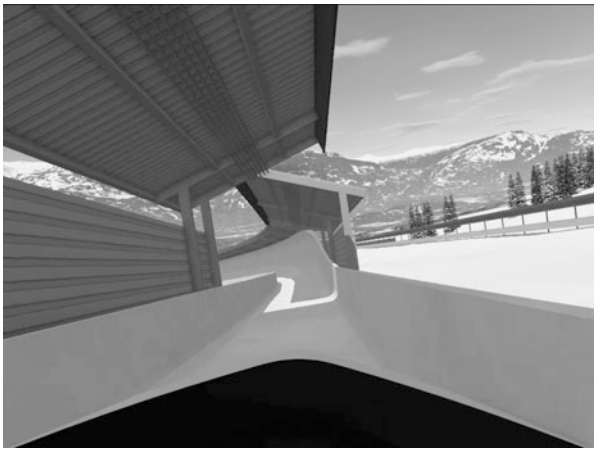

Fig. 19.10 Bobsled simulator software: Pilot's view to the Whistler track between turns 13 and 14

References

1. Aeberhard, U., Glocker, Ch.: Energy considerations for excited perfect collisions. In: van Campen, D.H., Lazurko, M.D., van den Oever, W.P.J.M. (eds.) Proc. 5th EUROMECH Nonlinear Oscillations Conference, pp. 422–431. Eindhoven Univ. of Technology, Eindhoven (2005)
2. Alart, P., Curnier, A.: A mixed formulation for frictional contact problems prone to Newton like solution methods. Comp. Meth. Appl. Mech. Eng. **92**(3), 353–357 (1991)

3. Anitescu, M., Potra, F.A., Stewart, D.E.: Time-Stepping for three-dimensional rigid body dynamics. Comp. Meth. Appl. Mech. Eng. **177**(3), 183–197 (1999)
4. Glocker, Ch.: Dynamik von Starrkörpersystemen mit Reibung und Stößen. VDI-Fortschrittberichte Mechanik/Bruchmechanik, Reihe 18, Nr. 182. VDI-Verlag, Düsseldorf (1995)
5. Glocker, Ch.: Set-Valued Force Laws: Dynamics of Non-Smooth Systems. Lecture Notes in Applied Mechanics, Vol. 1. Springer, Berlin/Heidelberg (2001)
6. Glocker, Ch.: Reduction techniques for distributed set-valued force laws. In: Baniotopoulos, C.C. (ed.) Nonsmooth/Nonconvex Mechanics with Applications in Engineering, pp. 173–180. Editions Ziti, Thessaloniki (2006)
7. Glocker, Ch.: An introduction to impacts. In: Haslinger, J., Stavroulakis, G. (eds.) Nonsmooth Mechanics of Solids, CISM Courses and Lectures, vol. 485, pp. 45–102. Springer, Wien/New York (2006)
8. Glocker, Ch., Aeberhard, U.: The geometry of Newton's cradle. In: Alart, P., Maisonneuve, O., Rockafellar, R.T. (eds.) Nonsmooth Mechanics and Analysis: Theoretical and Numerical Advances, AMMA, vol. 12, pp. 185–194. Springer, New York (2006)
9. Glocker, Ch., Studer, C.: Formulation and preparation for numerical evaluation of linear complementarity systems in dynamics. Multibody Syst. Dynam. **13**(4), 447–463 (2005)
10. Jean, M.: The non-smooth contact dynamics method. Comput. Methods Appl. Mech. Eng. **177**, 235–257 (1999)
11. Leine, R.I., Glocker, Ch.: A set-valued force law for spatial coulomb-contensou friction. Eur. J. Mech. A/Solids **22**, 193–216 (2003)
12. Leine, R.I., Nijmeijer, H.: Dynamics and bifurcations of non-smooth mechanical systems. Lecture Notes in Applied and Computational Mechanics, vol. 18. Springer, Berlin/Heidelberg (2004)
13. Möller, M.: Consistent integrators for non-smooth dynamical systems. Diss. ETH No. 19715. ETH Zurich, Zurich (2011)
14. Moreau, J.J.: Unilateral contact and dry friction in finite freedom dynamics. In: Moreau, J.J., Panagiotopoulos, P.D. (eds.) Non-Smooth Mechanics and Applications, CISM Courses and Lectures, vol. 302, pp. 1–82. Springer, Wien (1988)
15. Moreau, J.J.: Bounded variation in time. In: Moreau, J.J., Panagiotopoulos, P.D., Strang, G. (eds.) Topics in Nonsmooth Mechanics, pp. 1–74. Birkhäuser, Basel (1988)
16. Moreau, J.J.: Numerical aspects of the sweeping process. Comput. Meth. Appl. Mech. Eng. **177**, 329–349 (1999)
17. Nützi, G.: Computing non-smooth dynamics on the GPU. Master thesis, Center of Mechanics. ETH Zurich, Zurich (2011)
18. Panagiotopoulos, P.D.: A nonlinear programming approach to the unilateral contact-, and friction-boundary value problem in the theory of elasticity. Ingenieur Archiv **44**, 421–432 (1975)
19. Pang, J.S., Trinkle, J.C.: Complementarity formulations and existence of solutions of dynamic multi-rigid-body contact problems with coulomb friction. Math. Program. **30**, 199–226 (1996)
20. Paoli, L., Schatzman, M.: A numerical scheme for impact problems I: The one-dimensional case. SIAM J. Numer. Anal. **40**(2), 702–733 (2002)
21. Paoli, L.u., Schatzman, M.: A numerical scheme for impact problems II: The multidimensional case. SIAM J. Numer. Anal. **40**(2), 734–768 (2002)
22. Pfeiffer, F., Glocker, Ch.: Multibody Dynamics with Unilateral Contacts. Wiley, New York (1996)
23. Rockafellar, R.T.: Convex Analysis. Princeton University Press, Princeton (1972)
24. Stewart, D.E.: Convergence of a time-stepping scheme for rigid body dynamics and resolution of Painlevé's problem. Arch. Ration. Mech. Anal. **145**, 215–260 (1998)
25. Stewart, D.E., Trinkle, J.C.: An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and coulomb friction. Int. J. Numer. Meth. Eng. **39**(15), 2673–2691 (1996)

26. Studer, C., Glocker, Ch.: Simulation of non-smooth mechanical systems with many unilateral constraints. In: van Campen, D.H., Lazurko, M.D., van den Oever, W.P.J.M. (eds.) Proc. 5th EUROMECH Nonlinear Oscillations Conference, pp. 1597–1606. Eindhoven University of Technology, Eindhoven (2005)
27. Studer, C., Glocker, Ch.: Representation of normal cone inclusion problems in dynamics via non-linear equations. Arch. Appl. Mech. **76**, 327–348 (2006)
28. Studer, C.: Numerics of unilateral contacts and friction: modeling and numerical time integration in non-smooth dynamics. Lecture Notes in Applied and Computational Mechanics, vol. 47. Springer, Berlin/Heidelberg (2009)

Chapter 20
Dynamics of Second Grade Fluids: The Lagrangian Approach

M. Paicu and G. Raugel

This paper is dedicated to Professor Jürgen Scheurle on the occasion of his 60th birthday

Abstract This article is devoted to the mathematical analysis of the second grade fluid equations in the two-dimensional case. We first begin with a short review of the existence and uniqueness results, which have been previously proved by several authors. Afterwards, we show that, for any size of the material coefficient $\alpha > 0$, the second grade fluid equations are globally well posed in the space $V^{3,p}$ of divergence-free vector fields, which belong to the Sobolev space $W^{3,p}(\mathbb{T}^2)^2$, $1 < p < +\infty$, where $\mathbb{T}^2$ is the two-dimensional torus. Like previous authors, we introduce an auxiliary transport equation in the course of the proof of this existence result. Since the second grade fluid equations are globally well posed, their solutions define a dynamical system $S_\alpha(t)$. We prove that $S_\alpha(t)$ admits a compact global attractor $\mathscr{A}_\alpha$ in $V^{3,p}$. We show that, for any $\alpha > 0$, there exists $\beta(\alpha) > 0$, such that $\mathscr{A}_\alpha$ belongs to $V^{3+\beta(\alpha),p}$ if the forcing term is in $W^{1+\beta(\alpha)}(\mathbb{T}^2)^2$. We also show that this attractor is contained in any Sobolev space $V^{3+m,p}$ provided that α is small enough and the forcing term is regular enough. The method of proof of the existence and regularity of the compact global attractor is new and rests on a Lagrangian method. The use of Lagrangian coordinates makes the proofs much simpler and clearer.

M. Paicu
Univ. Bordeaux, IMB, UMR 5251, F-33400 Talence, France
e-mail: Marius.Paicu@math.u-bordeaux1.fr

G. Raugel (✉)
CNRS, Laboratoire de Mathématiques d'Orsay, Orsay Cedex, F-91405; Univ. Paris-Sud, Orsay Cedex, F-91405, France
e-mail: Genevieve.Raugel@math.u-psud.fr

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_20,

20.1 Introduction

In petroleum industry, in polymer technology, in problems of liquid crystals suspensions, non-Newtonian (also called Rivlin–Ericksen) fluids of differential type often arise. The constitutive law of incompressible homogeneous fluids of grade 2 is given by

$$\sigma = -pI + 2\nu A_1 + \alpha_1 A_2 + \alpha_2 A_1^2,$$

where σ is the Cauchy tensor, A_1 and A_2 are the first two Rivlin–Ericksen tensors:

$$A_1(u) = \frac{1}{2}[\nabla u + \nabla u^T], \quad A_2(u) = \frac{DA_1}{Dt} + (\nabla u)^T A_1 + A_1(\nabla u)$$

and

$$\frac{D}{Dt} = \partial_t + u.\nabla.$$

is the material derivative.

In 1974, Dunn and Fosdick [16] established that a fluid modelled by the above relations is compatible with thermodynamics (that is, the Clausius–Duhem inequality and the assumption that the Helmholtz free energy is a minimum when the fluid is at rest) if the following conditions

$$\alpha_1 + \alpha_2 = 0, \quad \alpha_1 \geq 0\,,$$

are imposed.

Writing then the equation $\frac{Du}{Dt} = u_t + u.\nabla u = \operatorname{div}\sigma$, one obtains the second grade fluid equations (20.2) below.

If $\alpha_1 \geq 0$, the fluid has asymptotic stability properties. In [18], it was showed that if $\alpha_1 + \alpha_2$ is arbitrary and $\alpha_1 < 0$, then the second grade fluid has an anomalous behaviour (unstable behaviour). There has been an extensive discussion on the modelling of the second grade fluids and on the restrictions, which have to be imposed on the coefficients α_1 and α_2 (see [16–18], for example).

If one does not impose the condition $\alpha_1 + \alpha_2 = 0$, the system of second grade can be written as

$$\begin{aligned}
&\partial_t(u - \alpha_1 \Delta u) - \nu \Delta u + \operatorname{rot}(u - (2\alpha_1 + \alpha_2)\Delta u) \times u \\
&\qquad + (\alpha_1 + \alpha_2)(-\Delta(u \cdot \nabla u) + 2u \cdot \nabla(\Delta u)) + \nabla p = f, \quad t > 0,\ x \in \Omega, \\
&\operatorname{div} u = 0, \quad t > 0,\ x \in \Omega, \\
&u(0, x) = u_0(x), \quad x \in \Omega\,,
\end{aligned} \tag{20.1}$$

where $\alpha_1 \geq 0$. When $\alpha_1 + \alpha_2 = 0$, setting $\alpha = \alpha_1$, we obtain the system of second grade fluids in the simplified form,

$$\begin{aligned} &\partial_t(u - \alpha\Delta u) - \nu\Delta u + \operatorname{rot}(u - \alpha\Delta u) \times u + \nabla p = f, \quad t > 0,\ x \in \Omega, \\ &\operatorname{div} u = 0, \quad t > 0,\ x \in \Omega, \\ &u(0, x) = u_0(x), \quad x \in \Omega\,, \end{aligned} \tag{20.2}$$

where Ω is either a bounded simply connected regular enough domain in $\mathbb{R}^d$, or the d-dimensional torus $\mathbb{T}^d$, $d = 2, 3$. In the two-dimensional case, we use the convention that $\operatorname{rot} u \equiv \operatorname{curl} u = (0, 0, \partial_1 u_2 - \partial_2 u_1)$ and we identify each 2-component vector-field $u = (u_1, u_2)$ with the 3-component vector field $u = (u_1, u_2, 0)$ and each scalar m with the 3-component vector field $w = (0, 0, m)$. If Ω is a bounded domain in $\mathbb{R}^d$, the equations (20.2) are completed with boundary conditions. In most of the papers, one assumes that the fluid adheres to the boundary $\partial\Omega$, that is, one requires homogeneous Dirichlet boundary conditions

$$u(x, t) = 0\,, \quad t > 0, x \in \partial\Omega\,. \tag{20.3}$$

The condition (20.3) is sufficient to determine a unique local solution of the system (20.2) despite the fact that the nonlinearity in (20.2) contains derivatives of higher order than 2. One can also consider the system (20.2) with non-homogeneous Dirichlet boundary conditions

$$u(x, t) = g(x, t)\,, \quad t > 0, x \in \partial\Omega\,, \tag{20.4}$$

where g must satisfy the compatibility condition $\int_{\partial\Omega} g \cdot n ds = 0$, n being the outward normal to the boundary $\partial\Omega$. For such boundary conditions, in the case of three-dimensional bounded domains Ω, Galdi et al. [26] proved the existence of local solutions of the system (20.2). They showed the uniqueness of the solutions, when the boundary is impermeable, that is, when $g \cdot n \equiv 0$. In the case where the boundary is impermeable, Girault and Scott [27] proved the existence of stationary solutions, when Ω is a two-dimensional domain. Under additional smallness conditions on the data, Girault and Scott obtained the uniqueness of the stationary solutions. The second grade fluid model with fully non-homogeneous Dirichlet boundary conditions is actually not well posed. For example, Gupta and Rajagopal [29] have given examples in which the stationary problem has multiple solutions. For this reason, it is important to require that $g \cdot n \equiv 0$.

C. Le Roux [45] has studied the system (20.2) subject to non-linear partial slip boundary conditions in a bounded simply-connected domain in $\mathbb{R}^3$. Under appropriate growth restrictions on the data, he has proved the existence and uniqueness of a classical solution.

Before describing the contents of this paper, we briefly recall the main known existence and uniqueness results of the solutions of (20.2) in the case of the

homogeneous Dirichlet boundary conditions (20.3). Since there are many papers devoted to this case, we cannot quote all of them. In particular, we will not recall the results concerning the stationary solutions (see, for example, [4, 7, 21, 24, 27]).

The first general existence and uniqueness results of solutions of (20.2) are due to Cioranescu and Ouazar in 1984 (see [13] and [14]). Assuming that the initial data u_0 belong to the space $W = \{v \in H^3(\Omega)^d \,|\, \operatorname{div} v = 0, v_{|\partial\Omega} = 0\}$ and that the force f belongs to $L^2((0,T), H^1(\Omega)^d)$ and using a Galerkin method with a special basis, Cioranescu and Ouazar proved that (20.2) has a unique (weak) solution $u \in L^\infty((0,T^*), W) \cap W^{1,\infty}((0,T^*), W')$, where $T^* = T$ in the case $d = 2$ and $0 < T^* \leq T$ in the case $d = 3$. Later, in 1997, Ciorasnecu and Girault [12] completed these results by showing the global existence of weak solutions in the three-dimensional case under the assumption that the data are small enough and showed that these solutions are more regular if the data are smoother. In 1993, Galdi, Grobbelaarvandalsen and Sauer [25] have shown the local existence and uniqueness of classical solutions of (20.1) and also the global existence of solutions of (20.1) under a smallness condition of the data, when α_1 is large enough. These local and global existence of classical solutions results have been improved in 1994 by Galdi and Sequeira [23] and, in particular, the requirement that α be large enough has been removed. In [25] (resp. [23]), the local (or global) existence of classical solutions has been proved by writing an equation for the auxiliary variable $v = u - \alpha\Delta u$ (resp. $v = curl(u - \alpha\Delta u)$) and by applying the Leray–Schauder fixed point theorem. For instance, in [23], assuming that the forcing term f vanishes and that $v_0 = curl(u_0 - \alpha\Delta u_0)$ belongs to $X_m = \{v \in H^m(\Omega)^3 \,|\, \operatorname{div} v = 0\}$, $m \geq 1$, the authors have proved the local existence and uniqueness of the solution u of (20.2) in $C^0((0,T), X_{m+2}) \cap L^\infty((0,T), H^{m+3}(\Omega)^3)$ with $\frac{du}{dt} \in L^\infty((0,T), H^{m+2}(\Omega)^3)$, where $T > 0$. Under a smallness condition on u_0, they proved that the solution is global.

Later in 1998, Bernard [3] has generalized the existence result of (weak) solutions of (20.2) to the system (20.1) by using a Galerkin method with the special basis as in [12] or [14]. Roughly speaking, assuming that f and u_0 belong to $L^1((0,+\infty), H^1(\Omega)^3) \cap L^\infty((0,+\infty), L^2(\Omega)^3)$ and to W respectively and are both small enough, Bernard has proved the global existence and uniqueness of a solution $u \in L^\infty((0,+\infty), W)$ and that $\frac{du}{dt} \in L^\infty((0,+\infty), H^1(\Omega)^3)$.

Also in 1998, in the three-dimensional case, Bresch and Lemoine [8] have obtained the existence and uniqueness of solutions of (20.2), when $f \in L^r((0,T), L^r(\Omega)^3)$ and u_0 is a divergence-free vector field in $X_1 \cap W^{2,r}(\Omega)^3$, where $r > 3$. More precisely, under these hypotheses, they showed that there exists a (unique) solution $u(t) \in C^0([0,T^*], W^{2,r}(\Omega)^3 \cap X_1)$, with $\frac{du}{dt} \in L^r((0,T^*), W^{1,r}(\Omega)^3)$ where $0 < T^* \leq T$. If f is in $L^\infty((0,+\infty), L^r(\Omega)^3)$, f and $u_0 \in X_1 \cap W^{2,r}(\Omega)^3$ are small enough and α is larger than a constant depending only on r and Ω, then the solution $u(t)$ is global and belongs to $C_b^0([0,+\infty), W^{2,r}(\Omega)^3 \cap X_1)$, with $\frac{du}{dt} \in L^\infty((0,+\infty), W^{1,r}(\Omega)^3)$. In their proof, given u, the authors introduce the unique solution w of the linear equation $w_t + (\nu/\alpha)w + u \cdot \nabla w + \nabla u \cdot w = (\nu/\alpha)u + f$, with $w(0) = u(0) - \alpha\Delta u(0)$. Then,

they consider the unique solution (z, π) of the "Stokes" problem $z - \alpha\Delta z + \nabla\pi = w$, where z is divergence-free and the mean value of π vanishes. Finally, applying the Leray–Schauder fixed point theorem, they show that the map $u \mapsto z$ has a fixed point. Of course, arguing in the same way, one can prove similar existence and uniqueness results when Ω is a bounded domain in $\mathbb{R}^2$ and $r > 2$. One notices that Bresch and Lemoine have used a similar strategy in [9] to prove the existence and uniqueness of a solution for third grade fluids. For other existence results in $W^{2,r}(\Omega)^3, r > 3$, see also [6].

In 2007, Girault and Saadouni [28] considered the equations of grade two (20.2) on a two-dimensional Lipschizian domain Ω. They proved the existence of a weak solution of (20.2) and obtained the uniqueness of the solution if Ω is a convex polygon. Introducing the auxiliary variable $z = \mathrm{rot}\,(u - \alpha\Delta u)$, they have replaced the system (20.2) by the equivalent system

$$\begin{aligned}
&\partial_t(u - \alpha\Delta u) - \nu\Delta u + z \times u + \nabla p = f, \quad t > 0,\ x \in \Omega,\\
&\alpha\partial_t z + \nu z + \alpha u \cdot \nabla z = \alpha\mathrm{rot}\, f + \nu\mathrm{rot}\, u, \quad t > 0,\ x \in \Omega,\\
&\mathrm{div}\, u = 0, \quad t > 0,\ x \in \Omega,\\
&u(0, x) = u_0(x), \quad x \in \Omega\,.
\end{aligned} \tag{20.5}$$

The authors proved the existence of a (weak) solution by using a semi-discretization in time of the system (20.5).

For the asymptotic behaviour in time of the solutions of (20.2), when Ω is replaced by $\mathbb{R}^2$ (respectively, $\mathbb{R}^3$), we refer the reader the papers [40, 41] and [15] (respectively [15] and [51]). Additional interesting related results about non-Newtonian second grade or third grade fluids are contained in [5, 22, 39, 48, 57, 59].

We would like to notice that the equations (20.2) differ from the so-called α-Navier–Stokes system (see, e.g., [20] and the references therein). Indeed, the α-Navier–Stokes model contains the strong regularizing term $-\nu\Delta(u - \alpha\Delta u)$ instead of $-\nu\Delta u$, and thus is a semilinear problem, which is much easier to solve than the second grade fluid equations where the dissipation is weaker.

In the inviscid case $\nu = 0$, the local existence and uniqueness of regular solutions still hold and, in the two-dimensional case, these solutions are global (see [10] for example). For the convergence of the solutions u_ν of (20.2) towards the solution u^* of Eq. (20.2) for $\nu = 0$, when ν goes to zero, we refer the reader to [60]. For additional results in the inviscid case, we also refer to [47].

Until now, only few papers have been devoted to the dynamics of second grade fluids. In 1998, Moise, Rosa and Wang [50] have considered the second grade fluid equations (20.2) with time-independent forcing term $f \in H^1(\Omega)^2$, where Ω is a bounded simply-connected domain in $\mathbb{R}^2$. In this case, we can introduce the dynamical system $S_\alpha(t)$ on W, defined by $S_\alpha(t)u_0 = u(t)$, where $u(t)$ is the solution of (20.2). Moise, Rosa and Wang have shown that the map $u_0 \in W \mapsto S_\alpha(t)u_0 \equiv u(t) \in W$ is continuous and that every solution u of (20.2) belongs to $C^0([0, +\infty), W)$. Applying the method of functionals of J. Ball, they

have proved that $S_\alpha(t)$ is asymptotically compact in W, which implies, since S_α has an absorbing set in W, that $S_\alpha(t)$ admits a compact global attractor $\mathscr{A}_\alpha$ in W (for the notions of asymptotic compactness and absorbing set, see Sect. 20.3 below).

In [55], Paicu, Raugel and Rekalo have proved that there exists a positive constant $\delta = \delta(\alpha, \|f\|_{H^1})$ such that the compact global attractor $\mathscr{A}_\alpha$ in W is actually bounded in $H^{3+\delta}(\Omega)^2$, when f belongs to $H^{1+\delta}(\Omega)^2$. Moreover, $\mathscr{A}_\alpha$ is bounded in $H^{3+m}(\Omega)^2$, $m \geq 0$, provided α is small enough and f belongs to $H^{1+m}(\Omega)^2$. They have also shown that, on the attractor, the second grade fluid equations (20.2) reduce to a finite number of ordinary differential equations with an infinite delay term [55, Sect. 5]. From these properties, they deduced that, as for the Navier–Stokes equations, the property of finite number of determining modes holds. Let us recall that the global attractor contains all the interesting asymptotic dynamics, in particular the equilibrium points and the periodic orbits. We would like to emphasize that the regularity property of the global attractor has important consequences. For example, they allow to prove persistence of non-degenerate equilibrium points or periodic orbits, when various parameters in the system (20.2) vary, such as the coefficient α or the domain Ω (see [35, 36, 40, 49]). In particular, if the Navier–Stokes equations admit a non-degenerate periodic orbit of minimal period $\omega > 0$, using these regularity properties, one obtains that, for $\alpha > 0$ small enough, (20.2) has a unique periodic orbit, which is close to the corresponding one of the Navier–Stokes equations and has minimal period ω_α close to ω [35]. If Ω is a three-dimensional bounded domain, there exists a compact attractor if f is small enough. But this attractor is a local one, since we do not know if the solutions exist globally for any size of the initial data. Thus, the study of this (local) attractor is less interesting. The above-mentioned regularity properties are certainly still true for the local attractor.

In this paper, we consider the equations of second grade, when the forcing term belongs to $L^\infty((0,+\infty), W^{1,p}(\Omega)^2)$ and the initial data are divergence free and belong to $W^{3,p}(\Omega)^2$, where $p > 1$. First, we prove the existence and uniqueness of the weak solution of (20.2), give some a priori estimates and show that the equation (20.2) generate a dynamical system $S_\alpha(t)$ on the subspace of divergence-free vector fields of $W^{3,p}(\Omega)^2$. In Sect. 20.3, we show that the dynamical system $S_\alpha(t)$ admits a compact global attractor $\mathscr{A}_\alpha$, which is bounded in a more regular space. We prove the existence and regularity of $\mathscr{A}_\alpha$ by using the Lagrangian coordinates. By adopting the Lagrangian approach, we simplify the previous proofs of the existence and the regularity of the compact global attractor.

For the sake of simplicity, we will only prove the results in the case where $\Omega = \mathbb{T}^2$.

Before we briefly describe these results, we introduce the needed notation. We denote $V^{m,p}$, $m \in \mathbb{N}$, $p \geq 1$, the closure of the space

$$\{u \in [C^\infty(\mathbb{T}^2)]^2 \,|\, u \text{ is periodic }, \operatorname{div} u = 0, \int_{\mathbb{T}^2} u\, dx = 0\},$$

in $W^{m,p}(\mathbb{T}^2)^2$. If $p = 2$, we set $V^m \equiv V^{m,2}$ and we simply write $H = V^0$. We equip the space $V^{m,p}$ with the classical $W^{m,p}(\mathbb{T}^2)^2$-norm, denoted $\|\cdot\|_{V^{m,p}} \equiv \|\cdot\|_{W^{m,p}}$. We will also use the usual $L^2(\mathbb{T}^2)^2$-scalar product $(\cdot,\cdot)$.

Finally, we denote $W^{m,p}_{per} \equiv W^{m,p}_{per}(\mathbb{T}^2)^2$ the space of vector fields $u \in W^{m,p}(\mathbb{T}^2)^2$, which are periodic and whose mean value vanishes.

If $m \in \mathbb{N}$, we define the spaces $W^{-m,p}_{per}$ as the dual space of W^{m,p^*}_{per}, where $\frac{1}{p} + \frac{1}{p^*} = 1$.

As several authors have already done it (see, e.g., [12, 23, 27, 45]), we consider the auxiliary variable $\omega = curl(u - \alpha\Delta u) \equiv \mathrm{rot}\,(u - \alpha\Delta u)$. Applying the curl (also called rotational) operator to the first equation in (20.2), we formally obtain the equation

$$\partial_t\omega + \frac{\nu}{\alpha}\omega + u\cdot\nabla\omega = \mathrm{rot}\, f + \frac{\nu}{\alpha}\mathrm{rot}\, u, \quad t > 0,\ x \in \Omega. \tag{20.6}$$

We thus replace the system (20.2) by the following system

$$\begin{aligned}
&\partial_t\omega + \frac{\nu}{\alpha}\omega + u\cdot\nabla\omega = \mathrm{rot}\, f + \frac{\nu}{\alpha}\mathrm{rot}\, u, \quad t > 0,\ x \in \Omega,\\
&\omega(0,x) = \mathrm{rot}\,(u_0(x) - \alpha\Delta u_0(x)), \quad x \in \Omega\,,\\
&\omega = \mathrm{rot}\,(u - \alpha\Delta u), \quad t > 0,\ x \in \Omega,\\
&\mathrm{div}\, u = 0, \quad t > 0,\ x \in \Omega,
\end{aligned} \tag{20.7}$$

where $\mathrm{rot}\, f \in L^\infty((0,+\infty), L^p_{per})$ and $u_0 \in V^{3,p}$.

In Sect. 20.2, we will prove that (20.2) (or (20.7)) has a solution u by showing that the map $J : u \in L^\infty((0,+\infty), V^{3,p}) \mapsto \omega \mapsto z \in L^\infty((0,+\infty), V^{3,p})$ has a fixed point, where, given u, ω is the solution of the affine equation (20.6) and z is the solution of the equation $\omega = \mathrm{rot}\,(z - \alpha\Delta z)$. The fixed point is obtained by applying the Leray–Schauder fixed point theorem and by adopting a Lagrangian point of view. As it was recalled in the above lines, the idea of applying the Leray–Schauder fixed point theorem is not new (however, the existence result below is new to our knowledge). Elementary a priori estimates will show that the solution u is unique in $L^\infty((0,+\infty), V^{2,2})$. This will lead us to the following theorem.

Theorem 20.1. *(i) Assume that $p > 1$ and that the forcing term f is in $L^\infty((0,+\infty), W^{1,p}_{per})$. Then, for every $u_0 \in V^{3,p}$, there exists a unique solution $u(t)$ of the equations (20.2) such that $u(t) \in C^0([0,+\infty), V^{3,p})$ and $\frac{d}{dt}u(t) \in L^\infty((0,+\infty), V^{2,p})$. Moreover, for any $t \geq 0$, the map $u_0 \in V^{3,p} \mapsto u(t) \in V^{3,p}$ is continuous.*

(ii) Likewise, if f belongs to $L^\infty(\mathbb{R}, W^{1,p}_{per})$, then, for every $u_0 \in V^{3,p}$, there exists a unique solution $u(t)$ of the equations (20.2) such that $u(t) \in C^0(\mathbb{R}, V^{3,p})$ and $\frac{d}{dt}u(t) \in L^\infty((0,+\infty), V^{2,p}) \cap L^\infty_{loc}(\mathbb{R}, V^{2,p})$. Moreover, for any $t \in \mathbb{R}$, the map $u_0 \in V^{3,p} \mapsto u(t) \in V^{3,p}$ is continuous.

More precise upper bounds of the solutions are given in Sect. 20.2.4. Likewise, by adopting the Lagrangian point of view, one could also prove the existence of a unique solution $u(t)$ and the boundedness of it, when the viscosity ν vanishes. We will give the details in this case in a subsequent paper.

Assume now that $f \in W^{1,p}_{per}$ is time-independent, then (20.2) is an autonomous system and the map $S_\alpha(t) : u_0 \in V^{3,p} \mapsto S_\alpha(t)u_0 \equiv u(t) \in V^{3,p}$ (where $u(t)$ is the solution of (20.2)) is a dynamical system and even a non-linear continuous group, that is, $S_\alpha(t)$ has the following properties

1. $S_\alpha(t)S_\alpha(s) = S_\alpha(t+s)$, for any $t, s \in \mathbb{R}$,
2. $u_0 \in V^{3,p} \mapsto S_\alpha(t)u_0 \equiv u(t) \in V^{3,p}$ is continuous from $V^{3,p}$ into $V^{3,p}$, for any $t \in \mathbb{R}$,
3. $t \mapsto S_\alpha(t)u_0 \in V^{3,p}$ belongs to $C^0(\mathbb{R}, V^{3,p})$, for any $u_0 \in V^{3,p}$.

The proof of Theorem 20.1 implies that $S_\alpha(t)$ admits a bounded absorbing set, that is, there exists a bounded set $\mathscr{B}_\alpha$ in $V^{3,p}$, such that, for any bounded set $B \in V^{3,p}$, there exists a time $\tau(B)$ such that, for $t \geq \tau(B)$,

$$S_\alpha(t)B \subset \mathscr{B}_\alpha \, .$$

From the proof of Theorem 20.1, one deduces that, in the case where f is time-independent, $\frac{d}{dt}u$ belongs to $C^0(\mathbb{R}, V^{2,p})$, which allows to state the following corollary.

Corollary 20.1. *Assume that $p > 1$ and that the forcing term $f \in W^{1,p}_{per}$ is time-independent, then for every $u_0 \in V^{3,p}$, there exists a unique solution $u(t)$ of the equations (20.2) such that $u(t) \in C^0(\mathbb{R}, V^{3,p})$ and $\frac{d}{dt}u(t) \in C^0(\mathbb{R}, V^{2,p}) \cap L^\infty((0,+\infty), V^{2,p}) \cap L^\infty_{loc}(\mathbb{R}, V^{2,p})$. Moreover, the dynamical system $S_\alpha(t)$ admits a bounded absorbing set in $V^{3,p}$.*

A dynamical system which has an absorbing set is called *bounded dissipative* (for further details, see [31], [32] or [58], for example). If a dynamical system is bounded dissipative, one may wonder if it has also asymptotic compactness properties, which will imply that it admits a compact global attractor (see [32, Theorem 3.4.6] or [58, Theorem 2.26], for example). Before stating the existence theorem of a compact global attractor, we recall its definition.

Definition 20.1. Let X be a Banach space and $S(t)$ be a dynamical system on X. A compact set $\mathscr{A} \in X$ is a compact *global attractor* if

- $\mathscr{A}$ is invariant, that is, $S(t)\mathscr{A} = \mathscr{A}$, for any $t \geq 0$,
- $\mathscr{A}$ attracts all bounded sets of X, that is, for any $\varepsilon > 0$, for any bounded set B in X, there exists a time $T = T(\varepsilon, B)$ such that

$$S(t)B \subset \mathscr{N}_X(\mathscr{A}; \varepsilon) \, , \qquad \text{for any t } \geq T \, ,$$

where $\mathscr{N}_X(\mathscr{A}; \varepsilon)$ denotes the ε-neighbourhood of $\mathscr{A}$ in X.

The compact global attractor plays an important role, since all the asymptotic (and interesting) dynamics are contained in it. In Sect. 20.3, we are going to show that $S_\alpha(t)$ is *asymptotically smooth* or *asymptotically compact*.

We recall that a dynamical system $S(t)$ on a Banach space X is *asymptotically compact* (or *asymptotically smooth*; for an equivalent definition of asymptotic smoothness, see [32, Chap. 3.2] or [58, Definition 2.12 and Proposition 2.15]) if, for any bounded subset B of X such that $\cup_{t\geq 0} S(t+\tau)(B)$ is bounded for some $\tau \geq 0$, every set of the form $\{S(t_n)z_n\}$, with $z_n \in B$ and $t_n \geq \tau$, $t_n \to_{n\to+\infty} +\infty$, is relatively compact in X. For further general concepts of dissipative systems, we refer the reader to [31–33, 38, 58, 61].

Since the equation (20.2) is fully non-linear (and not only semi-linear), the asymptotic compactness of $S_\alpha(t)$ is not straightforward. In [50], for the case $p = 2$, Moise, Rosa and Wang had proved it by using the method of functionals of J. Ball. Here, using the Lagrangian point of view, we will be able to write $S_\alpha(t)$ as the sum $S_\alpha(t)u_0 = \Sigma_\alpha(t)u_0 + K_\alpha(t)u_0$, where $\Sigma_\alpha(t)u_0$ is a map, which is "asymptotically contracting" on $V^{3,p}$ and $K_\alpha(t)$ is a compact map from $V^{3,p}$ into itself (see Sect. 20.3 for more details). This property implies by [32, Lemma 3.2.6] or [58, Theorem 2.31] that $S_\alpha(t)$ is asymptotically compact. Since $S_\alpha(t)$ is also bounded dissipative, [32, Theorem 3.4.6] or [58, Theorem 2.26] then imply that $S_\alpha(t)$ has a compact global attractor in $V^{3,p}$, which is also connected.

Theorem 20.2. *For $p > 1$, if the forcing term f is time-independent and belongs to $W^{1,p}_{per}$, then $S_\alpha(t)$ admits a compact global attractor $\mathscr{A}_\alpha$ in $V^{3,p}$ and $\mathscr{A}_\alpha$ is connected.*

The fact that $S_\alpha(t)$ is a non-linear group prevents smoothing properties in finite time. Thus, in view of the applications (persistence of equilibrium points, of periodic orbits, of local stable and unstable manifolds under perturbations of the equation (20.2), it is interesting to know if the elements or trajectories on the global attractor $\mathscr{A}_\alpha$ are more regular.

Numerous authors have shown regularity properties of the compact global attractor in the case of dynamical systems which are not smoothing in finite time. Such results were obtained already more than 30 years ago for retarded functional differential equations in $\mathbb{R}^n$ with finite delay or neutral functional differential equations by Hale [30] and Nussbaum [53]. For dissipative evolutionary equations, which admit a compact global attractor, regularity results have later been proved by several authors, using different methods (see [34] and [55] for references). We recall that one of the first regularity results applicable to partial differential evolutionary equations has been shown by Hale and Scheurle [37] in 1985, who considered the equation

$$\dot{u} = Au + f(u)\,, \quad u(0) = u_0 \in X\,, \tag{20.8}$$

on a Banach space X, where A is the generator of a (linear) C^0 semi-group and $f(\cdot)$ is a smooth map on X. It is known that, for any $u_0 \in X$, there exists a unique local

mild solution $u(t) \in C^0([0,T);X)$ of (20.8). Let us assume that all the solutions exist on $[0,+\infty)$. Then, (20.8) defines a dynamical system $S(t)$ on X, given by $S(t)u_0 = u(t)$ where $u(t)$ is the solution of (20.8). Hale and Scheurle have proved that if $S(t)$ has a compact invariant set $\mathscr{J}$ in X, then there exists a positive number η such that if $\|Df(v)\|_{L(X,X)} \leq \eta$ for any v in a small neighborhood of $\mathscr{J}$, the mapping $t \in \mathbb{R} \to S(t)u \in X$, for any $u \in \mathscr{J}$, is as smooth as f. The smoothness in the time variable implies smoothness in the spatial variable if (20.8) is the abstract version of a PDE. In particular, if the restriction of $S(t)$ to $\mathscr{J}$ is of class C^1, then $\mathscr{J}$ is bounded in the domain $D(A)$, which usually is a smoother space than X.

The system of second grade (20.2) is more complex than the abstract equation (20.8) and one cannot deduce spatial regularity properties from the time regularity results. In [60], using Lagrangian coordinates, Shkoller has proved time regularity properties of all the solutions of (20.2). However, from these time regularity results, one cannot deduce spatial regularity properties.

In [55, Sect. 2], in the special case where $p = 2$, we have proved the regularity of the attractor $\mathscr{A}_\alpha$ by establishing a series of appropriate a priori estimates for the solutions of the linear equation (which is the analogous of the transport equation (20.6))

$$\begin{aligned}
&\partial_t(w^* - \alpha\Delta w^*) - \nu\Delta w^* + \operatorname{rot}(w^* - \alpha\Delta w^*) \times u^* + \nabla p^* = f, \quad t > 0,\ x \in \mathbb{T}^2,\\
&\operatorname{div} w^* = 0, \quad t > 0,\ x \in \mathbb{T}^2,\\
&w^*(0,x) = u_0(x), \quad x \in \mathbb{T}^2\,,
\end{aligned} \tag{20.9}$$

where $f \in H^{m+1}_{per}$ and $u^* \in L^\infty((0,+\infty), V^{m+2}) \cap C^0([0,+\infty), V^2)$ and by using the decomposition of $S_\alpha(t)u_0$ into $S_\alpha(t)u_0 = v_n(t) + (S_\alpha(t)u_0 - v_n(t))$, where $v_n(t)$ is the solution at time t of the equation (20.2) satisfying $v(s_n) = 0$ and where s_n is a sequence converging to $-\infty$. In the course of this proof, we have obtained "good" estimates of the size of the elements of $\mathscr{A}_\alpha$ in various norms. However, the proofs were long.

Here, using the system (20.7) for $p > 1$ and the Lagrangian coordinates, we are proving the regularity of $\mathscr{A}_\alpha$ in a more elegant way (see Sect. 20.3). Notice that, in the case $m = 1$ below, we recover the same condition as in [55, Theorem 1.1]. In the case $m > 1$, we obtain a better condition for the regularity than in [55, Theorem 1.1].

Theorem 20.3. *Let $p > 1$.*

1) Let $f \in W^{2,p}_{per}$. Assume that $\sup_{v\in\mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} < \frac{\nu}{\alpha}$ and let $a_1 \equiv \frac{\nu}{\alpha} - \sup_{v\in\mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} > 0$. Then, the following upper bound holds for any u belonging to the global attractor

$$\|\nabla(\operatorname{rot} u - \alpha\Delta \operatorname{rot} u)\|_{L^p} \leq a_1^{-1}(\|\operatorname{rot} f\|_{W^{1,p}} + \frac{\nu}{\alpha} M_\alpha(p))\,,$$

where $M_\alpha(p)$ is given in (20.84) below.

2) *There always exists* $0 < \theta \leq 1$ *such that* $a_{1,\theta} \equiv \frac{\nu}{\alpha} - \theta \sup_{v \in \mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} > 0$. *If* $f \in W^{1+\theta,p}_{per}$, *then the following estimate is true for any* u *belonging to the global attractor*

$$\|\mathrm{rot}\, u - \alpha \Delta \mathrm{rot}\, u\|_{W^{\theta,p}} \leq a^{-1}_{1,\theta}(\|\mathrm{rot}\, f\|_{W^{\theta,p}} + \frac{\nu}{\alpha} M_\alpha(p)) \, .$$

3) *More generally, if* $f \in W^{1+m,p}_{per}$ *and* $a_m \equiv \frac{\nu}{\alpha} - (2m-1) \sup_{v \in \mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} > 0$, *then the following upper bound holds for any* u *belonging to the global attractor*

$$\|\mathrm{rot}\, u - \alpha \Delta \mathrm{rot}\, u\|_{W^{m,p}} \leq a^{-1}_m M_{m,\alpha}(p) \, ,$$

where $M_{m,\alpha}(p)$ *is a positive constant.*

The paper is organized as follows. Section 20.2 is devoted to the proof of Theorem 20.1 and to several remarks about the solutions of (20.2). In Sect. 20.3, we first prove that $S_\alpha(t)$ is asymptotically smooth in $V^{3,p}$ and thus admits a compact global attractor $\mathscr{A}_\alpha$ in $V^{3,p}$. Afterwards, we prove Theorem 20.3, that is, the regularity properties of $\mathscr{A}_\alpha$ if the forcing term is smoother.

20.2 Existence Results for the Second Grade Fluid Equations

Theorem 20.1 can be proved in different ways. For example, we could remark that the local existence result [8, Theorem 1] can be extended to the two-dimensional case and the periodic boundary conditions, when the initial data belong to $V^{2,q}$, $q > 2$ and the forcing term f is in $L^\infty((0,+\infty), L^q_{per})$. Since $W^{1,p}(\mathbb{T}^2)$, $p > 1$, is continuously embedded into the space $L^{q_0}(\mathbb{T}^2)$, where $q_0 > 2$, we could deduce from Theorem 1 of [8] that, for every $u_0 \in V^{3,p}$, there exists a unique local solution $u(t) \in C^0([0,T), V^{2,q_0})$ of (20.2), where $T > 0$. Afterwards, we could show that this solution is unique and is actually more regular.

However, since we want to emphasize the important role of the transport equation (20.6), we will give a complete direct proof of Theorem 20.1.

20.2.1 The Transport Equation

Since the existence of the solution of (20.2) will be proved by a fixed point argument involving the solution ω of the transport equation (20.6), we first study the following general transport equation (where $\nu > 0$ and $\alpha > 0$),

$$\begin{aligned} &\partial_t w + \frac{\nu}{\alpha} w + u \cdot \nabla w = g \, , \quad t > 0, x \in \mathbb{T}^2 \, , \\ &w(0,x) = w_0(x) \, , \quad x \in \mathbb{T}^2 \, , \end{aligned} \tag{20.10}$$

where, for the sake of simplicity (and in view of the applications), $u \in C^0([0,+\infty), V^{2,p}) \cap L^\infty((0,+\infty), V^{3,p})$, $p > 1$.

Before stating an existence and uniqueness result of solutions of (20.10), we introduce the "Lagrangian coordinates", that is, the following ordinary differential equation, for $t, \tau \in [0,+\infty)$, $x \in \mathbb{T}^2$,

$$\partial_t \varphi(t;\tau,x) = u(t,\varphi(t;\tau,x)) \ , \quad \varphi(\tau;\tau,x) = x \in \mathbb{T}^2 \ . \tag{20.11}$$

Since $u \in C^0([0,+\infty) \times \mathbb{T}^2, \mathbb{T}^2) \cap L^\infty((0,+\infty), V^{1,\infty})$, the classical Cauchy–Lipschitz theorem implies that, for every $x \in \mathbb{T}^2$, there exists a unique solution $\varphi(t;\tau,x) \in C^1([0,+\infty), \mathbb{T}^2)$ and the function $\varphi(t;\tau,x) : x \in \mathbb{T}^2 \mapsto \varphi(t;\tau,x) \in \mathbb{T}^2$ is Lipschitz-continuous with respect to x, where the Lipschitz constant may depend on t. Moreover, the function $\varphi(t;\tau,x) : (t,\tau,x) \mapsto \varphi(t;\tau,x)$ belongs to $C^1([0,+\infty)^2 \times \mathbb{T}^2, \mathbb{T}^2)$. The integral form of the equation (20.11) is as follows

$$\varphi(t;\tau,x) = x + \int_\tau^t u(s,\varphi(s;\tau,x))ds \ . \tag{20.12}$$

Of course, the solution $\varphi(t;\tau,x)$ also depends on u. If we want to emphasize that $\varphi(t;\tau,x)$ also depends on u or when we let u vary, we will use the notation $\varphi_u(t;\tau,x)$ instead of $\varphi(t;\tau,x)$.

In what follows, we will often use the following estimates without further notice. Below Jac φ denotes the Jacobian matrix of φ.

Lemma 20.1. *Let $u \in C^0([0,+\infty), V^{2,p}) \cap L^\infty((0,+\infty), V^{3,p})$, $p > 1$.*

1) Then,

$$(\det \ \textit{Jac}\ \varphi)(t;\tau,x)) = 1 \ , \forall \tau \ , \forall t \ , \forall x \ , \tag{20.13}$$

2) The following estimate holds, for any $1 \le q \le +\infty$, any $t \ge \tau$ (resp. $\tau \ge t$)

$$\|\nabla \varphi(t;\tau,\cdot)\|_{L^q} \le \exp\Big(\int_\tau^t \|\nabla u(s)\|_{L^\infty} ds \ \Big) \tag{20.14}$$

(resp. $\|\nabla \varphi(t;\tau,\cdot)\|_{L^q} \le \exp(\int_t^\tau \|\nabla u(s)\|_{L^\infty} ds)$).

3) Let u_i, $i = 1,2$ be two elements in $C^0([0,+\infty), V^{2,p}) \cap L^\infty((0,+\infty), V^{3,p})$, $p > 1$ and denote $\varphi_{u_i}(t;\tau,x)$ the corresponding solutions of (20.10). Then, for any $1 \le q$, any $t \ge \tau$ (resp. $\tau \ge t$)

$$\|\varphi_{u_1}(t;\tau,\cdot) - \varphi_{u_2}(t;\tau,\cdot)\|_{L^q} \le \|u_1 - u_2\|_{L^\infty((\tau,t),L^q)} \exp\Big(\int_\tau^t \|\nabla u_1(s)\|_{L^\infty} ds \ \Big) \tag{20.15}$$

(resp. $\|\varphi_{u_1}(t;\tau,\cdot) - \varphi_{u_2}(t;\tau,\cdot)\|_{L^q} \le \|u_1 - u_2\|_{L^\infty((t,\tau),L^q)} \exp(\int_t^\tau \|\nabla u_1(s)\|_{L^\infty} ds)$).

Proof. 1) The property (20.13) is well known. It is a consequence of the fact that $\operatorname{div} u = 0$ (see, for example, [11]).

2) Let $t \geq \tau$. We set

$$\psi_k(t;\tau,x) = \frac{\partial}{\partial x_k}\varphi(t;\tau,x) \ .$$

and notice that

$$\partial_t \psi_k(t;\tau,x) = \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}(t,\varphi(t;\tau,x))\frac{\partial \varphi_i}{\partial x_k}(t;\tau,x) \ ,$$

which implies that, for $t \geq \tau$,

$$\|\psi_k(t;\tau,\cdot)\|_{L^q} \leq \|\psi_k(\tau;\tau,\cdot)\|_{L^q} + \int_\tau^t \|\nabla u(s)\|_{L^\infty}\|\psi_k(s;\tau,\cdot)\|_{L^q} ds \ .$$

Noticing that $\psi_k(\tau;\tau,x) = I$ for any x and using the Gronwall inequality, we deduce the estimate (20.14) from the above inequality.

The statement of 3) is proved in the same way.

□

Theorem 20.4. *1) Let $p > 1$. Let $k = 0, 1$, for any $w_0 \in W^{k,p}_{per}$ and any $g \in L^\infty((0,T), W^{k,p}_{per})$, there exists a unique (mild) solution $w(t) \in C^0([0,T], W^{k,p}_{per})$ of (20.10) and $\partial_t w$ belongs to $L^\infty((0,T), W^{k-1,p}_{per})$, where $T > 0$.*

2) For $k \geq 2$, assume that u belongs to $C^0([0,+\infty), V^{k+1,p}) \cap L^\infty((0,+\infty), V^{k+2,p})$, then, for any $w_0 \in W^{k,p}_{per}$ and any $g \in L^\infty((0,T), W^{k,p}_{per})$, there exists a unique (mild) solution $w(t) \in C^0([0,T], W^{k,p}_{per})$ of (20.10) and $\partial_t w$ belongs to $L^\infty((0,T), W^{k-1,p}_{per})$, where $T > 0$.

3) Moreover, we have the following estimate, for any $0 \leq t \leq T$,

$$\begin{aligned}
\|w(t)\|_{L^p} \leq & e^{-\frac{\nu}{\alpha}t}\|w_0\|_{L^p} + \frac{\alpha}{\nu}(1 - e^{-\frac{\nu}{\alpha}t})\|g\|_{L^\infty(I,L^p)} \\
\|\partial_t w(t)\|_{W^{-1,p}} \leq & \big(\frac{\nu}{\alpha} + \|u\|_{L^\infty(I,L^\infty)}\big)\big(e^{-\frac{\nu}{\alpha}t}\|w_0\|_{L^p} + \frac{\alpha}{\nu}(1 - e^{-\frac{\nu}{\alpha}t})\|g\|_{L^\infty(I,L^p)}\big) \\
& + \|g\|_{L^\infty(I,L^p)} \ ,
\end{aligned} \tag{20.16}$$

where $I = (0,T)$. These inequalities hold for any $t \geq 0$, if $w_0 \in W^{0,p}_{per}$ and $g \in L^\infty((0,+\infty), W^{0,p}_{per})$.

Proof. To prove this theorem, we proceed as Beirão da Veiga [2], but replace the Dirichlet boundary conditions by the periodic ones. Let us consider the equation

$$\begin{aligned} &\partial_t w + aw + u \cdot \nabla w = g \ , \quad t > 0, x \in \mathbb{T}^2 \ , \\ &w(0,x) = w_0(x) \ , \quad x \in \mathbb{T}^2 \ , \end{aligned} \tag{20.17}$$

where for simplicity a is a given constant. To solve this equation, Beirão da Veiga considered the differential operator

$$\tilde{A}_a(t)w \equiv aw + u \cdot \nabla w \ , \quad t \in [0,T] \ ,$$

acting in the distributional sense on the functions w on $\Omega \equiv \mathbb{T}^2$. For $k \geq 1$, he introduced the space

$$D^k(t) \equiv \{w \in W^{k,p}_{per} \,|\, u \cdot \nabla w \in W^{k,p}_{per}\},$$

and defined the operator

$$A^k_a(t)w \equiv \tilde{A}_a(t)w \ , \quad \forall w \in D^k(t) \ . \tag{20.18}$$

In the case $k = 0$, one defines the operator A^0_a as the closure in L^p of the operator $A^1_a : D^k(t) \to W^{1,p}_{per}$.

In [2], Beirão da Veiga proved that, under the above regularity hypotheses made on $u(t)$, the family $\{A^k_a(t)\}_{t\in I}$, where $I = [0,T]$ is $(1,\theta_k)$-stable in the sense of Kato ([42, 43] and also [56]), with $\theta_k \geq 0$. Thus, the evolution operator $U_a(t,s)$ associated with the family $\{A^k_a(t)\}_{t\in I}$ is strongly continuous in $W^{k,p}_{per}$, for $k \geq 0$ (see [2, Theorem 2.2 and Sects. 3 and 4]) and, for any $w_0 \in W^{k,p}_{per}$ and any $g \in L^\infty((0,T), W^{k,p}_{per})$, there exists a unique (mild) solution $w(t) \in C^0([0,T], W^{k,p}_{per})$ of (20.17) given by

$$w(t) = U_a(t,0)w_0 + \int_0^t U_a(t,s)g(s)ds \ . \tag{20.19}$$

Moreover, $w(t)$ is a strong solution in $W^{k-1,p}_{per}$, that is, the equality (20.17) holds in $W^{k-1,p}_{per}$ a.e. in t.

One remarks that

$$U_a(t,s) = e^{-a(t-s)}U(t,s) \ , \tag{20.20}$$

where $U(t,s) \equiv U_0(t,s)$ and thus

$$w(t) = e^{-at}U(t,0)w_0 + \int_0^t e^{-a(t-s)}U(t,s)g(s)ds \ . \tag{20.21}$$

Theorem 2.2 of [2] implies that, for any $k \geq 0$, one has, for $0 \leq t \leq T$,

$$\|w(t)\|_{W^{k,p}} \leq \exp(\theta_k T)(\|w_0\|_{W^{k,p}} + \int_0^T \|g(s)\|_{W^{k,p}} ds) \ . \tag{20.22}$$

In [2], Beirão da Veiga also proved that the evolution operator $U_a(t,s)$ is strongly continuous from $W^{-k,p}_{per}$ into itself, for $k \geq 0$, which implies that the estimate (20.22) still holds if k is replaced by $-k$, that is, one has, for $0 \leq t \leq T$,

$$\|w(t)\|_{W^{-k,p}} \leq \exp(\theta_k T)(\|w_0\|_{W^{-k,p}} + \int_0^T \|g(s)\|_{W^{-k,p}} ds) \ . \tag{20.23}$$

We apply the above results with $a = \frac{\nu}{\alpha}$. In our case, we obtain a better estimate for $k = 0$. Indeed, assume first that $w_0 \in W^{1,p}_{per}$ and $g \in L^\infty((0,T), W^{1,p}_{per})$. We first take the inner product of the equality (20.10) with $(\delta + |w|^2)^{(p-2)/2} w$, where $\delta > 0$ is small, then integrate by parts by taking into account that $\operatorname{div} u = 0$ and that $u \in L^\infty((0,T), W^{1,\infty}(\mathbb{T}^2)^2)$, and finally let δ go to zero. Then, we obtain that, for $0 \leq t \leq T$,

$$\partial_t \|w(t)\|_{L^p} + \frac{\nu}{\alpha} \|w(t)\|_{L^p} \leq \|g\|_{L^p} \ . \tag{20.24}$$

Integrating (20.24) with respect to the time variable and applying the Gronwall lemma, we deduce from (20.24) that, for $0 \leq t \leq T$,

$$\|w(t)\|_{L^p} \leq e^{-\frac{\nu t}{\alpha}} \|w_0\|_{L^p} + \int_0^t e^{\frac{\nu}{\alpha}(s-t)} \|g(s)\|_{L^p} ds \leq e^{-\frac{\nu t}{\alpha}} \|w_0\|_{L^p} + \frac{\alpha}{\nu} \|g\|_{L^\infty(I,L^p)} \ . \tag{20.25}$$

This inequality is also valid in the case where the interval $I \equiv (0,T)$ is replaced by $[0,+\infty)$ in the statement of the theorem.

Arguing by density, one readily shows that the inequality (20.25) still holds if w_0 and g only belong to L^p_{per} and $L^\infty((0,T), L^p_{per})$.

By the general theory developed in [43] or [56, Chap. 5], we also know that $\partial_t w$ belongs to $L^\infty((0,T), W^{k-1,p})$. If moreover g is in $C^0([0,T], W^{k,p})$, then $\partial_t w$ belongs to $C^0([0,T], W^{k-1,p})$. Using the equality (20.10) and the inequality (20.25), we also show by density as above that, for $t \geq 0$,

$$\begin{aligned} \|\partial_t w(t)\|_{W^{-1,p}} \leq \big(\frac{\nu}{\alpha} + \|u\|_{L^\infty(I,L^\infty)}\big)\big(e^{-\frac{\nu}{\alpha}t}\|w_0\|_{L^p} + \frac{\alpha}{\nu}(1 - e^{-\frac{\nu}{\alpha}t})\|g\|_{L^\infty(I,L^p)}\big) \\ + \|g\|_{L^\infty(I,L^p)} \ . \end{aligned} \tag{20.26}$$

Finally, let $\tilde{w}$ be the solution of the equation (20.10), where u, g, and w_0 are replaced by $\tilde{u}$, $\tilde{g}$ and $\tilde{w}_0$, respectively. Then, $W = \tilde{w} - w$ is a solution of the equation

$$\begin{aligned}&\partial_t W+\frac{\nu}{\alpha}W+\tilde{u}\cdot\nabla W=\tilde{g}-g+(u-\tilde{u})\cdot\nabla w\,,\quad t>0, x\in\mathbb{T}^2\,,\\&W(0,x)=\tilde{w}_0(x)-w_0(x)\,,\quad x\in\mathbb{T}^2\,,\end{aligned}\tag{20.27}$$

Assume that w_0, $\tilde{w}_0$ and g, $\tilde{g}$ belong to $W^{1,p}_{per}$ and $L^\infty((0,T),W^{1,p}_{per})$, respectively. Applying the estimate (20.24) to the equation (20.27), we obtain, that, for $0\le t\le T$,

$$\begin{aligned}\partial_t\|W(t)\|_{L^p}+\frac{\nu}{\alpha}\|W(t)\|_{L^p}\le&\big(\|(\tilde{u}-u)\nabla w\|_{L^p}+\|\tilde{g}-g\|_{L^p}\big)\\\le&\|(\tilde{u}-u)(t)\|_{L^\infty}\|w(t)\|_{W^{1,p}}+\|(\tilde{g}-g)(t)\|_{L^p}\,.\end{aligned}\tag{20.28}$$

Integrating with respect to t and taking into account the inequality (20.22), we finally get the following estimate, for $0\le t\le T$,

$$\begin{aligned}\|W(t)\|_{L^p}\le&e^{-\frac{\nu t}{\alpha}}\|w_0-\tilde{w}_0\|_{L^p}+\frac{\alpha}{\nu}\|g-\tilde{g}\|_{L^\infty(I,L^p)}\\&+\frac{\alpha}{\nu}\|(\tilde{u}-u)(t)\|_{L^\infty(I,L^\infty)}\big[\exp(\theta_1 T)(\|w_0\|_{W^{1,p}}\\&+\int_0^T\|g(s)\|_{W^{1,p}}ds)\big]\,.\end{aligned}\tag{20.29}$$

Using the estimates (20.23) and (20.25), we also show that, for $0\le t\le T$,

$$\begin{aligned}\|W(t)\|_{W^{-1,p}}\le\,&e^{\theta_1 T}\Big[\|w_0-\widetilde{w_0}\|_{W^{-1,p}}+\|g-\tilde{g}\|_{L^1(I,W^{-1,p})}\\&+\|(\tilde{u}-u)(t)\|_{L^1(I,L^\infty)}(e^{-\frac{\nu t}{\alpha}}\|w_0\|_{L^p}+\tfrac{\alpha}{\nu}\|g\|_{L^\infty(I,L^p)})\Big]\,.\end{aligned}\tag{20.30}$$

By density, the above inequality also holds if w_0, $\tilde{w}_0$ and g, $\tilde{g}$ only belong to L^p_{per} and $L^\infty((0,T),L^p_{per})$, respectively. □

Remark 20.1. In [44], Ladyzenskaya and Solonnikov proved that, if the data w_0 and f are regular enough, the solution w of (20.17) with $a=0$ is given by

$$w(t,x)=w_0(\varphi(0;t,x))+\int_0^t g(s,\varphi(s;t,x))ds\,.\tag{20.31}$$

This implies by uniqueness of the solution that, if $w_0\in W^{k,p}_{per}$ and $g\in L^\infty((0,T),W^{k,p}_{per})$, for $k\ge 0$, the solution w of the equation (20.10) is given by

$$w(t,x)=e^{-\frac{\nu}{\alpha}t}w_0(\varphi(0;t,x))+\int_0^t e^{-\frac{\nu}{\alpha}(t-s)}g(s,\varphi(s;t,x))ds\tag{20.32}$$

The integral formula allows to prove the above estimates in another (elegant) way, without using the inequalities of [2]. Let W be the solution of (20.27). It satisfies the integral equation:

$$\begin{aligned} W(t,x) =& e^{-\frac{\nu}{\alpha}t}\big(\tilde{w}_0(\tilde{\varphi}(0;t,x)) - w_0(\tilde{\varphi}(0;t,x))\big) \\ &+ \int_0^t e^{-\frac{\nu}{\alpha}(t-s)}(\tilde{g}(s,\tilde{\varphi}(s;t,x)) - g(s,\tilde{\varphi}(s;t,x)))ds \\ &+ \int_0^t e^{-\frac{\nu}{\alpha}(t-s)}(u-\tilde{u})(s,\tilde{\varphi}(s;t,x))\cdot\nabla w(\tilde{\varphi}(s;t,x))ds\,, \end{aligned} \tag{20.33}$$

where φ and $\tilde{\varphi}$ are the solutions of the equation (20.11) associated with u and $\tilde{u}$, respectively. From the equality (20.33), we at once deduce, by applying Lemma 20.1 and (20.25), that, for $0 \le t \le T$,

$$\begin{aligned} \|W(t)\|_{W^{-1,p}} \le& C^*(T,\tilde{u})\Big(e^{-\frac{\nu}{\alpha}t}\|w_0 - \widetilde{w_0}\|_{W^{-1,p}} + \frac{\alpha}{\nu}\|g-\tilde{g}\|_{L^\infty(I,W^{-1,p})} \\ &+ \frac{\alpha}{\nu}\|(\tilde{u}-u)(t)\|_{L^\infty(I,L^\infty)}\|w\|_{L^\infty(L^p)}\Big) \\ \le& C^*(T,\tilde{u})\Big[e^{-\frac{\nu}{\alpha}t}\|w_0 - \widetilde{w_0}\|_{W^{-1,p}} + \frac{\alpha}{\nu}\|g-\tilde{g}\|_{L^\infty(I,W^{-1,p})} \\ &+ \frac{\alpha}{\nu}\|(\tilde{u}-u)(t)\|_{L^\infty(I,L^\infty)}\big(\|w_0\|_{L^p} + \frac{\alpha}{\nu}\|g\|_{L^\infty(I,L^p)}\big)\Big]\,, \end{aligned} \tag{20.34}$$

where $C^*(T,\tilde{u}) = \exp\int_0^T \|\nabla\tilde{u}(s)\|_{L^\infty}ds$.

We point out that the above inequality also holds if w_0, $\tilde{w}_0$ and g, $\tilde{g}$ only belong to L^p_{per} and $L^\infty((0,T),L^p_{per})$, respectively.

We are actually interested in the solution of the following transport equation

$$\begin{aligned} &\partial_t\omega + \frac{\nu}{\alpha}\omega + u\cdot\nabla\omega = \operatorname{rot} f + \frac{\nu}{\alpha}\operatorname{rot} u, \quad t>0,\ x\in\mathbb{T}^2 \\ &\omega(0,x) = \omega_0(x)\,, \quad x\in\mathbb{T}^2\,, \end{aligned} \tag{20.35}$$

where $u \in C^0([0,+\infty),V^{2,p}) \cap L^\infty((0,+\infty),V^{3,p})$, $p>1$.

As an immediate consequence of Theorem 20.4, we obtain the following corollary.

Corollary 20.2. *Let $p>1$. For any $\omega_0 \in L^p_{per}$ and $\operatorname{rot} f \in L^\infty((0,T),L^p_{per})$, there exists a unique (mild) solution $\omega \in C^0([0,T],L^p_{per})$ (with $\partial_t\omega \in L^\infty((0,T),W^{-1,p}_{per})$) of the equation (20.35). Moreover, the following estimate holds, for $t\ge 0$,*

$$\|\omega(t)\|_{L^p} \le e^{-\frac{\nu}{\alpha}t}\|\omega_0\|_{L^p} + \frac{\alpha}{\nu}\|\operatorname{rot} f\|_{L^\infty(I,L^p)} + (1-e^{-\frac{\nu}{\alpha}t})\|\operatorname{rot} u\|_{L^\infty(I,L^p)}\,. \tag{20.36}$$

This inequality holds for any $t \geq 0$, if $I \equiv (0,T)$ is replaced above by $I \equiv (0,+\infty)$.

The upper bound for $\|\partial_t \omega\|_{W^{-1,p}}$ follows from (20.26).

20.2.2 An Auxiliary Problem

In Corollary 20.2 we have obtained the solution ω of the equation (20.35). We next want to show that there exists a unique divergence-free vector field z such that $\omega = z - \alpha \Delta z$. This will be an easy consequence of the following two lemmas.

Lemma 20.2. *1) For any $w \in W^{k,p}_{per}$, $k \geq 0$, there exists a unique vector field $\psi \in V^{k+1,p}$ such that*

$$\operatorname{rot} \psi(x) = w(x) \ , \quad \forall x \in \mathbb{T}^2. \tag{20.37}$$

Moreover, there exists a positive constant $C_0(k)$ such that,

$$\|\psi\|_{W^{k+1,p}} \leq C_0(k) \|u\|_{W^{k,p}} \ . \tag{20.38}$$

2) Likewise, for any $w \in W^{m,\infty}((0,T), W^{k,p}_{per})$ (resp. in $C^m([0,T], W^{k,p}_{per})$), $k \geq 0$, $m \geq 0$, there exists a unique vector field $\psi \in W^{m,\infty}((0,T), V^{k+1,p})$ (resp. in $C^m([0,T], V^{k+1,p})$) such that the equality (20.37) holds.

Proof. 1) For any $w \in W^{k,p}_{per}$, following [1, Lemma 2.3] for example, we construct the vector field

$$\psi = \nabla^{\perp} G(w) \ , \tag{20.39}$$

where $G(w)$ is the solution of the problem: to find $G(w) \in W^{1,p}_{per}$ such that,

$$\Delta G(w) = w \ . \tag{20.40}$$

The solution $G(w)$ is unique in $W^{1,p}_{per}$. The regularity of ψ is a consequence of the regularity properties of the solutions of the Laplace equation.

We remark that the vector field ψ is unique in $V^{0,p}_{per}$. Indeed, if ψ_1 and ψ_2 are two solutions of (20.37), then $\Delta(\psi_1 - \psi_2) = 0$, which has a unique solution in $V^{0,p}_{per}$. Statement 2) is proved in the same way.

□

We next show that $w \in W^{k,p}_{per}$ can be written in the form $w = z - \alpha \Delta z$, where $z \in V^{k+3,p}$.

Lemma 20.3. *1) For any $w \in W^{k,p}_{per}$, $k \geq 0$, there exists a unique vector field $z \in V^{k+3,p}$ such that*

$$\operatorname{rot}(z - \alpha \Delta z)(x) = w(x) \ , \quad \forall x \in \mathbb{T}^2 \ . \tag{20.41}$$

Moreover, there exists a positive constant $C_1(k,\alpha)$ *such that,*

$$\|z\|_{W^{k+3,p}} \leq C_1(k,\alpha)\|u\|_{W^{k,p}} \,. \tag{20.42}$$

2) Likewise, for any $w \in W^{m,\infty}((0,T),W^{k,p}_{per})$ *(resp. in* $C^m([0,T],W^{k,p}_{per})$*),* $k \geq 0$*,* $m \geq 0$*, there exists a unique vector field* $z \in W^{m,\infty}((0,T),V^{k+3,p})$ *(resp. in* $C^m([0,T],V^{k+3,p})$*) such that the equality* (20.41) *holds.*

Proof. 1) By Lemma 20.2, we know that there exists a unique vector field $\psi \in V^{k+1,p}$ such that $\operatorname{rot}\psi = w$ (and ψ is unique in $V^{1,p}$). But, it is well known that the problem: to find $z \in V^{1,p}$ such that

$$z - \alpha\Delta z = \psi \tag{20.43}$$

has a unique solution. Moreover, the regularity properties of the Laplacian operator imply that $z \in V^{k+3,p}$ and that the inequality (20.42) holds. Statement 2) is proved in the same way.

□

From the Corollary 20.2 and the Lemmata 20.2 and 20.3, we at once deduce the following corollary.

Corollary 20.3. *Let* $p > 1$*. For any* $\omega_0 \in L^p_{per}$ *and* $\operatorname{rot} f \in L^\infty((0,T),L^p_{per})$*, there exists a unique* $z \in C^0([0,T],V^{3,p})$ *(with* $\partial_t z \in L^\infty((0,T),V^{2,p})$*) such that*

$$\omega = \operatorname{rot}(z - \alpha\Delta z) \tag{20.44}$$

is the unique (mild) solution of the equation (20.35)*. Moreover, the following estimates hold, for* $0 \leq t \leq T$*,*

$$\begin{aligned}
\|z(t)\|_{W^{3,p}} \leq & C_1(0,\alpha)\big[e^{-\frac{\nu}{\alpha}t}\|\omega_0\|_{L^p} + (1-e^{-\frac{\nu}{\alpha}t})(\frac{\alpha}{\nu}\|\operatorname{rot} f\|_{L^\infty(I,L^p)} \\
& + \|\operatorname{rot} u\|_{L^\infty(I,L^p)})\big] \\
\|\partial_t z(t)\|_{W^{2,p}} \leq & C_2(\alpha)\big[\big(\frac{\nu}{\alpha} + \|u\|_{L^\infty(I,L^p)}\big) \\
& \times \big(e^{-\frac{\nu}{\alpha}t}\|\omega_0\|_{L^p} + (1-e^{-\frac{\nu}{\alpha}t})(\frac{\alpha}{\nu}\|\operatorname{rot} f\|_{L^\infty(I,L^p)} + \|\operatorname{rot} u\|_{L^\infty(I,L^p)})\big) \\
& + \|\operatorname{rot} f\|_{L^\infty(I,L^p)} + \frac{\nu}{\alpha}\|\operatorname{rot} u\|_{L^\infty(I,L^p)}\big] \,,
\end{aligned} \tag{20.45}$$

where $C_2(\alpha)$ *is a positive constant depending only on* α*.*

These inequalities hold for any $t \geq 0$*, if* $I \equiv (0,T)$ *is replaced above by* $I \equiv (0,+\infty)$*.*

Proof. The existence and uniqueness of $z(t) \in C^0([0,T],W^{3,p}_{per})$, such that $\omega = \operatorname{rot}(z - \alpha\Delta z)$ is the mild solution of (20.35), is a direct consequence of Corollary 20.2 and of Lemma 20.3.

Taking the derivative of (20.44) with respect to t, we obtain the equality

$$\mathrm{rot}\,\partial_t z - \alpha\Delta \mathrm{rot}\,\partial_t z = \partial_t \omega\ .$$

Since $\partial_t\omega$ belongs to $W^{-1,p}_{per}$, the regularity properties of the above equation imply that $\mathrm{rot}\,\partial_t z$ is in $W^{2,p}_{per}$ and thus $\partial_t z$ belongs to $W^{3,p}_{per}$. The inequalities (20.45) are a direct consequence of the inequalities of Corollary 20.2 and of Lemma 20.3 and of (20.26). □

20.2.3 *Local Existence and Uniqueness of Solutions in $V^{3,p}$, $p > 1$*

Let $u_0 \in V^{3,p}$ be given. We first remark that, if ω is a solution of the transport equation (20.35) with $\omega_0 = \mathrm{rot}(u_0 - \alpha\Delta u_0)$ and z is the solution of (20.44), then there exists a unique pressure $p \in W^{1,p}_{per}$ such that,

$$\begin{aligned}
&\partial_t(z - \alpha\Delta z) - \nu\Delta z + \mathrm{rot}\,(z - \alpha\Delta z)\times u + \nabla p = f, \quad t > 0,\ x \in \Omega,\\
&\mathrm{div}\,u = 0, \quad t > 0,\ x \in \Omega,\\
&z(0,x) = u_0(x), \quad x \in \Omega\ .
\end{aligned} \tag{20.46}$$

Local Existence of the Solution of (20.2)

Now we are ready to show the local existence of solutions of (20.2). Let $T > 0$ be fixed (the choice of T will be made more precise below). Let $u_0 \in V^{3,p}$ and $f \in L^\infty((0,T), W^{1,p}_{per})$ be given.

As we have explained in the introduction, we define the following map $J_T : L^\infty((0,T), V^{3,p}) \cap C^0([0,T], V^{2,p})$ into itself as follows

$$u \in L^\infty((0,T), V^{3,p}) \cap C^0([0,T], V^{2,p}) \mapsto \omega \in C^0([0,T], L^p_{per}) \mapsto z \in C^0([0,T], V^{2,p})\ , \tag{20.47}$$

where ω is the solution of the equation (20.35) with $\omega_0 = \mathrm{rot}(u_0 - \alpha\Delta u_0)$ and z is the solution of the equation (20.44). We will show that J_T is a continous compact map from a closed convex subset E_T of $L^\infty((0,T), V^{3,p}) \cap C^0([0,T], V^{2,p})$ into E_T. Then applying the Leray–Schauder fixed point theorem, we deduce that J_T has a fixed point u^*. We notice that the idea of introducing such a type of map and of using the Leray–Schauder theorem goes back to [25], where the local existence of solutions has been proved (see also [8, 23, 45], for example). The map, constructed in these papers differs from the one here. Indeed, these authors considered the map $\mathscr{F} : w \mapsto u \mapsto \mathscr{F}(w) = \omega$, where u satisfies $\mathrm{div}\,u = 0$ and $w = \mathrm{rot}\,(u - \alpha\Delta u)$ and where ω is the solution of (20.35). In [25], and [23], the authors work in more

regular spaces. Even if there are some differences, our proof follows the same main lines.

First we introduce the positive constant K given by

$$K \equiv C_1(0,\alpha)\big(\|\omega_0\|_{L^p} + \frac{\alpha}{\nu}\|\mathrm{rot}\, f\|_{L^\infty(I,L^p)}\big)\ , \tag{20.48}$$

where $C_1(0,\alpha)$ is given in Corollary 20.3 and then choose $T > 0$ such that

$$2KC_1(0,\alpha)(1 - e^{-\frac{\nu}{\alpha}T}) < K\ . \tag{20.49}$$

Finally, we define the (non empty) set

$$E_T \equiv \{v \in L^\infty((0,T),V^{3,p}) \cap C^0([0,T],V^{2,p}) \mid v(0) = u_0\ ,\ \|v\|_{L^\infty(W^{3,p})} \leq 2K\}\ . \tag{20.50}$$

We equip E_T with the classical topology of the space $X_T = C^0([0,T],V^{2,p})$.

First, one checks that E_T is a closed subset of X_T. As in [23] or in [9], one considers a sequence v_n, $n \in \mathbb{N}$, in X_T converging to v. Since the sequence v_n is bounded in $L^\infty((0,T),V^{3,p})$, it converges in $L^\infty((0,T),V^{3,p})$ weak $*$ to an element U in $L^\infty((0,T),V^{3,p})$ and

$$\|U\|_{L^\infty((0,T),V^{3,p})} \leq 2K\ .$$

Due to the uniqueness of the limit in the space of distributions in $(0,T)\times\mathbb{T}^2$, $U = v$ and thus v belongs to E_T.

With the above choice of K, Corollary 20.3 at once implies that $J(E_T) \subset E_T$.

Actually, $J(E_T)$ is relatively compact in E_T. Indeed, by Corollary 20.3, $J(E_T)$ is bounded in $W^{1,\infty}((0,T),W^{2,p}_{per}) \cap L^\infty((0,T),V^{3,p})$. Since the injection of $W^{3,p}$ into $W^{2,p}$ is compact, we deduce from [46, Assertion(12.10), page 142] that every bounded set in $W^{1,\infty}((0,T),W^{2,p}(\mathbb{T}^2)) \cap L^\infty((0,T),W^{3,p}(\mathbb{T}^2))$ is relatively compact in $C^0([0,T],W^{2,p}(\mathbb{T}^2))$. Thus, $J(E_T)$ is relatively compact in E_T.

It remains to verify that the map $J : u \in E_T \mapsto \omega \mapsto z \in E_T$ is continuous for the topology of X_T. Let u_1 and u_2 be two elements of E_T, let ω_1 and ω_2 be the two corresponding solutions of the equation (20.35) and finally let z_1, z_2 be the two corresponding solutions of (20.44). From the estimate (20.34) in Remark 20.1, we deduce that, for $0 \leq t \leq T$,

$$\begin{aligned}
&\|(\omega_1 - \omega_2)(t)\|_{W^{-1,p}} \\
&\leq C^*(T,u_2)\Big[\frac{\alpha}{\nu}\|u_1 - u_2\|_{L^\infty(I,L^\infty)}\big(\|\omega_1(0)\|_{L^p} + \|\mathrm{rot}\, f\|_{L^\infty(I,L^p)} \\
&\qquad + \frac{\nu}{\alpha}\|\mathrm{rot}\, u_1\|_{L^\infty(I,L^p)}\big) + \|u_1 - u_2\|_{L^\infty(I,L^p)}\Big] \\
&\leq C(K)\|u_1 - u_2\|_{L^\infty(I,W^{2,p})}\ ,
\end{aligned} \tag{20.51}$$

where $C(K)$ is a positive constant depending on K. Using the regularity properties of the Laplacian and arguing as in Corollary 20.3, one deduces from the inequality (20.51) that, for $0 \leq t \leq T$,

$$\|(z_1 - z_2)(t)\|_{W^{2,p}} \leq C\|(\omega_1 - \omega_2)(t)\|_{W^{-1,p}} \leq CC(K)\|u_1 - u_2\|_{L^\infty(I,W^{2,p})} . \tag{20.52}$$

From the inequality (20.52), one at once deduces that the map J is continuous for the topology of X_T.

Now we may apply the Leray–Schauder fixed point theorem to the map J. Thus, there exists a fixed point u of J, that is, a function $u \in E_T$ satisfying the system (20.7). Moreover, by Theorem 20.4 and by Corollary 20.3, u belongs to $C^0([0,T], V^{3,p}) \cap W^{1,\infty}((0,+\infty), W^{2,p}_{per})$ and $u \equiv z$ satisfies the estimates (20.45). Moreover, applying Theorem 2.2 of [2] in the "negative order Sobolev space" $W^{-1,p}_{per}$, we deduce that $\partial_t u$ actually belongs to $C^0([0,T], W^{2,p}_{per})$. Finally, introducing the pressure term as in (20.46), we have proved that the system (20.2) admits a solution (u,p) if $T > 0$ is small enough (T depending only on u_0 and f).

The propagation of the regularity of u is a direct consequence of Theorem 20.4. Assume that $u_0 \in V^{4,p}$ and $f \in L^\infty((0,T), W^{2,p}_{per})$, then, in the equation (20.35), ω_0 and $\mathrm{rot}\, f + \frac{\nu}{\alpha}\mathrm{rot}\, u$ belong to $V^{1,p}$ and $L^\infty((0,T), W^{1,p}_{per})$, respectively. Thus, by Theorem 20.4, the solution ω of (20.35) belongs to $C^0((0,T), W^{1,p}_{per})$. Since $\omega = \mathrm{rot}\,(u - \alpha\Delta u)$, it follows that u belongs to $C^0((0,T), V^{4,p}_{per})$. When $k \geq 2$, we proceed by recursion on k. Indeed, if $u_0 \in V^{k+3,p}$, $f \in L^\infty((0,T), W^{k+1,p}_{per})$, then, by Theorem 20.4, u belongs to $C^0([0,T], V^{k+3,p})$, provided that u is in $C^0([0,T], V^{k+1,p}) \cap L^\infty((0,T), W^{k+2,p}_{per})$. But this regularity property is known by application of Theorem 20.4 at the order $k-1$.

Uniqueness of the Solution of (20.2)

The proof of the uniqueness of the solutions of (20.2) is well known and goes back to [14]. For the sake of completeness, we give a quick proof of it. Actually, we will prove a more general continuity result. Let $u_i(t) \in C^0([0,T], V^{3,p})$ (with $\partial_t u_i(t) \in L^\infty((0,T), V^{2,p})$), $i = 0, 1$, be two solutions of (20.2). Then, $U = u_1 - u_2$ satisfies the equation

$$\begin{aligned} \partial_t(U - \alpha\Delta U) - \nu\Delta U + \mathrm{rot}\,(U - \alpha\Delta U) \times u_1 + \mathrm{rot}\,(u_2 - \alpha\Delta u_2) \times U \\ = -\nabla(p_1 - p_2)\,, \quad t > 0,\ x \in \Omega\,, \\ U(0,x) = u_1(0) - u_2(0)\,, \quad x \in \Omega\,. \end{aligned} \tag{20.53}$$

In [55, Theorem A.1], we have shown the following equality

$$(\operatorname{rot} \Delta U \times u_1, U) \equiv \int_{\mathbb{T}^2} \operatorname{rot} U \left(\Delta u_1^1 U^2 - \Delta u_1^2 U^1\right) dx + 2 \int_{\mathbb{T}^2} \operatorname{rot} U \left(\nabla u_1^1 \cdot \nabla U^2 - \nabla u_1^2 \cdot \nabla U^1\right) dx. \tag{20.54}$$

Taking the inner product of the first equation in (20.53) with U in L^2 and using the equality (20.54) together with classical Sobolev inequalities, we obtain, for $0 \leq t \leq T$,

$$\begin{aligned}
&\partial_t(\|U(t)\|_{L^2}^2 + \alpha\|\nabla U(t)\|_{L^2}^2) + \nu\|\nabla U(t)\|_{L^2}^2 \\
&\leq 2\Big|\int_{\mathbb{T}^2} (\operatorname{rot}(U - \alpha\Delta U) \times u_1) U dx\Big| \\
&\leq C_1\Big(\|u_1(t)\|_{L^\infty}\|U\|_{L^2}\|\nabla U\|_{L^2} + \alpha\|\nabla u_1(t)\|_{L^\infty}\|\nabla U\|_{L^2}^2 \\
&\qquad + \alpha\|\nabla U\|_{L^2}\|\Delta u_1\|_{W^{1,p}}\|U\|_{L^{\frac{p}{p-1}}}\Big) \\
&\leq C_2\Big(\|u_1(t)\|_{L^\infty}\|U\|_{L^2}\|\nabla U\|_{L^2} + \alpha\|\nabla u_1(t)\|_{L^\infty}\|\nabla U\|_{L^2}^2 + \alpha\|\nabla U\|_{L^2}^2\|\Delta u_1\|_{W^{1,p}}\Big) \\
&\leq C_3\|u_1(t)\|_{W^{3,p}}\Big(\frac{1+\alpha}{\alpha}\Big)\big(\|U\|_{L^2}^2 + \alpha\|\nabla U\|_{L^2}^2\big)\,.
\end{aligned} \tag{20.55}$$

Integrating with respect to t and applying the Gronwall lemma, we obtain that, for $0 \leq t \leq T$,

$$\|U(t)\|_{L^2}^2 + \alpha\|\nabla U(t)\|_{L^2}^2 \leq \big[\|u_1(0) - u_2(0)\|_{L^2}^2 + \alpha\|\nabla(u_1 - u_2)(0)\|_{L^2}^2\big] \times \exp\int_0^T C_3\Big(\frac{1+\alpha}{\alpha}\Big)\|u_1(s)\|_{W^{3,p}} ds\,. \tag{20.56}$$

If $u_1(0) = u_2(0)$, then (20.56) implies that $U(t) \equiv 0$, that is, that the solution $u(t)$ of (20.2) is unique.

Continuity of the Map $u_0 \in V^{3,p} \mapsto u(t) \in V^{3,p}$

Let $f \in L^\infty((0,T), W^{1,p}_{per})$ (resp. $f \in L^\infty((0,+\infty), W^{1,p}_{per})$) be given. In the next section, we will show that the solution $u(t,x) \equiv u(t,x;u_0)$, with $u(0,x;u_0) = u_0(x)$ of (20.2) exists on $(0,T)$ (resp. $(0,+\infty)$) and is uniformly bounded in time for u_0 belonging to bounded sets of $V^{3,p}$. So we do not need to worry about blow-

up in finite time. To simplify the notation, we will sometimes only write $u(t;u_0)$ instead of $u(t,x;u_0)$.

Assume that f belongs to $L^\infty((0,T),W^{1,p}_{per})$. The estimate (20.56) implies that the map $u_0 \in V^{3,p} \mapsto u(t;u_0) \in V^{1,p}$ is continuous and even Lipschitzian on the bounded sets of $V^{3,p}$. Since, for any bounded set B_0 in $V^{3,p}$, there exists a bounded set $\gamma^+(B_0) \in V^{3,p}$ such that $u(t;u_0) \in \gamma^+(B_0)$ for any $0 \le t \le T$ and any $u_0 \in B_0$, we deduce, by interpolation, that, for every $0 \le \theta < 3$, the map $u_0 \mapsto u(t;u_0) \in V^{\theta,p}$ is Hölder continuous on the bounded sets of $V^{3,p}$ and, in particular, $u_0 \to u(t)$ belongs to $C^0(V^{3,p},V^{\theta,p}) \cap L^\infty(V^{3,p},V^{3,p})$. We next prove that actually $u_0 \to u(t)$ belongs to $C^0(V^{3,p},V^{3,p})$.

Below, we set $\omega(t,x;u_0) = \mathrm{rot}\,(u(t,x;u_0) - \alpha\Delta u(t,x;u_0))$, $\omega(x;u_0) = \mathrm{rot}\,(u_0(x) - \alpha\Delta u_0(x))$ and we denote $\varphi_{u_0}(t;\tau,x)$ the solution of the equation (20.11), where $u(t)$ is replaced by $u(t,x;u_0)$. We recall that $\omega(t,x;u_0)$ writes, for $0 \le t \le T$,

$$\begin{aligned}\omega(t,x;u_0) = {} & e^{-\frac{\nu}{\alpha}t}\omega(\varphi_{u_0}(0;t,x);u_0) \\ & + \int_0^t e^{-\frac{\nu}{\alpha}(t-s)}\big(\mathrm{rot}\,f(s,\varphi_{u_0}(s;t,x)) + \frac{\nu}{\alpha}\mathrm{rot}\,u(s,\varphi_{u_0}(s;t,x);u_0)\big)ds\,.\end{aligned} \tag{20.57}$$

Let u_{0n} be a sequence converging to u_0 in $V^{3,p}$; we want to show that $\omega(t;u_{0n})$ converges to $\omega(t;u_0)$ in $L^\infty((0,T),L^p)$ when n goes to $+\infty$. We at once remark that, for $0 \le t \le T$,

$$\begin{aligned}&\|\omega(\varphi_{u_0}(0;t,x);u_0) - \omega(\varphi_{u_{0n}}(0;t,x);u_{0n})\|_{L^p} \\ &\quad \le \|\omega(\varphi_{u_0}(0;t,x);u_0) - \omega(\varphi_{u_{0n}}(0;t,x);u_0)\|_{L^p} + (1+\alpha)\|u_0 - u_{0n}\|_{V^{3,p}} \\ &\quad \le \|\omega_m(\varphi_{u_0}(0;t,x)) - \omega_m(\varphi_{u_{0n}}(0;t,x))\|_{L^p} + (1+\alpha)\|u_0 - u_{0n}\|_{V^{3,p}} \\ &\qquad + 2\|\omega(\cdot;u_0) - \omega_m(\cdot)\|_{L^p}\,,\end{aligned} \tag{20.58}$$

where ω_m is a sequence in $W^{3,p}_{per}$ converging to ω in L^p. Next we use the Taylor formula and apply Lemma 20.1 to obtain, for $0 \le t \le T$,

$$\begin{aligned}\|\omega_m(\varphi_{u_0}(0;t,x)) - \omega_m(\varphi_{u_{0n}}(0;t,x))\|_{L^p} &\le C(T)\|\nabla\omega_m\|_{L^\infty}\|\varphi_{u_0}(0;t,\cdot) - \varphi_{u_{0n}}(0;t,\cdot)\|_{L^p} \\ &\le C(T,\|u_0\|_{V^{3,p}})C(T)\|\nabla\omega_m\|_{L^\infty}\|u_{0n}-u_0\|_{W^{2,p}}\end{aligned} \tag{20.59}$$

The inequalities (20.58) and (20.59) show that the map $u_0 \in V^{3,p} \mapsto \omega(t,x;u_0) \in L^p$ is continuous. In the same way, we prove that the map $u_0 \in V^{2,p} \mapsto \mathrm{rot}\,f(s,\varphi_{u_0}(s;t,x))$ is continuous (uniformly with respect to s). To show the continuity (uniformly with respect to s) of $\mathrm{rot}\,u(s,\varphi_{u_0}(s;t,x);u_0)$, we argue in the same way and in addition we use the fact that there exists $0 < \theta < 3$ such that $\|\mathrm{rot}\,u(s,y;u_0) - \mathrm{rot}\,u(s,y;u_{0n})\|_{L^p} \le C(\theta,u_0)\|u_0 - u_{0n}\|_{V^{\theta,p}}$.

Remark 20.2. We notice that the local existence of solutions as well as the continuity properties also hold for negative time, if the force f belongs to $L^\infty((-T,0),W^{k+1,p})$, $k \geq 0$, $1 < p < +\infty$.

Remark 20.3. Mutadis mutandis, one can also use the above method of proof to show the corresponding local existence and continuity of the solutions of (20.2), when the periodic boundary conditions are replaced by homogeneous Dirichlet ones, provided the domain Ω is smooth enough (of class C^2) and simply connected. We emphasize that the proof of Bresch and Lemoine [8] of local existence of solutions u in the spaces $V^{2,q}$, $q > 2$, requires less regularity of the domain Ω since they do not consider the transport equation satisfied by ω.

20.2.4 Global Existence of Solutions in $V^{3,p}$, $p > 1$

Let u_0 be given in $V^{3,p}$. We assume here that f belongs to $L^\infty(\mathbb{R}^+, W^{1,p}_{per})$. We set

$$T^*(u_0) = \sup\{T > 0 \,|\, (20.7) \text{ has a solution } \omega = (u - \alpha\Delta u) \in C^0([0,T], L^p_{per})\}\ .$$

The proof of the local existence implies that $T^*(u_0) > 0$. If $T^*(u_0) < +\infty$, then $\|\omega(t)\|_{L^p}$ goes to infinity, when t goes to $T^*(u_0)$. Indeed, if this is not true, then there exist $r > 0$ and a sequence t_n converging to $T^*(u_0)$, with $t_n < T^*(u_0)$ such that $\|\omega(t_n)\|_{L^p} \leq r$, for any n. Due to the proof of the local existence (in particular, see the choices of K and T in (20.48) and in (20.49)), there exists $\tilde{T}(r) > 0$ such that, for any n, $\omega(t)$, which exists on $[0,t_n]$ extends to $[0,t_n + \tilde{T}(r)]$. But, for n large enough, $t_n + \tilde{T}(r) > T^*(u_0)$, which is a contradiction. Thus $\|\omega(t)\|_{L^p}$ goes to infinity, when t goes to $T^*(u_0)$.

By the inequality (20.16) in Theorem 20.4, $\omega = \mathrm{rot}\,(u - \alpha\Delta u)$ satisfies the following estimate for $0 \leq t < T^*(u_0)$,

$$\|\mathrm{rot}\,(u-\alpha\Delta u)(t)\|_{L^p} \leq e^{-\frac{\nu}{\alpha}t}\|\mathrm{rot}\,(u_0-\alpha\Delta u_0)\|_{L^p} + \frac{\alpha}{\nu}\|\mathrm{rot}\,f\|_{L^\infty(\mathbb{R}^+,L^p)} + \|\mathrm{rot}\,u\|_{L^\infty((0,t),L^p)}\ . \tag{20.60}$$

It remains to bound the term $\|\mathrm{rot}\,u\|_{L^\infty(\mathbb{R}^+,L^p)}$. We will first estimate this term for $1 < p \leq 2$. We proceed as in the proof of the uniqueness (see also [55]). Taking the inner product of the first equation in (20.2) with u and using the Young inequality, we obtain, for $0 \leq t < T^*(u_0)$,

$$\partial_t(\|u(t)\|^2_{L^2} + \alpha\|\nabla u(t)\|^2_{L^2}) + \nu\|\nabla u(t)\|^2_{L^2} \leq \frac{1}{\nu\lambda_1}\|f(t)\|^2_{L^2}\ ,$$

where $\lambda_1 > 0$ is the first eigenvalue of the Stokes operator in L^p. From the above inequality, we deduce that, for $0 \leq t < T^*(u_0)$,

$$\partial_t(\|u(t)\|_{L^2}^2+\alpha\|\nabla u(t)\|_{L^2}^2)+\frac{\nu}{2(\lambda_1^{-1}+\alpha)}(\|u(t)\|_{L^2}^2+\alpha\|\nabla u(t)\|_{L^2}^2)+\frac{\nu}{2}\|\nabla u(t)\|_{L^2}^2 \leq \frac{1}{\nu\lambda_1}\|f(t)\|_{L^2}^2 . \tag{20.61}$$

Integrating the inequality (20.61) and applying the Gronwall lemma, we obtain, for $0 \leq t < T^*(u_0)$,

$$\begin{aligned}\|u(t)\|_{L^2}^2+\alpha\|\nabla u(t)\|_{L^2}^2+\frac{\nu}{2}\int_0^t \exp\Big(\frac{\nu}{2(\lambda_1^{-1}+\alpha)}(s-t)\Big)\|\nabla u(s)\|_{L^2}^2 ds \\ \leq \exp\Big(-\frac{\nu}{2(\lambda_1^{-1}+\alpha)}t\Big)\big[\|u_0\|_{L^2}^2+\alpha\|\nabla u_0\|_{L^2}^2\big] \\ +\frac{2(1+\lambda_1\alpha)}{\lambda_1^2\nu^2}\|f\|_{L^\infty(\mathbb{R}^+,L^2)}^2 .\end{aligned} \tag{20.62}$$

From the estimates (20.60) and (20.62), we at once deduce that, for $0 \leq t < T^*(u_0)$, for $1 < p \leq 2$,

$$\begin{aligned}\|\mathrm{rot}\,(u-\alpha\Delta u)(t)\|_{L^p} \leq \exp\Big(-\frac{\nu}{\alpha}t\Big)\|\mathrm{rot}\,(u_0-\alpha\Delta u_0)\|_{L^p} \\ +\alpha^{-1/2}\exp\Big(-\frac{\nu\lambda_1}{4(1+\lambda_1\alpha)}t\Big)\big[\|u_0\|_{L^2}+\sqrt{\alpha}\|\nabla u_0\|_{L^2}\big] \\ +\frac{\alpha}{\nu}\|\mathrm{rot}\,f\|_{L^\infty(\mathbb{R}^+,L^p)}+\frac{\sqrt{2}(1+\lambda_1\alpha)^{1/2}}{\lambda_1\nu\sqrt{\alpha}}\|f\|_{L^\infty(\mathbb{R}^+,L^2)} ,\end{aligned} \tag{20.63}$$

This inequality implies the global existence of u in the case $1 < p \leq 2$.

In the case where $p = 2$, we obtain a better estimate than (20.63). Indeed, replacing ω by $\mathrm{rot}\,(u-\alpha\Delta u)$ in the equality (20.35) and taking the inner product of this equation with $\mathrm{rot}\,(u-\alpha\Delta u)$, we readily obtain, for $t \geq 0$,

$$\begin{aligned}\|\mathrm{rot}\,(u-\alpha\Delta u)(t)\|_{L^2}^2 \leq \exp\Big(-\frac{\nu\lambda_1}{2(1+2\alpha\lambda_1)}t\Big)\|\mathrm{rot}\,(u_0-\alpha\Delta u_0)\|_{L^2}^2 \\ +\frac{2(1+2\alpha\lambda_1)^2}{\lambda_1^2\nu^2}\|\mathrm{rot}\,f\|_{L^\infty(\mathbb{R}^+,L^2)}^2 .\end{aligned} \tag{20.64}$$

For more details, we refer the reader to [55, Sect. 2.2]. In the case where $2 < p < +\infty$, we remark that the continuous Sobolev embedding $H^1(\mathbb{T}^2) \subset L^p(\mathbb{T}^2)$ holds for any $1 < p < +\infty$. Thus, we directly deduce from the inequalities (20.60) and (20.64) that, for $2 < p < +\infty$, for $0 \leq t < T^*(u_0)$,

$$
\begin{aligned}
&\|\mathrm{rot}\,(u-\alpha\Delta u)(t)\|_{L^p}\\
&\qquad\le \exp\,(-\frac{\nu}{\alpha}t)\|\mathrm{rot}\,(u_0-\alpha\Delta u_0)\|_{L^p}+\frac{\alpha}{\nu}\|\mathrm{rot}\,f\|_{L^\infty(\mathbb{R}^+,L^p)}\\
&\qquad\quad + C_S(p)\min(\alpha^{-1},\alpha^{-\frac{1}{2}})\Big[\exp(-\frac{\nu\lambda_1}{4(1+2\alpha\lambda_1)}t)\|\mathrm{rot}\,(u_0-\alpha\Delta u_0)\|_{L^2}\\
&\qquad\quad + \frac{\sqrt{2}(1+2\alpha\lambda_1)}{\lambda_1\nu}\|f\|_{L^\infty(\mathbb{R}^+,L^2)}\Big]\,,
\end{aligned}
\tag{20.65}
$$

where $C_S(p)$ is a positive constant depending on the above-mentioned Sobolev embedding. This inequality implies the global existence of u in the case where $2<p<+\infty$.

Notice that the existence of solutions on the time interval $(-\infty,0]$ also holds if the forcing term belongs to $L^\infty((-\infty,0),W^{k+1,p})$, $k\ge 0$, $1<p<+\infty$. But the solution $u(t)$ may blow-up at $-\infty$.

Assume now that the forcing term $f\in W^{1,p}_{per}$ does not depend on the time. Then, we introduce the map $S_\alpha(t):u_0\in V^{3,p}\mapsto u(t)\in V^{3,p}$, where $u(t)$ is the solution of the system (20.2). The properties that we obtained in Sects. 20.2.3 and 20.2.4 imply that $S_\alpha(t)$ is a dynamical system (and also a non-linear continuous group). Moreover, due to the estimates (20.63) to (20.65), $S_\alpha(t)$ admits a bounded absorbing set $\mathscr{B}_\alpha$. We can choose for $\mathscr{B}_\alpha$ the ball $B_{V^{3,p}}(0,C\alpha^{-1}R_\alpha(p))$ of center 0 and radius $C\alpha^{-1}R_\alpha(p)$ in $V^{3,p}$, where C is a positive constant and where

$$
\begin{aligned}
&R_\alpha(p)=\frac{\alpha}{\nu}\|\mathrm{rot}\,f\|_{L^p}+\frac{\sqrt{2}(1+\lambda_1\alpha)^{1/2}}{\lambda_1\nu\sqrt{\alpha}}\|f\|_{L^2}\ \text{if}\ 1<p<2\\
&R_\alpha(p)=\frac{\sqrt{2}(1+2\alpha\lambda_1)}{\lambda_1\nu}\|\mathrm{rot}\,f\|_{L^2}\ \text{if}\ p=2\\
&R_\alpha(p)=\frac{\alpha}{\nu}\|\mathrm{rot}\,f\|_{L^p}+C_S(p)\min(\alpha^{-1},\alpha^{-\frac{1}{2}})\frac{\sqrt{2}(1+2\alpha\lambda_1)}{\lambda_1\nu}\|f\|_{L^2}\ \text{if}\ 2<p<+\infty\,.
\end{aligned}
\tag{20.66}
$$

Remark 20.4. The norm $\|\mathrm{rot}\,(u-\alpha\Delta u)\|_{L^p}$ is appropriate for estimating the $V^{3,p}$-norm of the solution u of the second grade fluid equations (20.2). The estimates (20.63), (20.64) and (20.65) are good if $\alpha>0$ is fixed or bounded away from zero. However, when α goes to 0, these estimates can be improved as we did it in [55, Sect. 2, Estimates (2.27) and (2.28)]. In order to obtain better estimates in the case where α is small, one proceeds in the following way. Instead of introducing the variable $\omega=\mathrm{rot}\,(u-\alpha\Delta u)$, one introduces the variable $\omega^*=-\mathrm{rot}\,\Delta u$ and one performs a priori estimates for ω^* by considering the transport equation:

$$
\partial_t\omega^*+u\cdot\nabla\omega^*+\frac{\nu}{\alpha}\omega^*+\frac{1}{\alpha}\partial_t\mathrm{rot}\,u=\frac{1}{\alpha}u\cdot\nabla\mathrm{rot}\,u+\frac{1}{\alpha}\mathrm{rot}\,f\,,
\tag{20.67}
$$

and using the above estimates.

In [55, Sect. 4], using these better upper bounds, we have proved the convergence of the solutions of (20.2) to those of the Navier-Stokes equations on finite time intervals when α goes to 0. We have also obtained convergence results for the global attractors. For the comparison of periodic orbits or other invariant sets of (20.2) with those of the Navier-Stokes equations, when α is small, we refer to [35, 49]. For another convergence result of solutions of (20.2) to those of the Navier-Stokes equations, we refer to [39].

Remark 20.5. The above global existence of solutions of (20.2) is still true, when the periodic boundary conditions are replaced by homogeneous Dirichlet ones, provided the domain Ω is smooth enough (of class C^2) and simply connected.

20.3 Dynamics of the Second Grade Fluids in the 2D Torus

In the whole section, we assume that the forcing term $f \in W^{1,p}_{per}$ does not depend on the time variable t. By (20.32), the solution $u(t)$ of (20.2) writes, for any $t \in \mathbb{R}$,

$$(u - \alpha\Delta u)(t,x) = \omega(t,x) = e^{-\frac{\nu}{\alpha}t}\omega_0(\varphi(0;t,x)) + \int_0^t e^{-\frac{\nu}{\alpha}(t-s)}\Big(\operatorname{rot} f(\varphi(s;t,x)) + \frac{\nu}{\alpha}\operatorname{rot} u(s,\varphi(s;t,x))\Big)\,ds\,. \tag{20.68}$$

20.3.1 Existence of a Compact Global Attractor

In the previous section, we have seen that $S_\alpha(t)$ admits a bounded absorbing set $\mathscr{B}_\alpha$ and that the trajectories of bounded sets are bounded. Thus, by [32, Theorem 3.4.6] or [58, Theorem 2.26], in order to establish the existence of a compact global attractor in $V^{3,p}$, it suffices to show that $S_\alpha(t)$ is asymptotically compact (or asymptotically smooth). Due to [32, Lemma 3.2.6] or [58, Theorem 2.31]), it is enough to prove that $S_\alpha(t)$ can be written as a sum

$$S_\alpha(t) = \Sigma_\alpha(t) + K_\alpha(t)\,, \tag{20.69}$$

where $\Sigma_\alpha(t)$ is an asymptotically uniformly contracting map on the bounded sets of $V^{3,p}$ and $K_\alpha(t)$ is a compact map from $V^{3,p}$ into itself. Actually, due to the equality (20.68), it suffices to show that, for any $u_0 \in V^{3,p}$,

$$S_\alpha(t)u_0 - \alpha\Delta S_\alpha(t)u_0 \equiv \omega(t;u_0) = \Sigma^*_\alpha(t)u_0 + K^*_\alpha(t)u_0\,, \tag{20.70}$$

where $\Sigma^*_\alpha(t)$ is an asymptotically uniformly contracting map on the bounded sets of $V^{3,p}$ into $V^{0,p}$ and $K^*_\alpha(t)$ is a compact map from $V^{3,p}$ into $V^{0,p}$.

The proof of the property (20.70) is simple. According to the equality (20.68), we set, for any $u_0 \in V^{3,p}$,

$$\begin{aligned}\Sigma^*_\alpha(t)u_0 &\equiv e^{-\frac{\nu}{\alpha}t}\omega_0(\varphi_{u_0}(0;t,x);u_0)\\ K^*_\alpha(t)u_0 &\equiv K_{\alpha,1}(t)u_0 + K_{\alpha,2}(t)u_0\\ &\equiv \int_0^t e^{-\frac{\nu}{\alpha}(t-s)}\mathrm{rot}\, f(\varphi_{u_0}(s;t,x))ds + \int_0^t e^{-\frac{\nu}{\alpha}(t-s)}\frac{\nu}{\alpha}\mathrm{rot}\, u(s,\varphi_{u_0}(s;t,x);u_0)ds\,.\end{aligned} \tag{20.71}$$

Since $\|\omega_0(\varphi_{u_0}(0;t,x);u_0)\|_{L^p} = \|\omega_0(x;u_0)\|_{L^p}$, it follows that, for any bounded set $B_0 \in V^{3,p}$, for any $u_0 \in B_0$, for any $t \geq 0$,

$$\|\Sigma^*_\alpha(t)u_0\|_{L^p} \leq C_1(\|B_0\|_{V^{3,p}})e^{-\frac{\nu}{\alpha}t}\,, \tag{20.72}$$

where $C_1(\|B_0\|_{V^{3,p}})$ depends only on the norm of B_0 in $V^{3,p}$.

We next show that $K_{\alpha,1}(t)$ is a compact map from $V^{3,p}$ into L^p_{per}. Let B_0 be a bounded subset of $V^{3,p}$. The set $[0,1]\times\overline{B}_0$ is a compact subset of $\mathbb{R}^+\times V^{2,p}$. Since the map $(s,u_0)\in[0,t]\times\overline{B}_0 \mapsto \mathrm{rot}\, f(\varphi_{u_0}(s;t,\cdot)) \in L^p$ is a continuous mapping, the image is a compact subset of L^p_{per}. By Mazur's theorem, it follows that $\int_0^t e^{-\frac{\nu}{\alpha}(t-s)}\mathrm{rot}\, f(\varphi_{u_0}(s;t,x))ds$ belongs to a compact set of L^p_{per}. Thus $K_{\alpha,1}(t)$ is a compact map.

Finally, we prove that $K_{\alpha,2}(t)$ is a compact mapping from $V^{3,p}$ into L^p_{per} by showing that $K_{\alpha,2}(t)$ maps every bounded set $B_0 \subset V^{3,p}$ into a compact set of $W^{1,p}_{per}$ and thus into a relatively compact set of L^p_{per}. Since, by Lemma 20.1, the following estimate holds

$$\|\mathrm{rot}\, u(s,\varphi_{u_0}(s;t,\cdot);u_0)\|_{W^{1,p}} \leq \|\mathrm{rot}\, u(s,\cdot;u_0,\|_{W^{1,p}}\exp(\int_0^t \|\nabla u(\sigma,\cdot;u_0)\|_{L^\infty}d\sigma\,, \tag{20.73}$$

and that

$$\|\mathrm{rot}\, u(\cdot,\cdot;u_0)\|_{L^\infty((0,t),W^{1,p})}\exp(\int_0^t \|\nabla u(\sigma,\cdot;u_0)\|_{L^\infty}d\sigma \leq C_2(t,\|B_0\|_{V^{3,p}})\,, \tag{20.74}$$

(where $C_2(t,\|B_0\|_{V^{3,p}})$ depends only on t and on the norm of B_0 in $V^{3,p}$), it follows that $\int_0^t e^{-\frac{\nu}{\alpha}(t-s)}\frac{\nu}{\alpha}\mathrm{rot}\, u(s,\varphi_{u_0}(s;t,x);u_0)ds$ belongs to a bounded set of $W^{1,p}_{per}$ and thus to a compact set of L^p_{per}. And then $K_{\alpha,2}(t)$ is a compact mapping from $V^{3,p}$ into L^p_{per}.

We have thus proved that $S_\alpha(t)$ is asymptotically compact in $V^{3,p}$.

Remark 20.6. We can prove in the same way that $S_\alpha(t)$ admits a compact global attractor when the periodic boundary conditions are replaced by homogeneous Dirichlet ones, provided the domain Ω is smooth enough (of class C^2) and simply connected.

20.3.2 Regularity of the Compact Global Attractor

We now consider a complete bounded orbit $u(t)$ (with $u(0) = u_0$) contained in the global attractor $\mathscr{A}_\alpha$. In particular, we know that $\omega(t) \equiv rot(u - \alpha\Delta u)(t)$ satisfies the following inequality, for any $1 < p < +\infty$,

$$\|rot(u - \alpha\Delta u)(t)\|_{V^{1,p}} \leq R_\alpha(p) \, , \quad \forall t \in \mathbb{R} \, . \tag{20.75}$$

Let $\tau < 0$. Since $\omega(t) \in V^{1,p}$ exists for any $t \in \mathbb{R}$ and, by (20.75), is uniformly bounded on $\mathbb{R}$, using the formula (20.68), we can write, for $t \geq \tau$,

$$\begin{aligned} \omega(t,x) =& e^{-\frac{\nu}{\alpha}(t-\tau)}\omega(\tau, \varphi(\tau;t,x)) \\ &+ \int_\tau^t e^{-\frac{\nu}{\alpha}(t-s)} \Big(\mathrm{rot}\, f(\varphi(s;t,x)) + \frac{\nu}{\alpha} \mathrm{rot}\, u(s, \varphi(s;t,x)) \Big) ds \, . \end{aligned} \tag{20.76}$$

Letting τ go to $-\infty$, one deduces from (20.75) and (20.76) that, for any $t \in \mathbb{R}$,

$$\omega(t,x) = \int_{-\infty}^t e^{-\frac{\nu}{\alpha}(t-s)} \Big(\mathrm{rot}\, f(\varphi(s;t,x)) + \frac{\nu}{\alpha} \mathrm{rot}\, u(s, \varphi(s;t,x)) \Big) ds \, . \tag{20.77}$$

If $f \in W^{2,p}_{per}$, the right-hand side member of (20.77) belongs to a smoother space than L^p_{per}. We now want to prove that indeed $\omega(t)$ is bounded in a smoother space $W^{\theta,p}$, where $0 < \theta \leq 1$. To this end, we introduce the integral

$$I(t,s,x) \equiv \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} g(\sigma, \varphi(\sigma;t,x)) d\sigma \, , \quad g(\sigma,x) = \mathrm{rot}\, f(x) + \frac{\nu}{\alpha} \mathrm{rot}\, u(\sigma,x) \, .$$

Since

$$\frac{\partial}{\partial x_k} I(t,s,x) = \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \Big(\sum_{i=1}^2 \partial_{x_i} g(\sigma, \varphi(\sigma;t,x)) \partial_{x_k} \varphi_i(\sigma;t,x) \Big) d\sigma \, , \tag{20.78}$$

thus, by Lemma 20.1,

$$\begin{aligned} \| \frac{\partial}{\partial x_k} I(t,s,x) \|_{L^p} \leq & \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \|\nabla g\|_{L^p} \|\nabla \varphi(\sigma;t,\cdot)\|_{L^\infty} d\sigma \\ \leq & \|\nabla g\|_{L^\infty(L^p)} \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \exp\Big(\int_\sigma^t \|\nabla u(\tau)\|_{L^\infty} d\tau \Big) d\sigma \, . \end{aligned} \tag{20.79}$$

Assume now that

$$\sup_{t \in \mathbb{R}} \|\nabla u(t)\|_{L^\infty} < \frac{\nu}{\alpha} \, , \tag{20.80}$$

and set $a_0 \equiv \frac{\nu}{\alpha} - \sup_{\tau\in\mathbb{R}} \|\nabla u(\tau)\|_{L^\infty} > 0$. From the estimate (20.79) and the hypothesis (20.80), we deduce that, for any $s \leq t$,

$$\|\frac{\partial}{\partial x_k} I(t,s,x)\|_{L^p} \leq a_0^{-1}\Big(\|\mathrm{rot}\, f\|_{W^{1,p}} + \frac{\nu}{\alpha}\sup_{\tau\in\mathbb{R}}\|\mathrm{rot}\, u(\tau)\|_{W^{1,p}}\Big)\,. \tag{20.81}$$

Since this inequality holds for any t, s, we conclude that, for any $t \in \mathbb{R}$,

$$\|\nabla(\mathrm{rot}\, u - \alpha\Delta \mathrm{rot}\, u)(t)\|_{L^p} \leq a_0^{-1}\Big(\|\mathrm{rot}\, f\|_{W^{1,p}} + \frac{\nu}{\alpha}\sup_{\tau\in\mathbb{R}}\|\mathrm{rot}\, u(\tau)\|_{W^{1,p}}\Big)\,. \tag{20.82}$$

Assume now that $\sup_{v\in\mathscr{A}_\alpha}\|\nabla v\|_{L^\infty} < \frac{\nu}{\alpha}$ and set $a_1 \equiv \frac{\nu}{\alpha} - \sup_{v\in\mathscr{A}_\alpha}\|\nabla v\|_{L^\infty} > 0$. From the estimates (20.82) and (20.64), we deduce the following upper bound for any element u_0 in the global attractor

$$\|\nabla(\mathrm{rot}\, u_0 - \alpha\Delta \mathrm{rot}\, u_0)\|_{L^p} \leq a_1^{-1}(\|\mathrm{rot}\, f\|_{W^{1,p}} + \frac{\nu}{\alpha}M_\alpha(p))\,, \tag{20.83}$$

where

$$\begin{aligned} M_\alpha(p) &= C_S\frac{\sqrt{2}(1+2\alpha\lambda_1)}{\max(\sqrt{\alpha},\alpha)\lambda_1\nu}\|\mathrm{rot}\, f\|_{L^2} \text{ if } 2 < p < +\infty \\ M_\alpha(p) &= C_S\frac{\sqrt{2}(1+2\alpha\lambda_1)}{\alpha\lambda_1\nu}\|\mathrm{rot}\, f\|_{L^2} \text{ if } p \geq 2\,, \end{aligned} \tag{20.84}$$

where $C_S > 0$ is a Sobolev embedding constant.

Assume now that, for the complete bounded orbit $u(t)$, the condition (20.80) is not true. Then we are not able to conclude that $\|\nabla(\mathrm{rot}\, u - \alpha\Delta\mathrm{rot}\, u)(t)\|_{L^p}$ is uniformly bounded for $t \in \mathbb{R}$. However, we can still show that there exists a positive number $0 < \theta < 1$, such that $\|(\mathrm{rot}\, u - \alpha\Delta\mathrm{rot}\, u)(t)\|_{W^{\theta,p}}$ is bounded by using an interpolation argument. Indeed, let $0 < \theta < 1$ such that

$$a_\theta \equiv \frac{\nu}{\alpha} - \theta\sup_{\tau\in\mathbb{R}}\|\nabla u(\tau)\|_{L^\infty} > 0\,. \tag{20.85}$$

We next define the continuous linear map $\mathscr{T} : h \in L^\infty(\mathbb{R}, L^p_{per}) \mapsto w \in L^\infty(\mathbb{R}, L^p_{per})$, which is the solution of the integral equation

$$\mathscr{T}(h)(t,x) \equiv w(t,x) = \int_{-\infty}^{t} e^{-\frac{\nu}{\alpha}(t-\sigma)}h(\sigma,\varphi(\sigma;t,x))d\sigma\,.$$

We remark that the vorticity ω, satisfying the equality (20.77), is given by $\omega = \mathscr{T}(g)$. The above computations show that $\mathscr{T}$ is also a continous map from $L^\infty(\mathbb{R}, W^{1,p}_{per})$ into itself, and thus, by interpolation, a continuous linear map from $L^\infty(\mathbb{R}, W^{\theta,p}_{per})$ into itself, for $0 \leq \theta \leq 1$. Moreover, for any $t \in \mathbb{R}$, we have,

$$\|\mathscr{T}(h)(t,\cdot)\|_{W^{\theta,p}} \leq \int_{-\infty}^{t} e^{-\frac{\nu}{\alpha}(t-\sigma)} \|h(\sigma,\varphi(\sigma;t,\cdot))\|_{W^{\theta,p}} d\sigma \ .$$

As we will see, due to the condition (20.85), $g(\sigma,\varphi(\sigma;t,\cdot))$ belongs to $L^\infty(\mathbb{R}, W^{\theta,p}_{per})$ and thus $\omega(t,\cdot)$ satisfies the above inequality.

Remarking that

$$\begin{aligned} \|g(\sigma,\varphi(\sigma;t,\cdot))\|_{L^p} \leq & \|g(\sigma,\cdot)\|_{L^p} \\ \|g(\sigma,\varphi(\sigma;t,\cdot))\|_{W^{1,p}} \leq & \|g(\sigma,\cdot)\|_{W^{1,p}} \exp\Big(\int_\sigma^t \|\nabla u(\tau)\|_{L^\infty} d\tau\Big) \end{aligned} \tag{20.86}$$

we obtain by interpolation that,

$$\begin{aligned} \|\omega(t,\cdot)\|_{W^{\theta,p}} \leq & \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \exp\theta\Big(\int_\sigma^t \|\nabla u(\tau)\|_{L^\infty} d\tau\Big) \|g(\sigma,\cdot)\|_{W^{\theta,p}} d\sigma \\ \leq & a_\theta^{-1}\Big(\|\mathrm{rot}\, f\|_{W^{\theta,p}} + \frac{\nu}{\alpha} \sup_{\tau\in\mathbb{R}} \|\mathrm{rot}\, u(\tau)\|_{W^{\theta,p}}\Big) \ . \end{aligned} \tag{20.87}$$

And we conclude that, for any $t \in \mathbb{R}$,

$$\|(\mathrm{rot}\, u - \alpha\Delta \mathrm{rot}\, u)(t)\|_{W^{\theta,p}} \leq a_\theta^{-1}\Big(\|\mathrm{rot}\, f\|_{W^{\theta,p}} + \frac{\nu}{\alpha} \sup_{\tau\in\mathbb{R}} \|\mathrm{rot}\, u(\tau)\|_{W^{\theta,p}}\Big) \ . \tag{20.88}$$

If $\sup_{v\in\mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} \geq \frac{\nu}{\alpha}$, we take $0<\theta<1$ so that

$$a_{1,\theta} \equiv \frac{\nu}{\alpha} - \theta \sup_{v\in\mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} > 0 \ .$$

We obtain the following upper bound for any u_0 in the global attractor

$$\|\mathrm{rot}\, u_0 - \alpha\Delta \mathrm{rot}\, u_0\|_{W^{\theta,p}} \leq a_{1,\theta}^{-1}(\|\mathrm{rot}\, f\|_{W^{\theta,p}} + \frac{\nu}{\alpha} M_\alpha(p)) \ , \tag{20.89}$$

Remark 20.7. One may wonder if the compact global attractors depend on p. Let $1 < p_1 < p_2 < +\infty$ and assume that the forcing term f belongs to $W^{1+\theta,p_2}$, $0<\theta<1$. We denote $\mathscr{A}_\alpha(p_1)$ and $\mathscr{A}_\alpha(p_2)$ the corresponding global attractors. It is clear that $\mathscr{A}_\alpha(p_2) \subset \mathscr{A}_\alpha(p_1)$. Taking into account the above regularity argument, we may show by using Sobolev embeddings and a bootstrap argument that $\mathscr{A}_\alpha(p_1) \subset \mathscr{A}_\alpha(p_2)$ and thus $\mathscr{A}_\alpha(p_1) = \mathscr{A}_\alpha(p_2)$.

Next we consider higher order derivatives of $\omega(t)$. Differentiating $\frac{\partial}{\partial x_k} I(t,s,x)$ with respect to x_l, we obtain

$$\frac{\partial^2}{\partial x_k \partial x_l} I(t,s,x) = \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \Big[\sum_{i,j=1}^{2} \partial^2_{x_i x_j} g(\sigma, \varphi(\sigma;t,x)) \partial_{x_k} \varphi_i(\sigma;t,x) \partial_{x_l} \varphi_j(\sigma;t,x) + \sum_{i=1}^{2} \partial_{x_i} g(\sigma, \varphi(\sigma;t,x)) \partial^2_{x_k x_l} \varphi_i(\sigma;t,x) \Big] d\sigma \,, \tag{20.90}$$

from which we deduce that, for any $s \leq t$,

$$\|D^2_x I(t,s,\cdot)\|_{L^p} \leq \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \big[\|\nabla g\|_{L^p} \|D^2_x \varphi(\sigma;t,\cdot)\|_{L^\infty} + \|D^2_x g\|_{L^p} \|\nabla \varphi(\sigma;t,\cdot)\|^2_{L^\infty} \big] d\sigma \,. \tag{20.91}$$

Arguing as in Lemma 20.1 and using the inequality (20.14) of Lemma 20.1, we get the following estimate, for any $\sigma \leq t$,

$$\|D^2_x \varphi(\sigma;t,\cdot)\|_{L^\infty} \leq \|D^2_x u(\sigma)\|_{L^\infty} \exp\Big(3 \int_\sigma^t \|\nabla u(\tau)\|_{L^\infty} d\tau\Big) \,. \tag{20.92}$$

The estimates (20.91) and (20.92) imply, for any $s \leq t$,

$$\begin{aligned} \|D^2_x I(t,s,\cdot)\|_{L^p} \leq & \|\nabla g\|_{L^\infty(L^p)} \|u\|_{L^\infty(W^{2,\infty})} \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \exp\Big(3 \int_\sigma^t \|\nabla u(\tau)\|_{L^\infty} d\tau\Big) d\sigma \\ & + \|D^2_x g\|_{L^\infty(L^p)} \int_s^t e^{-\frac{\nu}{\alpha}(t-\sigma)} \exp\Big(2 \int_\sigma^t \|\nabla u(\tau)\|_{L^\infty} d\tau\Big) d\sigma \,. \end{aligned} \tag{20.93}$$

Assume now that

$$\sup_{t \in \mathbb{R}} 3 \|\nabla u(t)\|_{L^\infty} < \frac{\nu}{\alpha} \,, \tag{20.94}$$

and set $a_2 \equiv \frac{\nu}{\alpha} - 3 \sup_{\tau \in \mathbb{R}} \|\nabla u(\tau)\|_{L^\infty} > 0$. Then, it follows from (20.93) and (20.94) that, for any $t \in \mathbb{R}$,

$$\begin{aligned} \|(\mathrm{rot}\, u - \alpha \Delta \mathrm{rot}\, u)(t)\|_{W^{2,p}} \leq & a_2^{-1} \Big[(\|\mathrm{rot}\, f\|_{W^{1,p}} + \frac{\nu}{\alpha} \|\mathrm{rot}\, u(\cdot)\|_{L^\infty(W^{1,p})}) \|u(\cdot)\|_{L^\infty(W^{2,\infty})} \\ & + (\|\mathrm{rot}\, f\|_{W^{2,p}} + \frac{\nu}{\alpha} \|\mathrm{rot}\, u(\cdot)\|_{L^\infty(W^{2,p})}) \Big] \\ \equiv & a_2^{-1} M_{2,\alpha}(p) \,. \end{aligned} \tag{20.95}$$

If $a_2 \equiv \frac{\nu}{\alpha} - 3 \sup_{v \in \mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} > 0$, then the estimate (20.95) holds for any element u of $\mathscr{A}_\alpha$. By a recursion argument, we finally obtain the third assertion of Theorem 20.3.

Remark 20.8. The above regularity results of $\mathscr{A}_\alpha$ still hold, when the periodic boundary conditions are replaced by homogeneous Dirichlet ones, provided the domain Ω is smooth enough (of class C^2) and simply connected.

Remark 20.9. In the above regularity proofs, in order to get the V^{3+m} regularity, we need to assume that $\frac{\nu}{\alpha} - m \sup_{v\in\mathscr{A}_\alpha} \|\nabla v\|_{L^\infty} > 0$. This method does not allow to show that the attractor is bounded in C^∞ or in the set of analytic functions, even if f is analytic. So these regularity properties remain an open question. Note that if f is integrable in time and is in a Gevrey class in the spatial variable and if the initial data are in a Gevrey class in the spatial variable, then the solutions of (20.2) also have Gevrey regularity (see [52] and [54]).

20.3.3 Finite-Dimensional Properties

We can also wonder if the dynamics of (20.2) has finite-dimensional properties. Using the methods of [61], we can certainly prove that the Hausdorff dimension of $\mathscr{A}_\alpha$ is finite. We leave it to the reader to check it.

We next want to recall a "finite-dimensional" property, which is well adapted to the Hilbert space setting, that is, to the case $p = 2$. Let P denote the classical orthogonal projection of $(L^2_{per}(\mathbb{T}^2))^2$ onto the subspace $H \equiv V^{0,2}$ of L^2-divergence-free vector fields. We also introduce the orthogonal projection P_n in H onto the space spanned by the eigenfunctions corresponding to the first n eigenvalues of the Stokes operator $A = -P\Delta$. Finally, we introduce the projection $Q_n = I - P_n$.

In [55], we have shown that there exists an integer N, such that, on the compact global attractor, the dynamics of (20.2) reduces to the dynamics of a system of N ordinary differential equations defined on $P_N V^{3,2}$ (see [55, Theorem 1.2]).

In [55], like in [34, Theorem 2.7], we deduced, from [55, Theorem 1.2], the so-called "finite number of determining modes property" for the system (20.2), when α is small enough. The property of "finite number of determining modes" was introduced and proved for the two-dimensional Navier–Stokes equations by Foias and Prodi in 1967 [19]. This property means that the asymptotic behaviour in time of the second grade fluid system depends only on a finite number of parameters (called the determining modes).

Theorem 20.5. *Let f be given in $W^{1+d,2}_{per}$, $d > 0$.*

We assume that $\nu - 2\alpha(\sup_{z\in\mathscr{A}_\alpha} \|\nabla z\|_{L^\infty}) > 0$. Then System (20.2) has the property of finite number of determining modes, that is, there exists a positive integer N_0 such that, for any u_0, u_1 in $V^{3,2}$, the property

$$\|P_{N_0} S_\alpha(t)u_0 - P_{N_0} S_\alpha(t)u_1\|_{V^3} \longrightarrow_{t\to+\infty} 0$$

implies that

$$\|S_\alpha(t)u_0 - S_\alpha(t)u_1\|_{V^3} \longrightarrow_{t\to+\infty} 0\,.$$

One also could directly prove Theorem 20.5, by performing appropriate a priori estimates. But, showing Theorem 20.5 as a consequence of [55, Theorem 1.2] and of the proof of [34, Theorem 2.7] is more elegant.

References

1. Bardos, C., Di Plinio, F., Temam, R.: The Euler equations in planar nonsmooth convex domains. J. Math. Anal. Appl. **407**, 69–89 (2013)
2. Beirão da Veiga, H.: Boundary-value problems for a class of first order partial differential equations in Sobolev spaces and applications to the Euler flow. Rend. Sem. Mat. Univ. Padova **79**, 247–273 (1988)
3. Bernard, J.-M.: Solutions globales variationnelles et classiques des fluides de grade deux. C. R. Acad. Sci. Paris Sér. I **327**, 953–958 (1998)
4. Bernard, J.-M.: Stationary problem of second-grade fluids in three dimensions: existence, uniqueness and regularity. Math. Meth. Appl. Sci. **22**, 655–687 (1999)
5. Bernard, J.M.: Weak and classical solutions of equations of motion for third grade fluids. Math. Model. Numer. Anal. **33**, 1091–1120 (1999)
6. Bernard, J.M.: Solutions $W^{2,p}, p > 3$, for second grade fluid equations with a boundary of class $C^{1,1}$. Commun. Appl. Nonlinear Anal. **9**, 1–29 (2002)
7. Bresch, D., Lemoine, J.: Stationary solutions for second grade fluids equations. Math. Models Meth. Appl. Sci. **8**, 737–748 (1998)
8. Bresch, D., Lemoine, J.: On the existence of solutions for non-stationary second grade fluids. In: Navier–Stokes Equations and Related Nonlinear Problems (Palanga, 1997), pp. 15–30. VSP, Utrecht (1998)
9. Bresch, D., Lemoine, J.: On the existence of solutions for non-stationary third-grade fluids. Int. J. Nonlinear Mech. **34**, 485–498 (1999)
10. Busuioc, V.: On second grade fluids with vanishing viscosity. C. R. Acad. Sci. Paris Ser. I Math. **328**, 1241–1246 (1999)
11. Chemin, J.-Y.: Fluides parfaits incompressibles. Astérisque 230 (1995)
12. Cioranescu, D., Girault, V.: Weak and classical solutions of a family of second grade fluids. Int. J. Nonlinear Mech. **32**, 317–335 (1997)
13. Cioranescu, D., Ouazar, E.H.: Existence et unicité pour les fluides de second grade. Note CRAS Sér. I **298**, 285–287 (1984)
14. Cioranescu, D., Ouazar, E.H.: Existence and uniqueness for fluids of second grade. In: Collège de France seminar, vol. VI (Paris, 1982/1983), pp. 178–197. Pitman, Boston, MA (1984)
15. Coulaud, O.: Asymptotic profiles for equations of second grade fluids equations on R^3 (2013), Manuscript
16. Dunn, J.E., Fosdick, R.L.: Thermodynamics, stability and boundedness of fluids of complexity 2 and fluids of second grade. Arch. Ration. Mech. Anal. **56**, 191–252 (1974)
17. Dunn, J.E., Rajagopal, K.R.: Fluids of differential type: critical review and thermodynamic analysis. Int. J. Eng. Sci. **33**, 689 (1995)
18. Fosdick, R.L., Rajagopal, K.R.: Anomalous features in the model of "second order fluids". Arch. Ration. Mech. Anal. **70**, 145–152 (1979)
19. Foias, C., Prodi, G.: Sur le comportement global des solutions non stationnaires des équations de Navier–Stokes en dimension deux. Rend. Sem. Mat. Univ. Padova **39**, 1–34 (1967)
20. Foias C., Holm D., Titi E.S.: The Navier–Stokes-alpha model of fluid turbulence. Phys. D (Special Issue in Honor of V. E. Zakharov on the Occasion of his 60th birthday), **152**, 505–519 (2001)
21. Galdi, G.P., Coscia, V.: Existence, uniqueness and stability of regular steady motions of a second grade fluid. Int. J. Nonlinear Mech. **29**, 493–506 (1994)

22. Galdi, G.P., Rajagopal, K.R.: Slow motion of a body in a fluid of second grade. Int. J. Eng. Sci. **35**, 33–54 (1997)
23. Galdi, G.P., Sequeira, A.: Further existence results for classical solutions of the equations of second grade fluids. Arch. Ration. Mech. Anal. **128**, 297–312 (1994)
24. Galdi, G.P., Padula, M., Rajagopal, K.R.: On the conditional stability of the rest state of a fluid of second grade in unbounded domains. Arch. Ration. Mech. Anal. **109**, 173–182 (1990)
25. Galdi, G.P., Grobbelaarvandalsen, M., Sauer, N.: Existence and uniqueness of classical-solutions of the equations of motion for 2nd-grade fluids. Arch. Ration. Mech. Anal. **124**, 221–237 (1993)
26. Galdi, G.P., Grobbelaarvandalsen, M., Sauer, N.: Existence and uniqueness of solutions of the equations of motion for a fluid of second grade with non-homogeneous boundary conditions. Int. J. Nonlinear Mech. **30**, 701–709 (1995)
27. Girault, V., Scott, L.R.: Analysis of a two-dimensional grade-two fluid model with a tangential boundary condition. J. Math. Pures Appl. **78**, 981–1011 (1999)
28. Girault, V., Saadouni, M.: On a time-dependent grade-two fluid model in two dimensions. Comput. Math. Appl. **53**, 347–360 (2007)
29. Gupta, A.S., Rajagopal, K.R.: An exact solution for the flow of a non-Newtonian fluid past an infinite porous plate. Mechanica **19**, 158–160 (1984)
30. Hale, J.K.: Smoothing properties of neutral equations. An. Acad. Brasil Ci. **45**, 49–50 (1973)
31. Hale, J.K.: Asymptotic behaviour and dynamics in infinite dimensions. In: Research Notes in Mathematics, vol. 132, pp 1–41. Pitman, Boston (1985)
32. Hale, J.K.: Asymptotic behavior of dissipative systems. Mathematical Surveys and Monographs, vol. 25. American Mathematical Society, Providence, RI (1988)
33. Hale, J.K., Joly, R., Raugel, G.: Infinite Dimensional Dissipative Systems, Book, Manuscript (2013)
34. Hale, J.K., Raugel, G.: Regularity, determining modes and Galerkin method. J. de Mathématiques Pures et Appl. **82**, 1075–1136 (2003)
35. Hale, J.K., Raugel, G.: A modified Poincaré method for the persistence of periodic orbits and applications. J. Dynam. Differ. Equat. **22**, 3–68 (2010)
36. Hale, J.K., Raugel, G.: Persistence of periodic orbits for perturbed dissipative dynamical systems. In: Fields Institute Communications, vol. 64, pp. 1–55. Springer, New York (2013)
37. Hale, J.K., Scheurle, J.: Smoothness of bounded solutions of nonlinear evolution equations. J. Differ. Equat. **56**, 142–163 (1985)
38. Henry, D.: Geometric theory of semilinear parabolic equations. Lecture Notes in Math. vol. 840. Springer, New York (1981)
39. Iftimie, D.: Remarques sur la limite $\alpha \to 0$ pour les fluides de grade 2. C. R. Acad. Sci. Paris Sér. I Math. **334**, 83–86 (2002)
40. Jaffal-Mourtada, B.: Dynamique des fluides de grade deux, PhD thesis, Univ. Paris-Sud, Orsay (2010)
41. Jaffal-Mourtada, B.: Long-time asymptotics of the second grade fluid equations. Dynam. Partial Differ. Equat. **8**, 185–223 (2011)
42. Kato, T.: Linear evolution equations of "hyperbolic" type, I. J. Fac. Sci. Univ. Tokyo **17**, 241–258 (1970)
43. Kato, T.: Linear evolution equations of "hyperbolic" type, II. J. Math. Soc. Jpn. **25**, 648–666 (1973)
44. Ladyzenskaya, O.A., Solonnikov, V.: Unique solvability of an initial and boundary value problem for viscous incompressible nonhomogeneous fluids. J. Sov. Math. **9**, 697–749 (1978)
45. Le Roux, C.: Existence and uniqueness of the flow of second-grade fluids with slip boundary conditions. Arch. Ration. Mech. Anal. **148**, 309–356 (1999)
46. Lions, J.L.: Quelques méthodes de résolution des Problèmes Non-linéaires. Dunod, Paris (1969)
47. Marsden, J.E., Ratiu, T., Shkoller, S.: A nonlinear analysis of the averaged Euler equations and a new diffeomorphism group. Geom. Funct. Anal. **10**, 582–599 (2000)

48. Massoudi, M., Vaidya, A.: On some generalizations of the second grade fluid model. Nonlinear Anal. Real World Appl. **9**, 1169–1183 (2008)
49. Mischaikow, K., Raugel, G.: Non Regular Perturbations of Dissipative Partial Differential Equations and Stability of the Conley Index, Manuscript (2013)
50. Moise, I., Rosa, R., Wang, X.: Attractors for non-compact semigroups via energy equations. Nonlinearity **11**, 1369–1393 (1998)
51. Nečasová, S., Penel, P.: Incompressible non-Newtonian fluids: Time asymptotic behaviour of weak solutions. Math. Meth. Appl. Sci. **29**, 1615–1630 (2006)
52. Ngo, V.S.: Effet dispersif pour les fluides anisotropes avec viscosité évanescente en rotation rapide, PhD thesis, Univ. Paris-Sud, Orsay (2009)
53. Nussbaum, R.: Periodic solutions of analytic functional differential equations are analytic. Michigan Math. J. **20**, 249–255 (1973)
54. Paicu, M., Vicol, V.: Analyticity and Gevrey-class regularity for the second-grade fluid equations. J. Math. Fluid Mech. **13**, 533–555 (2011)
55. Paicu, M., Raugel, G., Rekalo, A.: Regularity of the global attractor and finite-dimensional behavior for the second grade fluid equations. J. Differ. Equat. **252**, 3695–3751 (2012)
56. Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations, Appl. Math. Sci., vol. 44. Springer, New York (1983)
57. Rajagopal, K.R.: On the decay of vortices in a second grade fluid. Meccanica **15**, 185–186 (1980)
58. Raugel, G.: Global attractors in partial differential equations, In: Fiedler, B. (ed.) Handbook of Dynamical Systems, vol. 2, pp 885–982. North-Holland, Amsterdam (2002)
59. Rivlin, R.S., Ericksen, J.L.: Stress-deformation relations for isotropic materials. J. Ration. Mech. Anal. **4**, 323–425 (1955)
60. Shkoller, S.: Smooth global Lagrangian flow for the 2D Euler and second-grade fluid equations. Appl. Math. Lett. **14**, 539–543 (2001)
61. Temam, R.: Infinite Dimensional Dynamical Systems in Mechanics and Physics. Springer, New York (1988), Second edition (1997)

Chapter 21
Dissipative Quantum Mechanics Using GENERIC

Alexander Mielke

Dedicated to Jürgen Scheurle, who exited my love for geometric mechanics, on the occasion of his 60th birthday

Abstract Pure quantum mechanics can be formulated as a Hamiltonian system in terms of the Liouville equation for the density matrix. Dissipative effects are modeled via coupling to a macroscopic system, where the coupling operators act via commutators. Following Öttinger (Phys. Rev. A 82:052119(11), 2010) we use the GENERIC framework (General Equations for Non-Equilibrium Reversible Irreversible Coupling) to construct thermodynamically consistent evolution equations as a sum of a Hamiltonian and a gradient-flow contribution, which satisfy a particular non-interaction condition:

$$\dot{q} = \mathbb{J}(q)\mathrm{D}\mathscr{E}(q) + \mathbb{K}(q)\mathrm{D}\mathscr{S}(q).$$

One of our models couples a quantum system to a finite number of heat baths each of which is described by a time-dependent temperature. The dissipation mechanism is modeled via the canonical correlation operator, which is the inverse of the Kubo–Mori metric for density matrices and which is strongly linked to the von Neumann entropy for quantum systems. Thus, one recovers the dissipative double-bracket operators of the Lindblad equations but encounters a correction term for the consistent coupling to the dissipative dynamics. For the finite-dimensional and isothermal case we provide a general existence result and discuss sufficient conditions that guarantee that all solutions converge to the unique thermal equilibrium state. Finally, we compare our gradient flow structure for quantum

A. Mielke (✉)
Weierstraß-Institut für Angewandte Analysis und Stochastik, Mohrenstraße 39, 10117 Berlin, Germany
e-mail: alexander.mielke@wias-berlin.de

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_21,

systems with the Wasserstein gradient flow for the Fokker–Planck equation and the entropy gradient flow for reversible Markov chains.

21.1 Introduction

A fundamental problem in nanoscience is a consistent coupling of quantum mechanics with effects on larger scales. In particular, one is interested in combining quantum and continuum mechanical models with dissipative effects in such a way that the fundamental axioms of thermodynamics and quantum mechanics are still satisfied. We refer to [33, 35] for general physical modeling of quantum dissipative systems. We follow [27, 28] in modeling the coupling between classical dissipative systems and reversible quantum systems formulated in terms of the Liouville equation for the density matrix. The basis is the theory of GENERIC systems, which stands for the acronym General Equations for Non-Equilibrium Reversible Irreversible Coupling. It provides systems that are thermodynamically correct in the sense that the total energy is preserved while the total entropy is nondecreasing. Moreover, the evolution has an additive split into the reversible dynamics driven by a co-symplectic structure (Poisson bracket) acting on the energy and the irreversible dynamics given as a gradient flow with the total entropy as the driving functional.

Evolution of the quantum mechanical system is given in terms of the density matrix $\rho = \rho^* \geq 0$ such that it can be coupled to more macroscopic dissipative components z. Thus, the total state $q = (\rho, z)$ lies in the state space $\mathscr{Q} = \mathfrak{R} \times Z$ defined below, and the evolution is given by the GENERIC system $(\mathscr{Q}, \mathscr{E}, \mathscr{S}, \mathbb{J}, \mathbb{K})$. Here $\mathscr{E}$ and $\mathscr{S}$ are the conserved energy functional and the entropy functional. The mapping $\mathbb{J}$ defines a Poisson structure, i.e. $\mathbb{J}(q) = -\mathbb{J}(q)^*$ and the Jacobi identity holds, while $\mathbb{K}$ defines an Onsager structure, i.e. $\mathbb{K}(q) = \mathbb{K}(q)^* \geq 0$. The evolution is given via

$$\dot{q} = \mathbb{J}(q)\mathrm{D}\mathscr{E}(q) + \mathbb{K}(q)\mathrm{D}\mathscr{S}(q), \tag{21.1}$$

where the crucial structural condition is the mutual *non-interaction condition*

$$\mathbb{J}(q)\mathrm{D}\mathscr{S}(q) \equiv 0 \quad \text{and} \quad \mathbb{K}(q)\mathrm{D}\mathscr{E}(q) \equiv 0. \tag{21.2}$$

We discuss the main properties of GENERIC systems in Sect. 21.2. For the history and the motivation of the GENERIC framework, we refer to [7, 8, 24, 26]. For applications in continuum mechanics, see [19].

Here we follow the approach pioneered in [27, 28] and analyze the system proposed there mathematically. Our aim is to provide existence results as well as conditions guaranteeing that the solutions converge into thermal equilibrium. We will restrict to the case that the underlying complex Hilbert space $\boldsymbol{H}$ is finite dimensional, i.e. $\dim \boldsymbol{H} < \infty$, and thus can be identified by $\mathbb{C}^{\dim \boldsymbol{H}}$. However, we hope that in future works we can treat the infinite-dimensional case

as well. To be more precise we consider GENERIC systems where the coupling occurs only through the Onsager operator $\mathbb{K}$. With $q = (\rho, z)$ we assume that $\mathscr{E}$, $\mathscr{S}$, and $\mathbb{J}$ have the form

$$\mathscr{E}(q) = \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle + E(z),\ \ \mathscr{S}(q) = -k_{\mathrm{B}} \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle + S(z),\ \ \mathbb{J}(q) = \begin{pmatrix} \mathrm{j}[\rho, \square]\, 0 \\ 0 \quad 0 \end{pmatrix},$$

where $\mathrm{j} = \mathrm{i}/\hbar$ and "$\square$" indicates the slot where the argument should be inserted. By $\langle\!\langle\, A \,\|\, B \,\rangle\!\rangle = \mathrm{tr}(AB^*)$ we denote the operator scalar product on $\mathrm{Lin}(\boldsymbol{H})$. In particular, the quantum system is described by the Liouville equation $\dot{\rho} = \mathrm{j}[\rho, H]$ for the density matrix $\rho \in \mathrm{Lin}(\boldsymbol{H})$. Here, $-k_{\mathrm{B}} \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle$ is the quantum mechanical von Neumann entropy, while $S(z)$ is assumed to be a concave macroscopic entropy.

The coupling between the components ρ and z occurs through $\mathbb{K}$ which is most easily formulated in terms of the dual dissipation potential $\Psi^*(q; \xi) = \frac{1}{2} \langle \mathbb{K}(q)\xi, \xi \rangle$ and is assumed to have the form

$$\Psi^*(\rho, z; \eta, \zeta) = \frac{1}{2} \sum_{n=1}^{N} \| \left[Q_n(z), \eta - \langle \alpha_n(z), \zeta \rangle H \right] \|^2_{\boldsymbol{C}_n(\rho,z)},$$

where $Q_n(z)$, $n = 1, .., N$ are coupling operators and $\|A\|^2_{\boldsymbol{C}} := \langle\!\langle\, \boldsymbol{C} A \,\|\, A \,\rangle\!\rangle$. The dissipative coupling via Q, and hence the interaction of the quantum system with the z component, occurs only through the commutators

$$[Q, \Xi] := Q\,\Xi \,-\, \Xi\, Q.$$

The coefficients α_m are assumed to satisfy the relation $\langle \alpha_n(z), \mathrm{D}E(z) \rangle \equiv 1$ such that $\Psi^*(q, \mathrm{D}\mathscr{E}(q)) \equiv 0$, which implies the second relation in (21.2). Since $\mathbb{J}(q)\mathrm{D}\mathscr{S}(q) \equiv 0$ also holds we have a GENERIC system.

In particular, we consider the case that z describes a finite number of macroscopic heat baths, each of which is fully characterized by its temperature $\theta_m(t)$ and its fixed heat capacity $c_m > 0$, i.e. with $z = \theta := (\theta_1, \ldots, \theta_M) \in]0, \infty[^M$ we have

$$E(\theta) = c \cdot \theta = \sum_{m=1}^{M} c_m \theta_m \quad \text{and} \quad S(\theta) = \sum_{m=1}^{M} c_m \log \theta_m.$$

Note that $\mathscr{E}$ is a linear functional in $q = (\rho, \theta)$ while $\mathscr{S}$ is strictly concave. Thus, for each given energy level E_0 there is a unique maximizer of $\mathscr{S}$ subject to $\mathscr{E}(\rho, \theta) = E_0$ called the thermodynamic equilibrium, which is given as

$$q_{\mathrm{eq}} = (\rho_{\mathrm{eq}}, \theta_{\mathrm{eq}}) \quad \text{with } \theta_{\mathrm{eq}} = \theta_*(1, , \ldots, 1) \text{ and } \rho_{\mathrm{eq}} = \tfrac{1}{Z} \exp\Big(\frac{-1}{k_{\mathrm{B}} \theta_*} H \Big),$$

where Z and θ_* depend on E_0 only. By the abstract theory of GENERIC (cf. Sect. 21.2.2) this q_{eq} is an equilibrium independently of the choice of $\mathbb{J}$ and $\mathbb{K}$ if (21.2) holds.

To prove a global existence result for the associated evolutionary system (21.1) we further specify the dissipation by choosing the super-operators $\boldsymbol{C}_m$ suitably. We follow [6, 27, 28] and use the inverse of the so-called Bogoliubov–Kubo–Mori metric (cf. [23, 31, 32]) given by

$$\mathscr{C}_\rho A := \int_0^1 \rho^s A \rho^{1-s} \,\mathrm{d}s \quad \text{and} \quad C_\rho^{-1} B = \int_0^\infty (\rho+sI)^{-1} B (\rho+sI)^{-1} \,\mathrm{d}s.$$

We call $\mathscr{C}_\rho$ the *canonical correlation operator* and refer to the above references for the relevance of this metric. There is a strong relation between the von Neumann entropy and $\mathscr{C}_\rho$ encoded in the commutator relations

$$\big[\mathscr{C}_\rho A, \log \rho\big] = \big[A, \rho\big] = \mathscr{C}_\rho\big[A, \log \rho\big], \tag{21.3}$$

for all $A = A^*$ and $\rho \in \mathfrak{R}$. Recall that $\mathrm{D}_\rho \mathscr{S} = -k_{\mathrm{B}} \log \rho$ and note that $\mathrm{D}_\rho^2 \mathscr{S} = \mathscr{C}_\rho^{-1}$, see [3, 23]. A proof of (21.3) is given in Sect. 21.4.1, where also the continuity of the mapping $\mathfrak{R} \ni \rho \mapsto \mathscr{C}_\rho$ is established.

In the case of M heat baths the mth heat bath may be coupled to the quantum system by the N_m coupling operators Q_m^n and $\boldsymbol{C}_m^n$, such that $N = N_1 + \ldots + N_M$. Choosing $\boldsymbol{C}_m^n = \mathscr{C}_\rho$ for all n and m and $\alpha_j = \frac{1}{c_{m(j)}} \boldsymbol{e}_{m(j)} \in \mathbb{R}^M$ we arrive at the system

$$\dot\rho = \mathrm{j}[\rho, H] - \sum_{m=1}^{M} \sum_{n=1}^{N_m} \big[\, Q_m^n \,,\, k_{\mathrm{B}}[Q_m^n, \rho] + \frac{1}{\theta_m} \mathscr{C}_\rho[Q_m^n, H] \,\big], \tag{21.4a}$$

$$\dot\theta = \kappa\big(\tfrac{c_m}{\theta_m}\big)_{m=1,\ldots,M} + \Big(\frac{1}{c_m} \sum_{n=1}^{N_m} \langle\!\langle\, k_{\mathrm{B}}[Q_m^n, \rho] + \tfrac{1}{\theta_m} \mathscr{C}_\rho[Q_m^n, H] \,\|\, [Q_m^n, H] \,\rangle\!\rangle\Big)_m, \tag{21.4b}$$

which will be discussed in Sect. 21.4. The equation for ρ displays dissipative double commutators $[Q, [Q, \rho]]$ of the Lindblad type in (21.4a) with a correction term involving the nonlinear term $\rho \mapsto [Q, \mathscr{C}_\rho[Q, H]]$. The latter term is continuous but not Lipschitz continuous on $\mathfrak{R}$. Moreover, we see a clear coupling to the temperatures θ_m in the different heat baths.

In Sect. 21.6 we discuss the isothermal case where the underlying Hilbert space $\boldsymbol{H}$ is finite-dimensional, i.e. $\boldsymbol{H} = \mathbb{C}^{\dim \boldsymbol{H}}$, which leads to the system

$$\dot\rho = \mathrm{j}[\rho, H] - \sum_{n=1}^{N} \big[Q^n, k_{\mathrm{B}}[Q^n, \rho] + \frac{1}{\theta_*} \mathscr{C}_\rho[Q^n, H]\big], \tag{21.5}$$

where now the temperature $\theta_* > 0$ is fixed. This is a dissipative Hamiltonian system of the form $\dot\rho = \big(J(\rho) - \frac{1}{\theta_*}K(\rho)\big)\mathrm{D}\mathscr{F}(\rho)$ for the free energy

$$\mathscr{F}(\rho) = \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle + k_\mathrm{B}\theta_* \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle.$$

Global existence of solutions is established in Theorem 21.3, and Theorem 21.4 provides conditions on the coupling operators Q^n such that all solutions satisfy $\rho(t) \to \rho_\mathrm{eq}$ for $t \to \infty$. In Sect. 21.5.3 we discuss the case $\dim \boldsymbol{H} = 2$ in detail and display the dynamics in suitable coordinates as an ODE in $\mathbb{R}^3$.

The linear Lindblad systems (cf. [6,15,16]) can be understood as approximations of (21.5) after replacing the super-operator $\mathscr{C}_\rho$ by the constant $\mathscr{C}_{\rho_\mathrm{eq}}$. Using (21.3) we have $\mathscr{C}_{\rho_\mathrm{eq}}[Q,H] = -k_\mathrm{B}\theta_*[Q,\rho_\mathrm{eq}]$ and arrive at (see Sect. 21.4.4),

$$\dot\rho = \mathrm{j}[\rho,H] - \sum_{n=1}^{N}\big[Q^n, k_\mathrm{B}[Q^n,\rho-\rho_\mathrm{eq}]\big] \quad \text{with } \rho_\mathrm{eq} = \frac{1}{Z}\exp\big(\frac{-1}{k_\mathrm{B}\theta_*}H\big). \tag{21.6}$$

In the recent work [3] Carlen and Maas consider a special version of this equation. First they assume $H = 0$ giving $\rho_\mathrm{eq} = \frac{1}{Z}I$, and second they work on the special Hilbert space $\boldsymbol{H} = \mathscr{F}_\mathrm{a}(\mathbb{H})$, which is the antisymmetric Fock space of a given finite-dimensional Hilbert space $\mathbb{H}$, i.e. $\dim \boldsymbol{H} = 2^J$ where $J = \dim \mathbb{H}$. Choosing a set $\{\, Q_j \mid j = 1,..,J \,\}$ satisfying the canonical anticommutation relations one defines the set of all coupling operators $\{\, Q^{\boldsymbol{\alpha}} = Q_1^{\alpha_1}\cdots Q_J^{\alpha_J} \mid \boldsymbol{\alpha} \in \{0,1\}^J \,\}$ with $N := 2^J$ elements. Then, the equation (21.6) turns into the dissipative system $\dot\rho = -\mathscr{N}\rho$, where $\mathscr{N}$ denotes the Fermionic number operator.

Section 21.7 discusses analogies between the gradient structure $\dot\rho = \mathbb{K}^{\boldsymbol{Q}}(\rho)\mathrm{D}\mathscr{S}_\mathrm{qm}(\rho)$ used here and in [3] and two other entropy gradient structures for stochastic problems, namely that for the Fokker–Planck equation introduced in [11, 25] and that for reversible Markov chains introduced in [4, 17, 20]. All structures have the common feature that the mapping $\rho \mapsto K(\rho)$ is homogeneous of degree 1, i.e. $K(\lambda\rho) = \lambda K(\rho)$. Moreover, K is defined in terms of couplings in the discrete case or derivatives relating to transportation in the continuous case. In [34] there is a related transport formulation of the Schrödinger equation, which is based on the Madelung form of quantum mechanics. This is a single-particle model, while the approach here and in [3] uses the many-particle formulation in terms of density matrices, which is necessary for couplings to exterior dissipative systems.

21.2 The GENERIC Framework

The framework of GENERIC was introduced by Öttinger and Grmela in [8, 24]. It is based on a quintuple $(\mathscr{Q},\mathscr{E},\mathscr{S},\mathbb{J},\mathbb{K})$, where the smooth functionals $\mathscr{E}$ and $\mathscr{S}$ on the state space $\mathscr{Q}$ denote the total energy and the total entropy, respectively. Moreover, $\mathscr{Q}$ carries two geometric structure, namely a Poisson structure $\mathbb{J}$ and a

dissipative structure $\mathbb{K}$, i.e. for each $q \in \mathcal{Q}$ the operators $\mathbb{J}(q)$ and $\mathbb{K}(q)$ map the cotangent space $\mathrm{T}_q^*\mathcal{Q}$ into the tangent space $\mathrm{T}_q\mathcal{Q}$. The evolution of the system is given by the differential equation

$$\dot{q} = \mathbb{J}(q)\mathrm{D}\mathscr{E}(q) + \mathbb{K}(q)\mathrm{D}\mathscr{S}(q), \tag{21.7}$$

where $\mathrm{D}\mathscr{E}$ and $\mathrm{D}\mathscr{S}$ are the differentials taking values in the cotangent space.

We refer to [7, 8, 24] for applications in fluid mechanics, to [10] for electromagnetism, to [19] for applications of GENERIC in elastoplastic materials. The book [26] contains a general introduction with an emphasis towards numerical simulation, while [22] surveys modeling aspects. Subsequently we will mainly dwell on the quantum mechanical papers [27, 28].

21.2.1 The Structure of GENERIC

The basic conditions on the geometric structures $\mathbb{J}$ and $\mathbb{K}$ are the symmetries

$$\mathbb{J}(q) = -\mathbb{J}(q)^* \text{ and } \mathbb{K}(q) = \mathbb{K}(q)^* \tag{21.8}$$

and the structural properties

$$\begin{aligned} &\mathbb{J} \text{ satisfies Jacobi's identity,} \\ &\mathbb{K}(q) \text{ is positive semi-definite, i.e., } \langle \xi, \mathbb{K}(q)\xi \rangle \geq 0. \end{aligned} \tag{21.9}$$

Jacobi's identity for $\mathbb{J}$ holds, if for all functions $\mathscr{F}_j : \mathcal{Q} \to \mathbb{R}$ we have

$$\{\{\mathscr{F}_1, \mathscr{F}_2\}_{\mathbb{J}}, \mathscr{F}_3\}_{\mathbb{J}} + \{\{\mathscr{F}_2, \mathscr{F}_3\}_{\mathbb{J}}, \mathscr{F}_1\}_{\mathbb{J}} + \{\{\mathscr{F}_3, \mathscr{F}_1\}_{\mathbb{J}}, \mathscr{F}_2\}_{\mathbb{J}} \equiv 0,$$

where the Poisson bracket is defined via $\{\mathscr{F}, \mathscr{G}\}_{\mathbb{J}}(q) := \langle \mathrm{D}\mathscr{F}(q), \mathbb{J}(q)\mathrm{D}\mathscr{G}(q) \rangle$. Finally, the central condition states that the energy functional does not contribute to dissipative mechanisms and that the entropy functional does not contribute to reversible dynamics, which is the following *non-interaction condition (NIC)*:

$$\forall\, q \in \mathcal{Q}: \quad \mathbb{J}(q)\mathrm{D}\mathscr{S}(q) = 0 \quad \text{and} \quad \mathbb{K}(q)\mathrm{D}\mathscr{E}(q) = 0. \tag{21.10}$$

Of course, the structure of GENERIC is geometric in the sense that it is invariant under coordinate transformations, see [19, 22].

21.2.2 Properties of GENERIC Systems

The first observation is that (21.9) and (21.10) imply energy conservation and entropy increase:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{E}(q(t)) = \langle \mathrm{D}\mathscr{E}(q), \dot{q}\rangle = \langle \mathrm{D}\mathscr{E}(q), \mathbb{J}\mathrm{D}\mathscr{E} + \mathbb{K}\mathrm{D}\mathscr{S}\rangle = 0 + 0 = 0, \tag{21.11}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{S}(q(t)) = \langle \mathrm{D}\mathscr{S}(q), \dot{q}\rangle = \langle \mathrm{D}\mathscr{S}(q), \mathbb{J}\mathrm{D}\mathscr{E} + \mathbb{K}\mathrm{D}\mathscr{S}\rangle = 0 + \langle \mathrm{D}\mathscr{S}, \mathbb{K}\mathrm{D}\mathscr{S}\rangle \geq 0. \tag{21.12}$$

Note that we would need much less than the two conditions (21.9) and (21.10) to guarantee these two properties. However, the next property needs (21.10) in its full strength.

Next, we show that equilibria can be obtained by the *maximum entropy principle*. If x_{eq} maximizes $\mathscr{S}$ under the constraint $\mathscr{E}(q) = E_0$, then we obtain a Lagrange multiplier $\lambda_{\mathrm{eq}} \in \mathbb{R}$ such that $\mathrm{D}\mathscr{S}(q_{\mathrm{eq}}) = \lambda_{\mathrm{eq}}\mathrm{D}\mathscr{E}(q_{\mathrm{eq}})$. Assuming that $\lambda_{\mathrm{eq}} \neq 0$ we immediately find that q_{eq} is an equilibrium of (21.7). Indeed, $\mathbb{J}(q_{\mathrm{eq}})\mathrm{D}\mathscr{E}(q_{\mathrm{eq}}) = \frac{1}{\lambda_{\mathrm{eq}}}\mathbb{J}(q_{\mathrm{eq}})\mathrm{D}\mathscr{S}(q_{\mathrm{eq}}) = 0$ and $\mathbb{K}(q_{\mathrm{eq}})\mathrm{D}\mathscr{S}(q_{\mathrm{eq}}) = \lambda_{\mathrm{eq}}\mathbb{K}(q_{\mathrm{eq}})\mathrm{D}\mathscr{E}(q_{\mathrm{eq}}) = 0$, where we have used the NIC (21.10).

Vice versa, for every steady state q_{eq} of (21.7) we must have

$$\mathbb{J}(q_{\mathrm{eq}})\mathrm{D}\mathscr{E}(q_{\mathrm{eq}}) = 0 \text{ and } \mathbb{K}(q_{\mathrm{eq}})\mathrm{D}\mathscr{S}(q_{\mathrm{eq}}) = 0. \tag{21.13}$$

Thus, in a steady state there cannot be any balancing between reversible and irreversible forces, both have to vanish independently. To see this we simply recall the entropy production relation (21.12), which implies $\langle \mathrm{D}\mathscr{S}(q_{\mathrm{eq}}), \mathbb{K}(q_{\mathrm{eq}})\mathrm{D}\mathscr{S}(q_{\mathrm{eq}})\rangle = 0$ for any steady state. Since $\mathbb{K}(q_{\mathrm{eq}})$ is positive semidefinite, this implies the second identity in (21.13). The first identity then follows from $\dot{q} \equiv 0$ in (21.7).

21.2.3 Isothermal Systems

Often temperature effects can be neglected and the model can be approximated by an isothermal system. We show here how this can be deduced consistently from the GENERIC form if we add some coupling to an external heat bath fixed to a given temperature $\theta_* > 0$. In particular, we will replace the two functionals $\mathscr{E}$ and $\mathscr{S}$ by one, namely the free energy $\mathscr{F}_*$ at the given temperature θ_*.

We start from a general system for the variable $q = (y, \theta)$ in the form

$$\begin{pmatrix}\dot{y}\\ \dot{\theta}\end{pmatrix} = \begin{pmatrix} J & \alpha \\ -\alpha^{\mathsf{T}} & 0\end{pmatrix}\begin{pmatrix}\mathrm{D}_y\overline{\mathscr{E}}\\ \mathrm{D}_\theta\overline{\mathscr{E}}\end{pmatrix} + \begin{pmatrix} K & \beta \\ \beta^{\mathsf{T}} & \Lambda\end{pmatrix}\begin{pmatrix}\mathrm{D}_y\overline{\mathscr{S}}\\ \mathrm{D}_\theta\overline{\mathscr{S}}\end{pmatrix} - \begin{pmatrix}0\\ A(\theta-\theta_*)\end{pmatrix}, \tag{21.14}$$

where the coupling operator A is assumed to be positive definite. Defining the functional $\mathscr{F}^\circ(y,\theta) = \overline{\mathscr{E}}(y,\theta) - \theta_* \overline{\mathscr{S}}(y,\theta)$, the system (21.14) takes the form

$$\begin{pmatrix} \dot y \\ \dot\theta \end{pmatrix} = \begin{pmatrix} J - \frac{1}{\theta_*}K & \alpha - \frac{1}{\theta_*}\beta \\ -\alpha^{\mathsf T} - \frac{1}{\theta_*}\beta^{\mathsf T} & -\frac{1}{\theta_*}\Lambda \end{pmatrix} \begin{pmatrix} \mathrm{D}_y \mathscr{F}^\circ \\ \mathrm{D}_\theta \mathscr{F}^\circ \end{pmatrix} - \begin{pmatrix} 0 \\ A(\theta - \theta_*) \end{pmatrix}.$$

The equation for $\dot y$ reads $\dot y = \big(J - \frac{1}{\theta_*}K\big)\mathrm{D}\mathscr{F}^\circ + \big(\alpha - \frac{1}{\theta_*}\beta\big)\mathrm{D}_\theta \mathscr{F}^\circ$, where the last term vanishes to order $O(\|\theta - \theta_*\|)$, see [19, Sect. 2.6] for more details. Defining the isothermal free energy $\mathscr{F}_*(y) = \mathscr{F}^\circ(y,\theta_*)$ and neglecting all terms of order $O(\theta - \theta_*)$ we arrive at the isothermal damped Hamiltonian system

$$\dot y = \big(J(y,\theta_*) - \frac{1}{\theta_*}K(y,\theta_*)\big)\mathrm{D}\mathscr{F}_*(y). \tag{21.15}$$

Note that J still defines a Poisson structure and that K is positive semi-definite. Hence, $\mathscr{F}_*$ is a Liapunov function for (21.15).

21.3 Coupling of Quantum and Dissipative Mechanics

21.3.1 Quantum Mechanics

The quantum mechanical system is described by states in a complex Hilbert space $\boldsymbol{H}$ with scalar product $\langle \cdot \,|\, \cdot \rangle$ and a Hamiltonian (operator) $H : D(H) \to \boldsymbol{H}$, which is assumed to be selfadjoint and semi-bounded, namely

$$\exists\, h_{\min} \in \mathbb{R}\ \forall\, \psi \in D(H): \quad \langle H\psi | \psi \rangle \geq h_{\min} \|\psi\|^2.$$

The associated Hamiltonian dynamics is given via the Schrödinger equation

$$\dot\psi = -\mathrm{j} H\psi, \quad \text{where } \mathrm{j} = \frac{\mathrm{i}}{\hbar}, \tag{21.16}$$

which has the solution $\psi(t) = \mathrm{e}^{-\mathrm{j}tH}\psi(0)$.

We denote by $\mathscr{L}^p(\boldsymbol{H})$ the Banach space of compact operators A from $\boldsymbol{H}$ into itself, such that

$$\|A\|_p := \Big(\sum_{j=1}^{\infty} \sigma_j(A)^p\Big)^{1/p} < \infty,$$

where σ_j is the jth singular value of A (i.e., the jth largest eigenvalue of $(A^*A)^{1/2}$). We refer to [5, Ch. III] for this and the following standard properties.

Moreover, $\mathscr{L}^\infty(\boldsymbol{H})$ is the set of bounded linear operators with $\|A\|_\infty$ denoting the standard operator norm. For $1 \le p_1 < p_2 < \infty$ we have

$$\mathscr{L}^{p_1}(\boldsymbol{H}) \subset \mathscr{L}^{p_2}(\boldsymbol{H}) \text{ and } \|A\|_{p_1} \ge \|A\|_{p_2} \text{ for all } A \in \mathscr{L}^{p_1}(\boldsymbol{H}).$$

Moreover, if $1/p = 1/p_1 + 1/p_2 \le 1$, then Hölder's estimate holds

$$\|AB\|_q \le \|A\|_{p_1} \|B\|_{p_2} \text{ for all } A \in \mathscr{L}^{p_1}(\boldsymbol{H}),\ B \in \mathscr{L}^{p_2}(\boldsymbol{H}). \tag{21.17}$$

On $\mathscr{L}^1(\boldsymbol{H})$ the trace operator $\mathrm{tr} : \mathscr{L}^1(\boldsymbol{H}) \to \mathbb{C},\ A \mapsto \sum_{j=1}^\infty \langle A\phi_j | \phi_j \rangle$ is well defined, where $\{\, \phi_j \mid j \in \mathbb{N} \,\}$ is an arbitrary complete orthonormal system in $\boldsymbol{H}$. Using the dual pairing

$$\langle\!\langle\, A \,\|\, B \,\rangle\!\rangle := \mathrm{tr}(AB^*), \text{where } \mathrm{tr}(\psi \otimes \overline{\phi}) = \langle \psi | \phi \rangle,$$

we see that $\mathscr{L}^2(\boldsymbol{H})$ is a Hilbert space and $\mathscr{L}^p(\boldsymbol{H})' \cong \mathscr{L}^q(\boldsymbol{H})$ for $1 < p, q < \infty$ with $1/p + 1/q = 1$. We will need the following elementary commutator relations:

$$\begin{aligned} &\langle\!\langle\, A \,\|\, BC \,\rangle\!\rangle = \langle\!\langle\, B^* A \,\|\, C \,\rangle\!\rangle = \langle\!\langle\, AC^* \,\|\, B \,\rangle\!\rangle, \text{ giving} \\ &\langle\!\langle\, A \,\|\, [B, C] \,\rangle\!\rangle = \langle\!\langle\, [B^*, A] \,\|\, C \,\rangle\!\rangle = -\langle\!\langle\, [A, B^*] \,\|\, C \,\rangle\!\rangle = \langle\!\langle\, [A, C^*] \,\|\, B \,\rangle\!\rangle. \end{aligned} \tag{21.18}$$

To couple a quantum system to a macroscopic one we need to describe it in terms of the multi-particle form using the density matrices

$$\rho \in \mathfrak{R} := \{\, \rho \in \mathscr{L}^1(\boldsymbol{H}) \mid \rho = \rho^* \ge 0,\ \mathrm{tr}\, \rho = 1 \,\}.$$

Hence, each $\rho \in \mathfrak{R}$ has the representation

$$\rho = \sum_{j=1}^\infty r_j\, \psi_j \otimes \overline{\psi}_j, \tag{21.19}$$

where $r_j \ge 0$, $\sum_1^\infty r_j = 1$, and $\{\, \psi_j \mid j \in \mathbb{N} \,\}$ is an orthonormal set. (Note that $(\psi \otimes \overline{\phi})a := \langle a | \phi \rangle \psi$ and $(\psi \otimes \overline{\phi})A = \psi \otimes \overline{A}^* \phi$.) Using (21.16) the evolution of ρ is given via the Liouville equation

$$\dot{\rho} = \mathrm{j}[\rho, H], \quad \text{where } [\rho, H] := \rho H - H\rho. \tag{21.20}$$

This is consistent with (21.16) if each of the ψ_j solves (21.16) while all r_j are constant.

Below we will use the von Neumann entropy

$$S_{\mathrm{qm}}(\rho) = -k_{\mathrm{B}} \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle = -k_{\mathrm{B}}\, \mathrm{tr}(\rho \log \rho) = -k_{\mathrm{B}} \sum_{j=1}^\infty r_j \log r_j. \tag{21.21}$$

It is easy to see that the entropy remains constant for the Hamiltonian system (21.20).

While the single-commutator equation (21.20) gives rise to Hamiltonian dynamics, the following double-commutator equation leads to dissipative dynamics:

$$\dot{\eta} = -[Q,[Q,\eta]] = -(Q^2\eta - 2Q\eta Q + \eta Q^2), \tag{21.22}$$

where $Q \in \mathscr{L}^p_{\mathrm{S}}(\boldsymbol{H}) := \{\, A \in \mathscr{L}^p(\boldsymbol{H}) \mid A = A^* \,\}$ is a given operator. As Q can be written in the form $\sum_{n=1}^{\infty} q_n \phi_n \otimes \overline{\phi}_n$ with $q_n \in \mathbb{R}$ we easily find that the corresponding coefficients $\eta_{nm}(t) = \langle \rho(t)\phi_n | \phi_m \rangle \in \mathbb{C}$ satisfy the ODE $\dot{\eta}_{nm} = -(q_n - q_m)^2 \eta_{nm}$. Thus, the double-commutator evolution diminishes all off-diagonal elements, while the diagonal elements of η (and hence the trace) remain unchanged. Hence, an often used dissipative quantum system is the Lindblad equation of the form

$$\dot{\rho} = \mathrm{j}[\rho, H] - \sum_{n=1}^{N} \big[Q^n, [Q^n, \rho - \rho_{\mathrm{eq}}]\big]. \tag{21.23}$$

While it is well known that this equation preserves the property that $\rho(t) \in \mathfrak{R}$ it is less clear what the longtime dynamics is and what suitable Liapunov function are. We will see later that the GENERIC framework leads to a correction of (21.23), see Sect. 21.4.4.

21.3.2 Dissipative Evolution

We assume that an additional variable z is present in the model that is dissipative. For simplicity, we assume that z lies in a closed subset $Z \subset \mathbb{R}^N$. The evolution is assumed to be purely dissipative in the sense that it is a gradient flow with respect to the entropy $S : Z \to \mathbb{R}$, namely

$$\dot{z} = K(z)\mathrm{D}S(z), \text{ where } K(z) = K(z)^{\mathsf{T}} \geq 0. \tag{21.24}$$

We immediately obtain entropy production in the form

$$\frac{\mathrm{d}}{\mathrm{d}t} S(z(t)) = \mathrm{D}S(z) \cdot K(z)\mathrm{D}S(z) \geq 0.$$

If an energy $E : Z \to \mathbb{R}$ is conserved along solutions z of (21.24), then the relation $\mathrm{D}E(z) \cdot K(z)\mathrm{D}S(z) \equiv 0$ has to hold. Very often one imposes the stronger condition $K(z)\mathrm{D}E(z) \equiv 0$, which is certainly sufficient, but not at all necessary.

21.3.3 Coupling of the Models

We now couple the quantum and the dissipative system in the GENERIC sense. The joint state space is $\mathscr{Q} = \mathfrak{R} \times Z \subset \mathscr{L}^1_{\mathrm{S}}(\boldsymbol{H}) \times \mathbb{R}^N$, where the state is given by the pair (ρ, z). The energy functional $\mathscr{E}$ and the entropy functional $\mathscr{S}$ take the form

$$\mathscr{E}(\rho, z) = \langle\!\langle\, H(z) \,\|\, \rho \,\rangle\!\rangle + E(z) \quad \text{and} \quad \mathscr{S}(\rho, z) = -k_{\mathrm{B}} \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle + S(z).$$

In the general case, the Hamiltonian H may depend on the dissipative variable z, but for our mathematical results we assume that H is independent of z. For the Poisson structure we assume that the variable z is totally dissipative, which means that $\mathbb{J}$ has block structure on the form

$$\mathbb{J}(\rho, z) = \begin{pmatrix} \mathrm{j}[\rho, \,\square\,] & 0 \\ 0 & 0 \end{pmatrix},$$

where $\square$ indicates, where the corresponding component of the vector applied from the right has to be inserted. The dissipation operator $\mathbb{K}$ will be defined in terms of the dissipation potential, which is quadratic in the driving forces

$$\begin{pmatrix} \mu \\ \zeta \end{pmatrix} = \mathrm{D}\mathscr{S}(\rho, z) = \begin{pmatrix} \mathrm{D}_\rho \mathscr{S}(\rho) \\ \mathrm{D}_z \mathscr{S}(\rho, z) \end{pmatrix} = \begin{pmatrix} -k_{\mathrm{B}} \log \rho \\ \mathrm{D}S(z) \end{pmatrix}.$$

In the following we use a special ansatz for the dissipation potential that is based on the physical observation that a quantum mechanical system interacts with its environment only via the commutators with respect to suitable coupling operators $Q_m(z) \in \mathscr{L}^\infty_{\mathrm{S}}(\boldsymbol{H})$. This leads to

$$\left\langle \mathbb{K}(\rho, z) \begin{pmatrix} \mu \\ \zeta \end{pmatrix} \middle| \begin{pmatrix} \mu \\ \zeta \end{pmatrix} \right\rangle = \zeta \cdot K_{\mathrm{diss}}(\rho, z)\zeta + \sum_{m=1}^{M} \| [Q_m(z), \mu - \alpha_m(\rho, z) \cdot \zeta\, H(z)] \|^2_{\boldsymbol{C}_m(\rho, z)},$$

where $\alpha_m(\rho, z) \in \mathbb{R}^N$, $K_{\mathrm{diss}}(\rho, z) \in \mathbb{R}^{N \times N}$ is symmetric and positive semidefinite, and the super-operators $\boldsymbol{C}_m(\rho, z)$ are symmetric and positive semidefinite on $\mathscr{L}^2(\boldsymbol{H})$. (Here super-operators are linear operator on $\mathscr{L}^p(\boldsymbol{H})$.) The norm $\| \cdot \|_{\boldsymbol{C}}$ is defined via

$$\| A \|_{\boldsymbol{C}} := \langle\!\langle\, \boldsymbol{C}A \,\|\, A \,\rangle\!\rangle^{1/2} = \| \boldsymbol{C}^{1/2} A \|_2.$$

To simplify the following notations we introduce another super-operator

$$\mathbb{K}^Q_{\boldsymbol{C}} : A \mapsto [Q^*, \boldsymbol{C}[Q, A]].$$

Using (21.18) the associated linear operator K takes the form

$$\mathbb{K}(q) = \begin{pmatrix} 0 & 0 \\ 0 & K_{\mathrm{diss}} \end{pmatrix} + \sum_{m=1}^{M} \begin{pmatrix} \mathbb{K}^{Q_m}_{\boldsymbol{C}_m} & -(\alpha_m \cdot \square)\, \mathbb{K}^{Q_m}_{\boldsymbol{C}_m} H \\ -\langle\!\langle\, \mathbb{K}^{Q_m}_{\boldsymbol{C}_m} \square \parallel H \,\rangle\!\rangle \alpha_m & \langle\!\langle\, \mathbb{K}^{Q_m}_{\boldsymbol{C}_m} H \parallel H \,\rangle\!\rangle \alpha_m \otimes \overline{\alpha}_m \end{pmatrix},$$

where we have omitted the arguments ρ and z for simplicity. The occurrence of double commutators $\mathbb{K}^{Q_m}_{\boldsymbol{C}_m}$ indicates the dissipative nature of K.

To satisfy the NIC (21.10) we still need an assumption on $\mathbb{K}$, namely

$$K_{\mathrm{diss}}(\rho, z)\mathrm{D}_z \mathscr{E}(\rho, z)) = 0 \text{ and } \alpha_m(\rho, z) \cdot \mathrm{D}_z \mathscr{E}(\rho, z)) = 1 \quad \text{for all } m, \rho, z. \tag{21.25}$$

These assumptions guarantee $\mathbb{K}(q)\mathrm{D}\mathscr{E}(q) \equiv 0$. Thus, $\mathbb{J}(\rho, z)$ and the symmetric and positive linear operator $\mathbb{K}$ satisfy the NIC (21.10), i.e. $\mathbb{J}(\rho, z)\mathrm{D}\mathscr{S}(\rho, z) \equiv 0$ and $\mathbb{K}(\rho, z)\mathrm{D}\mathscr{E}(\rho, z) \equiv 0$. Now the GENERIC formalism provides the evolutionary system for $q = (\rho, z)$, namely

$$\begin{aligned} \dot\rho &= \mathrm{j}[H(z), \rho] - \sum_{1}^{M} \mathbb{K}^{Q_m(z)}_{\boldsymbol{C}_m(\rho,z)} \Big(k_{\mathrm{B}} \log \rho + \alpha_m(\rho, z) \cdot \mathrm{D}S(z) H(z) \Big), \\ \dot z &= K_{\mathrm{diss}}(\rho, z)\mathrm{D}S(z) + \\ &\quad + \sum_{1}^{M} \langle\!\langle\, \mathbb{K}^{Q_m(z)}_{\boldsymbol{C}_m(\rho,z)} \big(k_{\mathrm{B}} \log \rho + \alpha_m(\rho, z) \cdot \mathrm{D}S(z) H(z) \big) \parallel H(z) \,\rangle\!\rangle \alpha_m(\rho, z). \end{aligned} \tag{21.26}$$

In the following we will reduce the generality in order to obtain more structure.

21.4 Canonical Correlation

21.4.1 The Kubo–Mori Metric

Following [6, 28] we introduce a *canonical correlation operator* which associates with the density matrix ρ in the following way: for each $\rho \in \mathfrak{R}$ we define

$$\mathscr{C}_\rho : \begin{cases} \mathscr{L}^\infty(\boldsymbol{H}) \to \mathscr{L}^1(\boldsymbol{H}), \\ A \mapsto \int_0^1 \rho^s A \rho^{1-s}\, \mathrm{d}s. \end{cases} \tag{21.27}$$

The boundedness follows from Hölder's estimate (21.17) giving

$$\|\mathscr{C}_\rho A\|_q \le \|\rho\|_{p_1} \|A\|_{p_2} \quad \text{for } \frac{1}{q} = \frac{1}{p_1} + \frac{1}{p_2}.$$

We further have the identity $(\mathscr{C}_\rho A)^* = \mathscr{C}_\rho(A^*)$ (using reparametrization and $\rho = \rho^*$). If ρ has the representation $\sum r_j \psi_j \otimes \overline{\psi}_j$, then we have

$$\mathscr{C}_\rho A = \sum_{j,k=1}^{\dim \boldsymbol{H}} \Lambda(r_j, r_k)\langle A\psi_k|\psi_j\rangle \psi_j \otimes \overline{\psi}_k \text{ and } \langle\!\langle\, \mathscr{C}_\rho A \,\|\, A \,\rangle\!\rangle = \sum_{j,k=1}^{\dim \boldsymbol{H}} \Lambda(r_j, r_k)\big|\langle A\psi_k|\psi_j\rangle\big|^2, \tag{21.28}$$

where the continuous function $\Lambda : [0,\infty[^2 \to [0,\infty[$ is given by

$$\Lambda(a,b) = \int_0^1 a^s b^{1-s}\,\mathrm{d}s = \begin{cases} \frac{a-b}{\log a - \log b} & \text{for } a,b>0 \text{ and } a \neq b, \\ a & \text{for } a = b \geq 0, \\ 0 & \text{for } \min\{a,b\} = 0. \end{cases} \tag{21.29}$$

Note that Λ satisfies the bounds $\min\{a,b\} \leq \sqrt{ab} \leq \Lambda(a,b) \leq \frac{1}{2}(a+b) \leq \max\{a,b\}$.

Thus, we have on $\mathscr{L}^\infty(\boldsymbol{H})$ the relations

$$\langle\!\langle\, \mathscr{C}_\rho A \,\|\, B \,\rangle\!\rangle = \langle\!\langle\, A \,\|\, \mathscr{C}_\rho B \,\rangle\!\rangle \quad \text{and} \quad \langle\!\langle\, \mathscr{C}_\rho A \,\|\, A \,\rangle\!\rangle \geq 0,$$

which induce the scalar product $(A,B) \mapsto \langle\!\langle\, \mathscr{C}_\rho A \,\|\, B \,\rangle\!\rangle$. This scalar product is called the *canonical correlation between A and B for the given state ρ* in [12,28].

For $\rho > 0$ and $\dim \boldsymbol{H} < \infty$ the operator $\mathscr{C}_\rho$ is invertible, namely

$$\mathscr{G}_\rho A := \mathscr{C}_\rho^{-1} A = \int_0^\infty (\rho + sI)^{-1} A (\rho + sI)^{-1}\,\mathrm{d}s. \tag{21.30}$$

This formula is easily derived from (21.28) using $1/\Lambda(a,b) = \int_0^\infty \big((a+s)(b+s)\big)^{-1}\,\mathrm{d}s$. The tensor $\mathscr{G}_\rho$ defines the *Bogoliubov–Kubo–Mori metric* on the set of density matrices as follows, see [23,29,30]. For a curve $[0,1] \ni s \mapsto \tilde{\rho}(s)$ we define its Kubo–Mori length $\ell(\tilde{\rho})$ via

$$\ell(\tilde{\rho})^2 = \int_0^1 \langle\!\langle\, \tilde{\rho}'(s) \,\|\, \mathscr{G}_{\tilde{\rho}(s)} \tilde{\rho}'(s) \,\rangle\!\rangle\,\mathrm{d}s.$$

The relevance of the Bogoliubov–Kubo–Mori metric as a generalization of the Fisher information metric is discussed in the above references. For our usage, the most important fact is the connection to the von Neumann entropy S_{qm}, see (21.21). In fact, $\mathscr{G}_\rho$ can be identified by its Hessian, namely

$$\langle\!\langle\, A \,\|\, \mathrm{D}^2 S_{\mathrm{qm}}(\rho) B \,\rangle\!\rangle = -k_{\mathrm{B}} \langle\!\langle\, A \,\|\, \mathscr{G}_\rho B \,\rangle\!\rangle.$$

Open problem 1. *Is the mapping* $\mathfrak{R} \ni \rho \mapsto \langle\!\langle A \,\|\, \mathscr{C}_\rho A \rangle\!\rangle$ *concave for all* A*? A positive answer would give good metric properties for the Riemannian manifold* $(\mathfrak{R}, \mathscr{G}_\rho)$*, see [14].*

The following identities, which go back to [13], play an important role in the field of dissipative effects in quantum mechanics and manifest the relation between the von Neumann entropy S_{qm} and the canonical correlation operator $\mathscr{C}_\rho$. See also [3, Lem. 3.1], where a general calculus for operator functions is developed.

Proposition 21.1. *For all* $A \in \mathscr{L}^\infty_{\mathrm{S}}(\boldsymbol{H})$ *and all* $\rho \in \mathfrak{R}$ *with* $\log \rho \in \mathscr{L}^\infty_{\mathrm{S}}(\boldsymbol{H})$ *we have*

$$\big[\mathscr{C}_\rho A, \log \rho\big] = \big[A, \rho\big] = \mathscr{C}_\rho\big[A, \log \rho\big]. \tag{21.31}$$

Proof. For the convenience of the reader we give a full proof. With $P = \log \rho$ we have $\rho = \mathrm{e}^P$ and

$$\begin{aligned}\big[\mathscr{C}_\rho A, \log \rho\big] &= \int_0^1 \big[\mathrm{e}^{sP} A \mathrm{e}^{(1-s)P}, P\big]\,\mathrm{d}s = \int_0^1 \mathrm{e}^{sP} A(\mathrm{e}^{(1-s)P} P) - (P\mathrm{e}^{sP}) A \mathrm{e}^{(1-s)P}\,\mathrm{d}s \\ &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\big(-\mathrm{e}^{sP} A \mathrm{e}^{(1-s)P}\big)\,\mathrm{d}s = -\big(\mathrm{e}^{sP} A \mathrm{e}^{(1-s)P}\big)_0^1 = \big[A, \mathrm{e}^P\big] = \big[A, \rho\big].\end{aligned}$$

With slightly different grouping, we also obtain the second identity, namely

$$\begin{aligned}\mathscr{C}_\rho\big[A, \log \rho\big] &= \int_0^1 \mathrm{e}^{sP}(AP - PA)\mathrm{e}^{(1-s)P}\,\mathrm{d}s = \int_0^1 \mathrm{e}^{sP} A(P\mathrm{e}^{(1-s)P}) - (\mathrm{e}^{sP} P) A \mathrm{e}^{(1-s)P}\,\mathrm{d}s \\ &= \int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\big(-\mathrm{e}^{sP} A \mathrm{e}^{(1-s)P}\big)\,\mathrm{d}s = \ldots = \big[A, \rho\big].\end{aligned}$$

Thus, both identities are established. □

The following result shows that $\mathscr{C}_\rho \in \mathrm{Lin}(\mathrm{L}^\infty_{\mathrm{S}}(\boldsymbol{H}), \mathrm{L}^1_{\mathrm{S}}(\boldsymbol{H}))$ depends continuously on $\rho \in \mathfrak{R} \in \mathrm{L}^1_{\mathrm{S}}(\boldsymbol{H})$ with respect to the norm topology. This result will be crucial for our existence theory. This property is derived from Phillips' result on Hölder continuity for $\rho \mapsto \rho^s$ in the appropriate norms.

Proposition 21.2. *For all* $\rho_1, \rho_2 \in \mathrm{L}^1_{\mathrm{S}}(\boldsymbol{H})$ *with* $\rho_1, \rho_2 \geq 0$ *and all* $A \in \mathrm{L}^\infty_{\mathrm{S}}(\boldsymbol{H})$ *we have*

$$\|\mathscr{C}_{\rho_1} A - \mathscr{C}_{\rho_2} A\|_1 \leq \omega\Big(\tfrac{\|\rho_1 - \rho_2\|_1}{\|\rho_1\|_1 + \|\rho_2\|_1}\Big)\big(\|\rho_1\|_1 + \|\rho_2\|_1\big)\|A\|_\infty, \quad \omega(\nu) = 2\tfrac{1-\nu}{|\log \nu|}. \tag{21.32}$$

Proof. The proof relies on Phillips' inequality (see [2, Thm. 4]):

$$X = X^* \geq Y = Y^* \geq 0 \text{ and } p \geq 1 \quad \Longrightarrow \quad \|X^{1/p} - Y^{1/p}\|_p \leq \|X - Y\|_1^{1/p}. \tag{21.33}$$

We need the result for the two non-ordered operators ρ_1 and ρ_2. For this we define $V = \rho_2 - \rho_1$ and decompose it into its positive and negative parts, namely $V = V_+ - V_-$ with $V_+, V_- \geq 0$. With $Z := \rho_1 + V_+$ we have

$$Z = \rho_1 + V_+ \geq \rho_1 \geq 0 \quad \text{and} \quad Z = \rho_2 + V_- \geq \rho_2 \geq 0.$$

Applying Phillips' inequality (21.33) to these two ordered pairs we obtain

$$\|Z^{1/p} - \rho_1^{1/p}\|_p \leq \|Z - \rho_1\|_1^{1/p} = \|V_+\|_1^{1/p}, \quad \|Z^{1/p} - \rho_2^{1/p}\|_p \leq \|Z - \rho_2\|_1^{1/p} = \|V_-\|_1^{1/p}.$$

Thus, the triangle estimate gives the desired generalization of (21.33), namely

$$\|\rho_2^{1/p} - \rho_1^{1/p}\|_p \leq \|V_+\|_1^{1/p} + \|V_-\|_1^{1/p} \leq 2^{1-1/p}\|V\|_1^{1/p} = 2^{1-1/p}\|\rho_1 - \rho_2\|_1^{1/p}, \tag{21.34}$$

where we have used $\|V_+\|_1 + \|V_-\|_1 = \|V\|_1$ and

$$\alpha^{1/p} + \beta^{1/p} \leq 2^{1-1/p}(\alpha+\beta)^{1/p} \text{ for all } p \geq 1,\ \alpha, \beta \geq 0. \tag{21.35}$$

We now estimate $C_{\rho_1} A - C_{\rho_2} A$ in $\mathrm{L}^1_{\mathrm{S}}(\boldsymbol{H})$ as follows:

$$\begin{aligned}
\|C_{\rho_1} A - C_{\rho_2} A\|_1 &\leq \int_{s=0}^1 \|\rho_1^s A \rho_1^{1-s} - \rho_2^s A \rho_1^{1-s}\|_1 + \|\rho_2^s A \rho_1^{1-s} - \rho_2^s A \rho_2^{1-s}\|_1 \,\mathrm{d}s \\
&\leq \int_{s=0}^1 \|\rho_1^s - \rho_2^s\|_{1/s} \|A\|_\infty \|\rho_1^{1-s}\|_{1/(1-s)} + \|\rho_2^s\|_{1/s} \|A\|_\infty \|\rho_1^{1-s} - \rho_2^{1-s}\|_{1/(1-s)} \,\mathrm{d}s \\
&\leq \|A\|_\infty \int_{r=0}^1 \|\rho_1^r - \rho_2^r\|_{1/r} \big(\|\rho_1\|_1^{1-r} + \|\rho_2\|_1^{1-r} \big) \,\mathrm{d}r \\
&\leq \|A\|_\infty \int_{r=0}^1 2^{1-r} \|\rho_1 - \rho_2\|_1^r \, 2^{1-(1-r)} \big(\|\rho_1\|_1 + \|\rho_2\|_1 \big)^{1-r} \,\mathrm{d}r,
\end{aligned}$$

where we have used (21.34) and (21.35) for the last estimate. Calculating the integral in the last expression gives the desired estimate (21.32) and the proposition is proved. □

We emphasize that $\mathfrak{R} \ni \rho \mapsto \mathscr{C}_\rho \in \mathrm{Lin}(\mathrm{L}^\infty_{\mathrm{S}}(\boldsymbol{H}), \mathrm{L}^1_{\mathrm{S}}(\boldsymbol{H}))$ is continuous, if $\mathfrak{R}$ is equipped with the norm of $\mathrm{L}^1_{\mathrm{S}}(\boldsymbol{H})$. Moreover, assuming $\dim \boldsymbol{H} < \infty$ and writing $\operatorname{int} \mathfrak{R} = \{ \rho \mid \rho > 0 \}$ we see that $\operatorname{int} \mathfrak{R} \ni \rho \mapsto \mathscr{C}_\rho$ is an analytic function. For this we use that $R = \log \rho$ is an analytic function on $\operatorname{int} \mathfrak{R}$ and that $\mathscr{C}_\rho A = \int_0^1 \mathrm{e}^{sR} A \mathrm{e}^{(1-s)R} \,\mathrm{d}s$.

21.4.2 GENERIC Systems with Canonical Correlation

We now specialize the general system (21.26) by choosing all the operators $\boldsymbol{C}_m$ equal to $\mathscr{C}_\rho$, as suggested in [28]. However, we note that the dissipative bracket

in [28, Eqn. (10)] differs from the form assumed here. The point is now that the interaction of $\mathscr{C}_\rho$ with $\log\rho$ simplifies the evolutionary system considerably by invoking (21.31) giving

$$\mathbb{K}^Q_{\mathscr{C}_\rho} \log\rho = \big[Q,[Q,\rho]\big],$$

where the right-hand side is continuous, in contrast to $\log\rho$. We arrive at the system

$$\begin{aligned}
\dot\rho &= \mathrm{j}\big[H(z),\rho\big] - \sum_1^M \Big(k_{\mathrm{B}}\big[Q_m(z),[Q_m(z),\rho]\big] + \alpha_m(\rho,z)\cdot \mathrm{D}S(z)\mathbb{K}^{Q_m(z)}_{\mathscr{C}_\rho} H(z)\Big),\\
\dot z &= K_{\mathrm{diss}}(\rho,z)\mathrm{D}S(z)\\
&\quad + \sum_1^M \langle\!\langle\, k_{\mathrm{B}}\big[Q_m(z),[Q_m(z),\rho]\big] + \alpha_m(\rho,z)\cdot \mathrm{D}S(z)\mathbb{K}^{Q_m(z)}_{\mathscr{C}_\rho} H(z) \,\|\, H(z)\,\rangle\!\rangle \alpha_m(\rho,z).
\end{aligned} \tag{21.36}$$

The most important feature of this system is that the singular term $\log\rho$ in the right-hand side has disappeared. Under suitable further assumptions the right-hand side forms a continuous vector field, which wasn't the case for (21.26).

We simplify the above system even further by assuming that it consists of M heat baths with temperatures $\theta = (\theta_1,\dots,\theta_M)$ and heat capacities $c_1,\dots,c_M > 0$. Each heat bath can interact with the quantum system by a finite number of coupling operators Q^n_m, which are independent of θ. Writing shortly $c = (c_m)$ we have the following GENERIC system $(\mathscr{Q},\mathscr{E},\mathscr{S},\mathbb{J},\mathbb{K})$ with

$$\mathscr{Q} = \mathfrak{R}\times\,]0,\infty[^M, \quad q = (\rho,\theta), \tag{21.37a}$$

$$\mathscr{E}(q) = \mathrm{tr}(\rho H) + c\cdot\theta, \quad \mathscr{S}(q) = -k_{\mathrm{B}}\,\mathrm{tr}(\rho\log\rho) + \textstyle\sum_{m=1}^M c_m\log\theta_m, \tag{21.37b}$$

$$\mathbb{J}(q) = \begin{pmatrix} \mathrm{j}[\rho,\square] & 0\\ 0 & 0\end{pmatrix}, \quad \mathbb{K}(q) = \mathrm{D}^2\Psi^*_{(\xi,\tau)}(q;\xi,\tau), \tag{21.37c}$$

where the dual dissipation potential Ψ^* is given by

$$\Psi^*(\rho,\theta;\xi,\tau) = \tfrac12 \textstyle\sum_{m=1}^M \sum_{n=1}^{N_m} \|[Q^n_m, \xi - \frac{\tau_m}{c_m}H]\|^2_{\mathscr{C}_\rho} + \tfrac12 \tau\cdot\kappa\tau,$$

where $\kappa\in\mathbb{R}^{M\times M}_{\mathrm{sym}}$ is positive semidefinite with $\ker\kappa = \mathrm{span}\,c$. A typical choice for κ is given by $\tau\cdot\kappa\tau = \sum_{m=1}^{M-1}\sum_{l=m+1}^M \kappa_{ml}\big(\frac{\tau_m}{c_m} - \frac{\tau_l}{c_l}\big)^2$, where the positive coefficients κ_{ml} give the direct heat transfer between the heat baths m and l. We have the NIC (21.10) because (i) $\mathrm{D}\mathscr{E}(q) = (H,c)$ and $\Psi^*(q;(H,c)) \equiv 0$ imply $\mathbb{K}(q)\mathrm{D}\mathscr{E}(q)\equiv 0$ and (ii) $\mathrm{D}\mathscr{S}(q) = (-k_{\mathrm{B}}\log\rho,*)$ implies $\mathbb{J}(q)\mathrm{D}\mathscr{S}(q)\equiv 0$.

The differential equation generated by the GENERIC system $(\mathscr{Q}, \mathscr{E}, \mathscr{S}, \mathbb{J}, \mathbb{K})$ is

$$\dot\rho = \mathrm{j}[\rho, H] - \sum_{m=1}^{M}\sum_{n=1}^{N_m}\big[\, Q_m^n \,,\, k_{\mathrm{B}}[Q_m^n, \rho] + \frac{1}{\theta_m}\mathscr{C}_\rho[Q_m^n, H]\,\big], \tag{21.38a}$$

$$\dot\theta = \kappa\big(\tfrac{c_m}{\theta_m}\big)_m + \Big(\frac{1}{c_m}\sum_{n=1}^{N_m}\langle\!\langle\, k_{\mathrm{B}}[Q_m^n, \rho] + \tfrac{1}{\theta_m}\mathscr{C}_\rho[Q_m^n, H] \,\|\, [Q_m^n, H]\,\rangle\!\rangle\Big)_m. \tag{21.38b}$$

We nicely see the correction terms $\frac{1}{\theta_m}\big[Q_m^n, \mathscr{C}_\rho[Q_m^n, H]\big]$ to the otherwise linear Lindblad system $\dot\rho = \mathrm{j}[\rho, H] - k_{\mathrm{B}}\sum_{m=1}^{M}\sum_{n=1}^{N_m}\big[Q_m^n, [Q_m^n, \rho]\big]$.

If $\dim \boldsymbol{H} < \infty$ the right-hand side of (21.38) is analytic in the interior of $\mathscr{Q}$, i.e. in $\operatorname{int}\mathfrak{R}\times]0,\infty[^M$, while it is continuous on $\mathscr{Q}$. Hence, solutions starting in the interior are unique as long as they stay in the interior. In [21] the global existence of solutions for (21.38) will be established. The difficulties arise if solutions reach the boundary of $\mathscr{Q}$, where the extension is nontrivial and may be nonunique. Section 21.6 provides an existence result for the simplified model presented in Sect. 21.5.

21.4.3 Steady States

We finally discuss the steady states for (21.38). By strict convexity of $\mathscr{S}$ there is a unique maximizer $q_{\mathrm{eq}} = (\rho_{\mathrm{eq}}, \theta_{\mathrm{eq}}) \in \mathscr{Q}$ subject to the constraint $\mathscr{E}(q) = E_0$, as soon as $E_0 > \lambda_{\min}(H)$. The latter condition is needed to make the admissible set $\{\, q \in \mathscr{Q} \mid \mathscr{E}(q) = E_0 \,\}$ nonempty. The unique maximizer takes the form

$$\theta_{\mathrm{eq}}(E_0) = \theta_*(E_0)(1, \dots, 1) \quad \text{and} \quad \rho_{\mathrm{eq}} = \frac{1}{Z(E_0)}\exp\big(\frac{-1}{k_{\mathrm{B}}\theta_*(E_0)}H\big). \tag{21.39}$$

By Sect. 21.2.2 we know that all $q_{\mathrm{eq}}(E_0)$ are steady states of (21.38).

The following result provides conditions showing that the family $q_{\mathrm{eq}}(E_0)$ provides the only steady states $q = (\rho, \theta)$ of (21.38) satisfying $\rho > 0$.

Theorem 21.1. *Assume* $\operatorname{kernel}\kappa = \operatorname{span} c$ *and that* $(Q_m^n)_{\substack{m=1,\dots,M\\ n=1,\dots,N_m}}$ *and* H *satisfy:*

$$\left.\begin{array}{l} \text{If } A = A^* \in \mathrm{Lin}(\boldsymbol{H}),\ [H, A] = 0, \\ \text{and } \forall\, m, n:\ [Q_m^n, A] = 0, \end{array}\right\} \quad \text{then } A = \alpha I \text{ for some } \alpha \in \mathbb{R}. \tag{21.40}$$

Then, all steady states $q = (\rho, \theta) \in \mathscr{Q}$ *of* (21.38) *satisfying* $\rho > 0$ *have the form* $q_{\mathrm{eq}}(E_0)$ *for some* $E_0 \in \mathbb{R}$.

Proof. If q is a steady state with $q = (\rho, \theta)$ with $\rho > 0$, then $2\Psi^*(q, \mathrm{D}\mathscr{S}(q)) = \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{S}(q) = 0$. Hence, we have

$$0 = \kappa \mathrm{D}_\theta \mathscr{S}(q) = \kappa\big(\tfrac{c_m}{\theta_m}\big)_{m=1,\dots,M}, \quad 0 = \|[Q_m^n, -k_\mathrm{B}\log\rho - \tfrac{1}{\theta_m}H]\|_{\mathscr{C}_\rho}.$$

Using $\operatorname{kernel}\kappa = \operatorname{span} c$ we obtain $\theta = (\theta_0, \dots, \theta_0)$. Since $\rho > 0$ we have $\|A\|_{\mathscr{C}_\rho} = 0$ if and only if $A = 0$, whence

$$\forall\, m = 1, \dots, M\ \forall\, n = 1, \dots, N_m: \quad [Q_m^n, k_\mathrm{B}\log\rho + \tfrac{1}{\theta_0}H] = 0.$$

Inserting this into (21.38a) with $\dot\rho = 0$ (for a steady state) we also find $[\rho, H] = 0$. The latter implies $[H, k_\mathrm{B}\log\rho + \frac{1}{\theta_0}H]$, and we can apply (21.40) to $A = k_\mathrm{B}\log\rho + \frac{1}{\theta_0}H$. Now $k_\mathrm{B}\log\rho + \frac{1}{\theta_0}H = \alpha I$ implies the desired result, and the theorem is proved. □

In general it seems difficult to exclude steady states $q = (\rho, \theta)$ with $\det\rho = 0$. We refer to Section 21.5.3 for an example where (21.40) does not hold and where a whole segment of equilibria exists for each E_0.

Open problem 2. *Provide minimal assumptions on H, Q_m^n, and κ to guarantee that there are no steady states with* $\det\rho = 0$.

21.4.4 Comparison to the Lindblad Equation

Most often dissipative quantum mechanics is described in terms of the Lindblad equation (cf. [6, 15, 16]). In the isothermal case $\theta_m = \theta_*$, our equation (21.38a) takes the nonlinear form

$$\dot\rho = \mathrm{j}[\rho, H] - \sum_{m=1}^{M}\sum_{n=1}^{N_m}\big[Q_m^n, k_\mathrm{B}[Q_m^n, \rho] + \frac{1}{\theta_*}\mathscr{C}_\rho[Q_m^n, H]\big], \tag{21.41}$$

where now the free energy $\mathscr{F}_*(\rho) = k_\mathrm{B}\theta_*\langle\!\langle\, \rho \,\|\, \log\rho \,\rangle\!\rangle + \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle$ is a Liapunov function.

The corresponding linear Lindblad equation is obtained from (21.41) simply by replacing $\mathscr{C}_\rho$ by $\mathscr{C}_{\rho_\mathrm{eq}}$ where we take the $\rho_\mathrm{eq} = \frac{1}{Z}\exp\big(\frac{-1}{k_\mathrm{B}\theta_*}H\big)$. Using the commutator relation (21.31) we have $\mathscr{C}_{\rho_\mathrm{eq}}[Q_m^n, H] = -k_\mathrm{B}\theta[Q_m^n, \rho_\mathrm{eq}]$ and arrive at

$$\dot\rho = \mathrm{j}[\rho, H] - \sum_{m=1}^{M}\sum_{n=1}^{N_m}\big[\, Q_m^n\,,\, k_\mathrm{B}\,[Q_m^n\,,\, \rho - \rho_\mathrm{eq}\,]\,\big]. \tag{21.42}$$

The striking advantage of the Lindblad equation is its linearity. We will see in Sect. 21.6 that it also leaves $\mathfrak{R}$ invariant, i.e. it preserves the trace condition $\operatorname{tr}\rho = 1$, the symmetry $\rho = \rho^*$, and the positivity $\rho \geq 0$.

However, for the coupling to the exterior like heat baths it is less useful. Note that ρ_{eq} has to be known already. In general, the equilibrium ρ_{eq} will depend in a complicated and nonlinear way on the other variables, e.g. $\rho_{\mathrm{eq}}(\theta)$ in (21.38). More importantly, one has to model the influence of $\rho - \rho_{\mathrm{eq}}(\theta)$ on the other variables in a self-consistent way. In particular, energy conservation and nonnegative entropy production rates have to be guaranteed.

21.5 A Simple Coupled System

21.5.1 The Case of One Heat Bath

We simplify the above problem even further, by assuming that there is only one dissipative interaction term (i.e., $M = 1$) with one coupling operator Q (i.e., $N_1 = 1$). Moreover, the matrix κ disappears as $c > 0$ and $\kappa c = 0$. The system reduces to

$$\begin{aligned} \dot\rho &= \mathrm{j}\big[H,\rho\big] - \Big[Q, k_{\mathrm{B}}\big[Q,\rho\big] + \frac{1}{\theta}\mathscr{C}_\rho\big[Q,H\big]\Big], \\ \dot\theta &= \frac{1}{c}\langle\!\langle\, k_{\mathrm{B}}\big[Q,\rho\big] + \frac{1}{\theta}\mathscr{C}_\rho\big[Q,H\big] \parallel \big[Q,H\big] \,\rangle\!\rangle. \end{aligned} \tag{21.43}$$

Energy and entropy are given by

$$\mathscr{E}(\rho,\theta) = \langle\!\langle\, \rho \parallel H \,\rangle\!\rangle + c\theta \quad \text{and} \quad \mathscr{S}(\rho,\theta) = -k_{\mathrm{B}} \operatorname{tr}(\rho\log\rho) + c\log\theta$$

It is easy to check that $\mathscr{E}$ is conserved along solutions while $\mathscr{S}$ is nondecreasing. For this, set $\Phi = -k_{\mathrm{B}}\log\rho - \frac{1}{\theta}H$, where Φ can be seen as the driving force for irreversible processes as $\mathrm{D}\mathscr{S} - \frac{1}{\theta}\mathrm{D}\mathscr{E} = (\Phi, 0)^{\mathsf{T}}$, see [19]. With these abbreviations system (21.43) takes the form

$$\dot\rho = \mathrm{j}\big[H,\rho\big] + \mathbb{K}_\rho^Q\Phi, \quad \dot\theta = -\frac{1}{c}\langle\!\langle\, \mathbb{K}_\rho^Q\Phi \parallel H \,\rangle\!\rangle.$$

Hence, we easily find

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{E}(\rho(t),\theta(t)) &= \langle\!\langle\, \dot\rho \parallel H \,\rangle\!\rangle + c\dot\theta = 0 + \langle\!\langle\, \mathbb{K}_\rho^Q\Phi \parallel H \,\rangle\!\rangle - \langle\!\langle\, \mathbb{K}_\rho^Q\Phi \parallel H \,\rangle\!\rangle = 0, \\ \frac{\mathrm{d}}{\mathrm{d}t}\mathscr{S}(\rho(t),\theta(t)) &= \langle\!\langle\, \dot\rho \parallel -k_{\mathrm{B}}\log\rho \,\rangle\!\rangle + c\frac{\dot\theta}{\theta} \end{aligned}$$

$$= 0 + \langle\!\langle\, \mathbb{K}_\rho^Q \Phi \,\|\, -k_{\mathrm{B}} \log \rho \,\rangle\!\rangle - \frac{1}{\theta} \langle\!\langle\, \mathbb{K}_\rho^Q \Phi \,\|\, H \,\rangle\!\rangle$$
$$= \langle\!\langle\, \mathbb{K}_\rho^Q \Phi \,\|\, \Phi \,\rangle\!\rangle = \langle\!\langle\, \mathscr{C}_\rho\big[Q, \Phi\big] \,\|\, \big[Q, \Phi\big] \,\rangle\!\rangle \geq 0.$$

21.5.2 Elimination of the Temperature

Since $\theta > 0$ is only a scalar, we may eliminate it by using the invariance of the energy by assuming $\mathscr{E}(\rho, \theta) = E_0$. Then, $\theta = (E_0 - \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle)/c$ and we can reduce system (21.43) to the single equation for ρ:

$$\dot\rho = \mathrm{j}\big[H, \rho\big] - \Big[Q, k_{\mathrm{B}}\big[Q, \rho\big] + \frac{c}{E_0 - \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle} \mathscr{C}_\rho\big[Q, H\big]\Big], \tag{21.44a}$$

which has the negative entropy $-\tilde{\mathscr{S}}$ as a Liapunov functional, where

$$\tilde{\mathscr{S}}(\rho) = -k_{\mathrm{B}} \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle + c \log\big(E_0 - \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle\big). \tag{21.44b}$$

Note that $\tilde{\mathscr{S}}$ is still a strictly concave function on the compact set $\mathfrak{R}$. Hence, assuming that E_0 is given such that $\tilde{\mathscr{S}}$ is finite at least at one point in $\mathfrak{R}$, then there is a unique maximizer ρ_* given via

$$\theta_* = (E_0 - \langle\!\langle\, \rho_* \,\|\, H \,\rangle\!\rangle)/c > 0 \quad \text{and} \quad \rho_* = \frac{1}{Z} \exp\big(-\tfrac{1}{k_{\mathrm{B}}\theta_*} H\big). \tag{21.45}$$

In fact, we are now easily able to pass to the isothermal limit in the sense of Sect. 21.2.3. We assume that E_0 and c tend to ∞ while H and Q are fixed. Assuming $E_0/c \to \theta_* > 0$ we find the expansion

$$c \log\big(E_0 - \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle\big) = c \log E_0 - \frac{c}{E_0} \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle + O(c/E_0^2),$$

where we use that $\rho \in \mathfrak{R}$ is bounded. Thus, we find the free energy

$$\mathscr{F}_*(\rho) = \lim_{\substack{c, E_0 \to \infty \\ E_0/c \to \theta_*}} \Big(E_0 \log E_0 - \theta_* \tilde{\mathscr{S}}(\rho)\Big) = \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle + k_{\mathrm{B}} \theta_* \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle, \tag{21.46a}$$

and the simplified isothermal evolutionary system

$$\dot\rho = \mathrm{j}\big[H, \rho\big] - \Big[Q, k_{\mathrm{B}}\big[Q, \rho\big] + \frac{1}{\theta_*} \mathscr{C}_\rho\big[Q, H\big]\Big]. \tag{21.46b}$$

21.5.3 *The Case* $\dim \boldsymbol{H} = 2$

In the case that the basic Hilbert space $\boldsymbol{H}$ is two-dimensional, i.e. $\boldsymbol{H} = \mathbb{C}^2$, we can write (21.46) explicitly by introducing suitable coordinates. While in the general case $\boldsymbol{H} = \mathbb{C}^n$ the set $\mathfrak{R}$ can be seen as a real (n^2-1)-dimensional manifold with piecewise smooth boundary, the case $n = 2$ is special, because $\mathfrak{R}$ has a smooth boundary and can be identified by the closed ball $\mathscr{A} = B_{1/2}(0) \subset \mathbb{R}^3$ as follows:

$$\mathfrak{R} = \{\, \rho = \hat{\rho}(\boldsymbol{a}) \mid \boldsymbol{a} \in \mathscr{A} \,\}, \quad \text{where } \hat{\rho}(\alpha,\beta,\gamma) = \begin{pmatrix} \frac{1}{2}+\alpha & \beta+\mathrm{i}\gamma \\ \beta-\mathrm{i}\gamma & \frac{1}{2}-\alpha \end{pmatrix}.$$

Using $z = \beta + \mathrm{i}\gamma$ and $r = |\boldsymbol{a}|$ the eigenvalues $r_\pm$ and eigenvectors $\psi_\pm$ of $\hat{\rho}(\boldsymbol{a})$ are

$$r_\pm = \frac{1}{2} \pm r, \qquad \psi_\pm = \frac{1}{\sqrt{2r(r\mp\alpha)}} \begin{pmatrix} z \\ -\alpha \pm r \end{pmatrix}.$$

Thus, $\mathscr{C}_{\hat{\rho}(\boldsymbol{a})}$ can be written out explicitly using (21.28). In particular, choosing H and Q we are able to write (21.44) and (21.46b) as explicit ODE systems in the three variables (α,β,γ). For didactical purposes we will do this for the special case

$$H = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} 0 & q \\ q & 0 \end{pmatrix},$$

where $h_1, h_2, q \in \mathbb{R}$. In treating this example, we will see some of the difficulties in proofing the existence of solutions and the convergence of all solutions into the unique thermodynamic equilibrium ρ_*. We first note that the free energy $\mathscr{F}_*$ has the form $\hat{\mathscr{F}}(\boldsymbol{a}) := \mathscr{F}_*(\hat{\rho}(\boldsymbol{a}))$ with

$$\hat{\mathscr{F}}(\boldsymbol{a}) = k_{\mathrm{B}}\theta_* \big((\tfrac{1}{2}+r)\log(\tfrac{1}{2}+r) + (\tfrac{1}{2}-r)\log(\tfrac{1}{2}-r) \big) + \tfrac{1}{2}(h_1+h_2) - (h_2-h_1)\alpha.$$

Some lengthy calculations, using the explicit form of $r_\pm$, $\psi_\pm$ and (21.28), give

$$\mathscr{C}_{\hat{\rho}(\boldsymbol{a})} \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = \begin{pmatrix} -\mathrm{i}\gamma\big(1 + \frac{1-2\lambda}{2r^2}\alpha\big) & \lambda + \frac{1-2\lambda}{2r^2}\gamma^2 - \mathrm{i}\frac{1-2\lambda}{2r^2}\beta \\ -\lambda - \frac{1-2\lambda}{2r^2}\gamma^2 - \mathrm{i}\frac{1-2\lambda}{2r^2}\beta & \mathrm{i}\gamma\big(\frac{1-2\lambda}{2r^2}\alpha - 1\big) \end{pmatrix},$$

where $\lambda = \lambda(r) = \Lambda(\frac{1}{2}+r, \frac{1}{2}-r)$. Using $\lambda(0) = 1/2$ we see that $r \mapsto (\lambda(r), \frac{1-2\lambda(r)}{2r^2})$ is analytic on $[0,1/2[$ and continuous on $[0,1/2]$. Thus, we see that $\boldsymbol{a} \mapsto \mathscr{C}_{\hat{\rho}(\boldsymbol{a})} \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$ is smooth in the interior of $\mathscr{A}$ and continuous on the closed ball $\mathscr{A}$, as predicted by Proposition 21.2. Inserting this into (21.46b) leads to the system

$$\begin{pmatrix} \dot{\alpha} \\ \dot{\beta} \\ \dot{\gamma} \end{pmatrix} = \begin{pmatrix} 0 \\ (h_2-h_1)\gamma \\ (h_1-h_2)\beta \end{pmatrix} - k_{\mathrm{B}}q^2 \begin{pmatrix} 4\alpha \\ 0 \\ 4\gamma \end{pmatrix} + \frac{q^2(h_2-h_1)}{\theta_*} \begin{pmatrix} 2\lambda + \frac{1-2\lambda}{r^2}\gamma^2 \\ 0 \\ -\frac{1-2\lambda}{r^2}\alpha\gamma \end{pmatrix}. \tag{21.47}$$

It is also instructive to see the isothermal form of the GENERIC system as given in (21.15). Using the coordinates $\boldsymbol{a} = (\alpha, \beta, \gamma)^{\mathsf{T}} \in \mathscr{A}$ and $r = |\boldsymbol{a}|$ we have

$$\mathrm{D}\hat{\mathscr{F}}(a) = \frac{2k_{\mathrm{B}}\theta_*}{\lambda(r)} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} - \begin{pmatrix} h_2 - h_1 \\ 0 \\ 0 \end{pmatrix}, \quad \hat{J}(\boldsymbol{a}) = \begin{pmatrix} 0 & \gamma & -\beta \\ -\gamma & 0 & \alpha \\ \beta & -\alpha & 0 \end{pmatrix},$$

$$\hat{K}(\boldsymbol{a}) = q^2 \begin{pmatrix} 2\lambda(r) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2\lambda(r) \end{pmatrix} + \frac{q^2(1-2\lambda(r))}{r^2} \begin{pmatrix} \gamma^2 & 0 & -\alpha\gamma \\ 0 & 0 & 0 \\ -\alpha\gamma & 0 & \alpha^2 \end{pmatrix}.$$

It is now easy to see that (21.47) is given in the form $\dot{\boldsymbol{a}} = \big(\hat{J}(\boldsymbol{a}) - \frac{1}{\theta_*}\hat{K}(\boldsymbol{a})\big)\mathrm{D}\hat{\mathscr{F}}(\boldsymbol{a})$. Moreover, we see that the Poisson structure $\hat{J}$ on $\mathscr{A}$ is the classical Lie–Poisson structure on $\mathbb{R}^3$ used for Euler's equation for rigid bodies, cf. [1, 18]. Moreover, we see that there is no dissipation in the β-component, since its direction is parallel to Q and, thus, vanishes in the commutator.

In particular, we see that all solutions starting in $\mathscr{A}$ remain there. In fact,

$$\tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}r^2 = \tfrac{1}{2}\tfrac{\mathrm{d}}{\mathrm{d}t}|\boldsymbol{a}|^2 = \boldsymbol{a}\cdot\dot{\boldsymbol{a}} = -4k_{\mathrm{B}}q^2(\alpha^2+\gamma^2) + \tfrac{2q^2}{\theta_*}(h_2-h_1)\alpha\lambda(|\boldsymbol{a}|) \quad (21.48)$$

implies that for $\boldsymbol{a} \in \partial\mathscr{A}$ (i.e., $r = |\boldsymbol{a}| = 1/2$) we have $\dot{r} \le 0$ because of $\lambda(1/2) = 0$. Moreover, solutions immediately move into the interior of $\mathscr{A}$ except when starting in $\boldsymbol{a}(0) = (0, \pm 1/2, 0)^{\mathsf{T}}$, where $\dot{\boldsymbol{a}} = \pm(0, 0, (h_1-h_2)/2)^{\mathsf{T}}$. Hence, except for the case $h_1 = h_2$ also in this case the solutions leave the boundary of $\mathscr{A}$.

The following general result on the dynamics of the simple ODE system (21.47) is an immediate consequence of the above derivations.

Theorem 21.2. *System (21.47) has for each initial condition $\boldsymbol{a}(0) \in \mathscr{A}$ a global solution, along which $\hat{\mathscr{F}}$ is nonincreasing. Solutions with $|\boldsymbol{a}(0)| < 1/2$ are unique and never touch the boundary $\partial\mathscr{A}$.*

In the purely Hamiltonian case $q = 0$ the solution are periodic with $\alpha(t) = \alpha(0)$ and $r(t) = r(0)$, i.e. the free energy $\hat{\mathscr{F}}(\boldsymbol{a}(t))$ is constant.

For $q \neq 0$ all solutions $\boldsymbol{a}(t)$ converge to a steady state for $t \to \infty$. If $h_1 \neq h_2$, this is the unique steady state $\boldsymbol{a}_*$ corresponding to $\rho_* = \hat{\rho}(\boldsymbol{a}_*)$ given in (21.45) and minimizing $\hat{\mathscr{F}}$.

If $h_1 = h_2$, then $\beta(t) = \beta(0)$ while $(\alpha(t), \gamma(t)) = \mathrm{e}^{-4k_{\mathrm{B}}q^2 t}(\alpha(0), \gamma(0))$.

Finally, we compare the solutions of the GENERIC system (21.47) with the corresponding linear Lindblad system, where the nonlinear term $[Q, \mathscr{C}_\rho[Q, H]]$ is replaced by the constant term $[Q, \mathscr{C}_{\rho_{\mathrm{eq}}}[Q, H]] = -k_{\mathrm{B}}\theta_*[Q, [Q, \rho_{\mathrm{eq}}]] = \text{const.}$ in (21.46b), namely

$$\begin{pmatrix} \dot{\alpha} \\ \dot{\beta} \\ \dot{\gamma} \end{pmatrix} = \begin{pmatrix} 0 \\ (h_2-h_1)\gamma \\ (h_1-h_2)\beta \end{pmatrix} - 4k_{\mathrm{B}}q^2 \begin{pmatrix} \alpha - \alpha_{\mathrm{eq}} \\ 0 \\ \gamma \end{pmatrix}, \quad (21.49)$$

where $\alpha_{\rm eq}$ depends on θ_* and h_2-h_1. Note that the dynamics of α and (β,γ) are mutually uncoupled, which is certainly not the case in (21.47).

21.6 Existence and Convergence into Equilibrium

Here we restrict ourselves to the finite-dimensional setting $2 \le \dim \boldsymbol{H} < \infty$. To keep notations simple we study the case without temperature, i.e. we either consider the isothermal case (21.46) or the full case (21.44) for a given value of E_0, see Sect. 21.5.2. In both cases the driving functional (now being a Liapunov functional) has the form

$$\mathscr{F}(\rho) = e\,\langle\!\langle\, \rho \,\|\, \log\rho \,\rangle\!\rangle + \Upsilon(\langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle),$$

where $e>0$ and $\Upsilon:\mathbb{R}\to\mathbb{R}\cup\{\infty\}$ is continuous convex function. Our state space is the compact set $\mathfrak{R}$ and the equation we study is

$$\begin{aligned}\dot\rho = \mathscr{V}(\rho) \;&:=\; \mathrm{j}[H,\rho] - \sum_{n=1}^{N}\big[Q^n, \mathscr{C}_\rho[Q^n, \mathrm{D}\mathscr{F}(\rho)]\big]\\ &=\; \mathrm{j}[H,\rho] - e\sum_{n=1}^{N}\big[Q^n,[Q^n,\rho]\big] - \Upsilon'(\langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle)\sum_{n=1}^{N}\big[Q^n,\mathscr{C}_\rho[Q^n,H]\big].\end{aligned} \tag{21.50}$$

21.6.1 Existence via a Modified Explicit Euler Scheme

The construction of solutions uses an explicit Euler scheme with a slight modification to keep the property $\rho\in\mathfrak{R}$. Consider the time interval $[0,t_0]$, choose an integer $k>1$, and define the time step $\delta = t_0/k$. For a given initial condition q_0 we define incrementally

$$\rho^{j+1} = \mathscr{P}_{\mathfrak{R}}\big(\rho^j + \delta\mathscr{V}(\rho^j)\big), \qquad j=0,\dots,k-1. \tag{21.51}$$

Here we introduced the projection operator $\mathscr{P}_{\mathfrak{R}}$ that maps arbitrary matrices with trace 1 to elements in $\mathfrak{R}$ in the following way

$$\mathscr{P}_{\mathfrak{R}}\Big(\sum_{i=1}^{\dim\boldsymbol{H}} r_i\psi_i\otimes\overline{\psi}_i\Big) = \frac{1}{\sum_{j=1}^{\dim\boldsymbol{H}}\max\{0,r_j\}}\sum_{i=1}^{\dim\boldsymbol{H}}\max\{0,r_i\}\psi_i\otimes\overline{\psi}_i.$$

The necessary properties of $\mathscr{P}_{\mathfrak{R}}$ are given in the following lemma.

Lemma 21.1. *The nonlinear projector $\mathscr{P}_{\mathfrak{R}}$ satisfies:*

(i) $\exists\, C>0\ \forall\, \rho_1,\rho_2\in\mathscr{L}^1_{\mathrm{S}}(\boldsymbol{H})$ *with* $\operatorname{tr}\rho_j=1$*:*
$\operatorname{dist}(\rho_j,\mathfrak{R})\le 1/2 \implies \|\mathscr{P}_{\mathfrak{R}}(\rho_1)-\mathscr{P}_{\mathfrak{R}}(\rho_2)\|\le C\|\rho_1-\rho_2\|$.

(ii) If $\mathfrak{R}\ni\rho^k\to\rho\in\mathfrak{R}$ *and* $\delta^k\searrow 0$*, then* $\frac{1}{\delta_k}\big(\mathscr{P}_{\mathfrak{R}}\big(\rho^k+\delta_k\mathscr{V}(\rho^k)\big)-\rho^k\big)\to\mathscr{V}(\rho)$.

Proof. Assertion (i) is clear by the classical Lipschitz continuity of orthogonal projections. For assertion (ii) we distinguish the cases $\rho>0$ and $\rho\in\partial\mathfrak{R}$. In the first case convergence (ii) follows easily by the continuity of $\mathscr{V}$, because for sufficiently large k the projection $\mathscr{P}_{\mathfrak{R}}$ acts as identity.

Now assume $\rho=\left(\begin{smallmatrix}\sigma&0\\0&0\end{smallmatrix}\right)$ with $\sigma>0$. In particular, we split the space $\boldsymbol{H}=\boldsymbol{H}_1\oplus\boldsymbol{H}_2$ and write all operators as 2×2-block operators with respect to this splitting. We find

$$\mathscr{C}_\rho\begin{pmatrix}A_{11}&A_{12}\\A_{21}&A_{22}\end{pmatrix}=\begin{pmatrix}\mathscr{C}_\sigma A_{11}&0\\0&0\end{pmatrix} \tag{21.52}$$

and conclude that $\mathscr{V}(\rho)=\left(\begin{smallmatrix}\alpha&\beta\\\beta^*&0\end{smallmatrix}\right)$. To show (ii) we use the decomposition

$$\begin{aligned}&\tfrac{1}{\delta_k}\Big(\mathscr{P}_{\mathfrak{R}}\big(\rho^k+\delta_k\mathscr{V}(\rho^k)\big)-\rho^k\Big)=V_k^1+V_k^2+\mathscr{V}(\rho),\\&\text{where } V_k^1:=\tfrac{1}{\delta_k}\Big(\mathscr{P}_{\mathfrak{R}}\big(\rho^k+\delta_k\mathscr{V}(\rho^k)\big)-\mathscr{P}_{\mathfrak{R}}\big(\rho^k+\delta_k\mathscr{V}(\rho)\big)\Big)\\&\text{and}\quad V_k^2:=\tfrac{1}{\delta_k}\Big(\mathscr{P}_{\mathfrak{R}}\big(\rho^k+\delta_k\mathscr{V}(\rho)\big)-\big(\rho^k+\delta_k\mathscr{V}(\rho)\big)\Big),\end{aligned}$$

which means that we have to show $V_k^1\to 0$ and $V_k^2\to 0$. By the Lipschitz continuity (i) and the continuity of $\mathscr{V}$ on $\mathfrak{R}$ we have $\|V_k^1\|\le C\|\mathscr{V}(\rho^k)-\mathscr{V}(\rho)\|\to 0$.

For the second term we establish the decomposition

$$\rho_k:=\rho^k+\delta_k\mathscr{V}(\rho)=\tilde\rho_k+\delta_k W_k\quad\text{with }\tilde\rho_k\in\mathfrak{R}\text{ and }W_k\to 0. \tag{21.53}$$

Then, $\mathscr{P}_{\mathfrak{R}}(\tilde\rho_k)=\tilde\rho_k$ implies the estimate

$$\|V_k^2\|=\tfrac{1}{\delta_k}\|\mathscr{P}_{\mathfrak{R}}(\rho_k)-\rho_k\|\le\tfrac{1}{\delta_k}\|\mathscr{P}_{\mathfrak{R}}(\rho_k)-\mathscr{P}_{\mathfrak{R}}(\tilde\rho_k)\|+\tfrac{1}{\delta_k}\|\tilde\rho_k-\rho_k\|\le(C+1)\|W_k\|\to 0,$$

which establishes the convergence in (ii).

It remains to show (21.53). We use the notation

$$\rho^k=\begin{pmatrix}\sigma_k&b_k\\b_k^*&c_k\end{pmatrix}\text{ and }\rho_k=\rho^k+\delta_k\mathscr{V}(\rho)=\begin{pmatrix}\Sigma_k&B_k\\B_k^*&c_k\end{pmatrix},$$

i.e. $\Sigma_k = \sigma_k + \delta_k\alpha$ and $B_k = b_k + \delta_k\beta$. As an intermediate decomposition we let

$$\rho_k = \hat{\rho}_k + \begin{pmatrix} 0 & 0 \\ 0 & \delta_k\gamma_k \end{pmatrix}, \text{ where } \gamma_k = \tfrac{1}{\delta_k}\big(b_k^*\sigma_k^{-1}b_k - B_k^*\Sigma_k^{-1}B_k\big).$$

Here $\hat{\rho}_k$ is positive semidefinite because of

$$\hat{\rho}_k = \begin{pmatrix} \Sigma_k & B_k \\ B_k^* & B_k^*\Sigma_k^{-1}B_k \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & c_k - b_k^*\sigma_k^{-1}b_k \end{pmatrix},$$

where the first and second operator are positive semidefinite because of $\sigma \approx \Sigma_k > 0$ and $\rho^k \geq 0$, respectively. Using $\operatorname{tr}\hat{\rho}_k = 1 - \delta_k \operatorname{tr}\gamma_k \approx 1$, we now let

$$\rho_k = \tilde{\rho}_k + \delta_k W_k \quad \text{with } \tilde{\rho}_k = \tfrac{1}{1-\delta_k \operatorname{tr}\gamma_k}\,\hat{\rho}_k \text{ and } W_k = \begin{pmatrix} 0 & 0 \\ 0 & \gamma_k \end{pmatrix} - (\operatorname{tr}\gamma_k)\,\tilde{\rho}_k.$$

By construction we have $\rho \approx \tilde{\rho}_k \in \mathfrak{R}$ and it remains to be shown that $\gamma_k \to 0$. Using $\rho^k \to 0$ we have $b_k \to 0$ and, hence, obtain

$$\|\gamma_k\| \leq \tfrac{C_1}{\delta_k}\big(\|B_k - b_k\|(\|b_k\| + \|B_k\|) + \|\sigma_k - \Sigma_k\|(\|B_k\| + \|b_k\|)^2\big) \leq C_2(\|b_k\| + \delta_k) \to 0.$$

This completes the proof of the lemma. □

With this lemma at hand, we obtain our global existence result. In contrast to the subsequent convergence result, we don't need any specific properties of H, $Q^1, ..Q^N$. In [21] we will show that the existence result can be extended to more general systems containing several heat baths, e.g. (21.38).

Theorem 21.3. *Consider e, Υ, H, and arbitrary coupling operators Q^n as defined at the beginning of this section. Then, for all initial conditions $\rho^0 \in \mathfrak{R}$ there is a global solution $\rho \in \mathrm{C}^1([0,\infty[\,;\mathfrak{R})$ of (21.50) with $\rho(0) = \rho^0$.*

Proof. The proof follows easily using the incremental approach based on the modified explicit Euler scheme (21.51). Since the vector field $\mathscr{V}$ is continuous on the compact set $\mathfrak{R}$ it is bounded. For a given time increment $\delta = t_0/k > 0$ we obtain the incremental approximations ρ_δ^j via (21.51) and form two interpolants, namely $\hat{\rho}_\delta : [0,\infty[\to \mathfrak{R}$ and $\overline{\rho}_\delta : [0,\infty[\to \mathfrak{R}$, where $\hat{\rho}_\delta$ is continuous and piecewise affine, whereas $\overline{\rho}_\delta$ is piecewise constant and continuous from the right, i.e. $\overline{\rho}_\delta(t) = \rho_\delta^j$ for $j\delta \leq t < (j+1)\delta$. These interpolants satisfy

$$\frac{\mathrm{d}}{\mathrm{d}t}\hat{\rho}_\delta(t) = \frac{1}{\delta}\Big(\mathscr{P}_{\mathfrak{R}}\big(\overline{\rho}_\delta + \delta\mathscr{V}_{\mathfrak{R}}(\overline{\rho}_\delta)\big) - \overline{\rho}_\delta\Big) \quad \text{for all } t \in [0,t_0] \setminus \{\,\delta j \mid j \in \mathbb{N}\,\}. \tag{21.54}$$

Moreover, the sequence $\hat{\rho}_\delta$ is uniformly Lipschitz continuous and, thus, admits a uniformly converging subsequence with $\delta = \delta_k \to 0$ and a limit $\rho : [0,t_0] \to \mathscr{R}$.

Moreover, we may assume $\frac{\mathrm{d}}{\mathrm{d}t}\hat{\rho}_\delta \overset{*}{\rightharpoonup} w$ in $\mathrm{L}^\infty([0,t_0];\mathrm{L}^2_{\mathrm{S}}(\boldsymbol{H}))$. Using Lemma 21.1 and the uniform convergence of ρ_δ to ρ, the right-hand side in (21.54) converges uniformly to $\mathscr{V}(\rho(t))$. Standard ODE arguments show that $\rho \in \mathrm{C}^1([0,t_0];\mathrm{L}^2_{\mathrm{S}}(\boldsymbol{H}))$ with $\dot\rho = w$, such that $\dot\rho = \mathscr{V}(\rho)$. Hence, ρ is a solution of (21.50). Since t_0 was arbitrary global existence follows, and the theorem is proved. □

Open problem 3. *Do we have uniqueness if every solution leaves the boundary of* $\mathfrak{R}$ *immediately? What are the conditions on* H *and* Q^n *to guarantee this property?*

21.6.2 Convergence into the Thermodynamic Equilibrium

One problem in our system is that the functional $\mathscr{F}$ (or more general the entropy $\mathscr{S}(\rho,\theta)$) is not differentiable on the boundary $\partial\mathfrak{R}$ of $\mathfrak{R}$. Thus, it is not clear how to show that $\mathscr{F}$ is a strict Liapunov function along solutions staying on $\partial\mathfrak{R}$. Thus, we will first provide some general conditions on the coupling operators Q^n such that $\partial\mathfrak{R}$ is transversally repelling, i.e. each solution leaves the boundary to the inside with a positive speed.

For a family $\boldsymbol{Q} = (Q^1,\dots,Q^N) \in \mathrm{L}^2_{\mathrm{S}}(\boldsymbol{H})^N$ we define

$$\mu(\boldsymbol{Q}) := \inf\{\, f_{\boldsymbol{Q}}(\psi,\phi) \mid |\psi| = |\phi| = 1,\ \langle\psi|\phi\rangle = 0 \,\} \text{ where } f_{\boldsymbol{Q}}(\psi,\phi) = \sum_{n=1}^{N} |\langle Q^n\psi|\phi\rangle|^2.$$

Clearly, we have $\mu(\boldsymbol{Q}) = 0$ if $\boldsymbol{Q} = (Q)$ and $\dim \boldsymbol{H} \ge 2$, since then we may choose two orthogonal eigenvectors of Q. However, for $\boldsymbol{Q} = (Q^1,Q^2)$ with

$$Q^1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \text{ and } Q^2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

we find $\mu(\boldsymbol{Q}) > 0$.

Open problem 4. *What is the minimal number* N *for a given* $\dim \boldsymbol{H}$ *such that there exists* $\boldsymbol{Q} = (Q^1,\dots,Q^N)$ *with* $\mu(\boldsymbol{Q}) > 0$ *?*

Proposition 21.3. *Assume that* (21.50) *satisfies additionally* $\mu(\boldsymbol{Q}) > 0$*, then all solutions* $\rho : [0,\infty[\to \mathfrak{R}$ *with* $\rho(0) \in \operatorname{int}\mathfrak{R}$ *satisfy* $\inf\{\operatorname{dist}(\rho(t),\partial\mathfrak{R}) \mid t \ge 0\} > 0$*. Moreover, there exist* $t_0 > 0$ *such that for each solutions with* $\rho(0) \in \partial\mathfrak{R}$ *we have* $\operatorname{dist}(\rho(t),\partial\mathfrak{R}) \ge e\mu(\boldsymbol{Q})t > 0$ *for* $t \in]0,t_0]$.

Proof. Since $\rho = \sum_{k=1}^{\dim \boldsymbol{H}} r_k\psi_k \otimes \overline{\psi}_k$ we have $\rho \in \partial\mathfrak{R}$ if and only if $\det\rho = \prod_{k=1}^{\dim \boldsymbol{H}} r_k = 0$. This implies $\operatorname{dist}(\rho,\partial\mathfrak{R}) \ge \min\{\, r_k \mid k = 1,\dots,\dim \boldsymbol{H}\,\}$. Moreover, from the equation $\dot\rho = \mathscr{V}(\rho)$ and $\rho(t) = \sum_{k=1}^{\dim \boldsymbol{H}} r_k(t)\psi_k(t) \otimes \overline{\psi}_k(t)$ we obtain easily the relation

$$\dot r_k(t) = \langle\!\langle\, \mathscr{V}(\rho(t)) \,\|\, \psi_k(t) \otimes \overline{\psi}_k(t) \,\rangle\!\rangle = \langle \mathscr{V}(\rho(t))\psi_k(t)|\psi_k(t)\rangle.$$

Now assume $\rho(t) \in \partial\mathfrak{R}$, i.e. that there exists k with $r_k(t) = 0$. For the three terms of $\mathscr{V}$ in (21.50) we find $\langle\!\langle\, [\rho, H] \,\|\, \psi_k(t) \otimes \overline{\psi}_k(t) \,\rangle\!\rangle = \langle\!\langle\, H \,\|\, [\rho, \psi_k(t) \otimes \overline{\psi}_k(t)] \,\rangle\!\rangle = 0$ by using (21.18). Invoking the block structure (21.52) with $\boldsymbol{H}_2 = \operatorname{span} \psi_k$ we obtain $(\mathscr{C}_\rho A)\psi_k(t) \otimes \overline{\psi}_k(t) = \psi_k(t) \otimes \overline{\psi}_k(t)(\mathscr{C}_\rho A) = 0$ for all A, and we conclude

$$\langle\!\langle\, [Q, \mathscr{C}_\rho B] \,\|\, \psi_k(t) \otimes \overline{\psi}_k(t) \,\rangle\!\rangle = \langle\!\langle\, [\mathscr{C}_\rho B, \psi_k(t) \otimes \overline{\psi}_k(t)] \,\|\, Q \,\rangle\!\rangle = \langle\!\langle\, 0 - 0 \,\|\, Q \,\rangle\!\rangle = 0.$$

Thus, only the Lindblad terms in $\mathscr{V}(\rho)$ remain. Using the abbreviation $E_j = \psi_j(t) \otimes \overline{\psi}_j(t)$ and $r_k(t) = 0$ we obtain the relation

$$\dot{r}_k(t) = -e \sum_{n=1}^{N} \langle\!\langle\, [Q^n, [Q^n, \rho]] \,\|\, E_k \,\rangle\!\rangle = -e \sum_{j=1}^{\dim \boldsymbol{H}} r_j \sum_{n=1}^{N} \langle\!\langle\, [Q^n, [Q^n, E_j]] \,\|\, E_k \,\rangle\!\rangle$$

$$= 2e \sum_{j \neq k} r_j \sum_{n=1}^{N} |\langle Q^n \psi_j | \psi_k \rangle|^2 \geq 2e \sum_{j \neq k} r_j \, \mu(\boldsymbol{Q}) = 2e\mu(\boldsymbol{Q}),$$

where we used $\langle\!\langle\, [Q, [Q, E_j]] \,\|\, E_k \,\rangle\!\rangle = -2|\langle Q\Psi_j | \Psi_k \rangle|^2$ for $j \neq k$ and $r_k = 0$ to drop the term for $j = k$. By the assumptions $e > 0$ and $\mu(\boldsymbol{Q}) > 0$ and the compactness of $\partial\mathfrak{R}$, the continuity of $\mathscr{V}$ guarantees that $\dot{r}_k(t) \geq e\mu(\boldsymbol{Q}) > 0$ whenever $\operatorname{dist}(\rho(t), \partial\mathfrak{R}) \leq \delta$ for some $\delta > 0$. Now all the assertions of the proposition follow by standard ODE arguments. □

The above proposition shows that we only need to consider solutions lying in the interior of $\mathfrak{R}$, where $\mathscr{F}$ is smooth and where the dissipation relation

$$\frac{\mathrm{d}}{\mathrm{d}t} \mathscr{F}(\rho(t)) = -2\Psi^*(\rho, \mathrm{D}\mathscr{F}(\rho)) = -\sum_{n=1}^{N} \| \big[Q^n, e \log \rho + \Upsilon'(\langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle) H \big] \|^2_{\mathscr{C}_\rho}$$

holds. Hence, $\mathscr{F}$ is a Liapunov function but not necessarily a strict Liapunov function. The latter means that $\mathscr{F}$ is strictly decreasing along $t \mapsto \rho(t)$ whenever ρ is not constant (recall that we have uniqueness of solutions in the interior of $\mathfrak{R}$). In fact, (21.50) still allows for (linear) Hamiltonian dynamics in a subspace $\boldsymbol{H}_0$ if this subspace is left invariant by H and $Q^n|_{\boldsymbol{H}_0} = 0$ for all n.

The following result shows that a slight strengthening of the commutator condition (21.40), which was used to establish the uniqueness of steady states, is sufficient to show that $\mathscr{F}$ is a strict Liapunov function.

Theorem 21.4. *Assume that the system* (21.50) *satisfies* $\mu(\boldsymbol{Q}) > 0$ *and the strengthened commutator relation*

$$\Big(A \in \mathrm{L}^2_{\mathrm{S}}(\boldsymbol{H}) \text{ and } \forall\, n = 1, \ldots, N \colon [Q^n, A] = 0 \Big) \quad \Longrightarrow \quad \exists\, \alpha \in \mathbb{R} \colon A = \alpha I. \tag{21.55}$$

Then, all solutions $\rho : [0,\infty[\to \mathfrak{R}$ *of* (21.50) *satisfy* $\rho(t) \to \rho_{\mathrm{eq}}$*, where* ρ_{eq} *is the unique minimizer of* $\mathscr{F}$.

Proof. It remains to show that $\mathscr{F}$ is a strict Liapunov function. For this we have to investigate the set $\Psi^*(\rho, \mathrm{D}\mathscr{F}(\rho)) = 0$. Using $\mathscr{C}_\rho > 0$ for $\rho \in \operatorname{int} \mathscr{R}$ condition (21.55) gives $e \log \rho + \Upsilon'(..)H = \alpha I$. Thus, we have $\rho = c \exp(-\gamma H)$ with $e\gamma = \Upsilon'(\langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle)$, where $c > 0$ and $\gamma \in \mathbb{R}$ have to satisfy $1 = c \operatorname{tr} \exp(-\gamma H)$ and $e\gamma = \Upsilon'(c\langle\!\langle \exp(-\gamma H) \,\|\, H \,\rangle\!\rangle)$. Because of the monotonicity of Υ' (convexity of Υ) there is exactly one solution which defines ρ_{eq}. Now, the convergence to the unique steady state follows by the principle of Krasovskii and La Salle, cf. [9], and the theorem is proved. □

Open problem 5. *What are weaker conditions that guarantee that all solutions converge into some steady state? When is* $\mathscr{F}$ *a strict Liapunov function?*

21.7 Comparison to Stochastic Gradient Structures

In this section we restrict ourselves to purely dissipative systems, which do not have any Hamiltonian part JDE. Our aim is to highlight analogies between the dissipative evolution of the density operator ρ and certain other gradient flows arising in probabilistic systems. First, we compare to the Fokker–Planck equation with the Wasserstein gradient structure introduced in [11, 25]. Second, we show the analogy to the entropic gradient structure for reversible Markov chains introduced in [4, 17, 20].

In general, a gradient system consists of a basic space $\boldsymbol{X}$ with a differential structure, a differential potential $\mathscr{F} : \boldsymbol{X} \to \mathbb{R}$, and a metric $\mathscr{G}$ such that $\mathscr{G}(u) : \mathrm{T}_u\boldsymbol{X} \to \mathrm{T}_u^*\boldsymbol{X}$ is a linear, symmetric, and positive definite operator. As in the GENERIC framework we will use the inverse operator $\mathbb{K}(u) = \mathscr{G}(u)^{-1}$ and will denote the gradient system by $(\boldsymbol{X}, \mathscr{F}, \mathbb{K})$. The gradient flow is then defined by the evolutionary equation

$$0 = \mathscr{G}(u)\dot{u} + \mathrm{D}\mathscr{F}(u) \quad \Longleftrightarrow \quad \dot{u} = -\nabla_{\mathscr{G}}\mathscr{F}(u) =: -\mathbb{K}(u)\mathrm{D}\mathscr{F}(u).$$

Thus neglecting the Hamiltonian part $\mathrm{j}[\rho, H]$ in our dissipative quantum system (21.50) we have the gradient system $(\mathfrak{R}, \mathscr{F}_{\mathrm{qm}}, \mathbb{K}_{\mathrm{qm}}^{\boldsymbol{Q}})$ with

$$\mathscr{F}_{\mathrm{qm}}(\rho) := \langle\!\langle\, \rho \,\|\, \log \rho \,\rangle\!\rangle + \langle\!\langle\, \rho \,\|\, H \,\rangle\!\rangle \quad \text{and} \quad \mathbb{K}_{\mathrm{qm}}^{\boldsymbol{Q}}(\rho)\Xi := \sum_{n=1}^{N} \big[Q^n, \mathscr{C}_\rho[Q^n, \Xi]\big].$$

Hence, the operator $\mathbb{K}_{\mathrm{qm}}^{\boldsymbol{Q}}$ is associated with the dual dissipation potential $\Psi_{\mathrm{qm}}^*(\rho, \Xi) = \frac{1}{2}\sum_1^N \|[Q^n, \Xi]\|_{\mathscr{C}_\rho}^2$.

We now compare this with the Fokker–Planck equation

$$\dot{u} = \operatorname{div}(M(\nabla u + u\nabla H)), \quad t > 0,\ x \in \mathbb{R}^d,$$

where $M \in \mathbb{R}^{d\times d}$ is a symmetric and positive definite matrix and $H \in \mathrm{C}^2(\mathbb{R}^d)$ is a suitable potential. This equation can be understood as the gradient system $(\mathrm{L}^2(\mathbb{R}^d), \mathscr{F}_{\mathrm{FP}}, \mathbb{K}_{\mathrm{FP}})$ with

$$\mathscr{F}_{\mathrm{FP}}(u) := \int_{\mathbb{R}^d} u\log u + uH\,\mathrm{d}x \quad \text{and} \quad \mathbb{K}_{\mathrm{FP}}(u)\xi := -\sum_{i,j}\partial_{x_i}\big(uM_{ij}\partial_{x_j}\xi\big).$$

The analogy between $\mathscr{F}_{\mathrm{qm}}$ and $\mathscr{F}_{\mathrm{FP}}$ is obvious, whereas for $\mathbb{K}_{\mathrm{qm}}$ and $\mathbb{K}_{\mathrm{FP}}$ we see that the operators $\Xi \mapsto [Q^n, \Xi]$ are replaced by directional derivatives $\xi \mapsto \boldsymbol{q}\cdot\nabla\xi$. Moreover the multiplication factor $u \geq 0$, which is the core of the Wasserstein theory, is replaced by the canonical correlation operator $\mathscr{C}_\rho \geq 0$, which also is homogeneous of degree 1 in the state variable ρ, i.e. $\mathscr{C}_{\lambda\rho} = \lambda\mathscr{C}_\rho$.

Finally, we consider Markov chains on a finite number of sites, namely $\{1,\dots,N\}$. If p_n denotes the probability to be in site n, then the states $\boldsymbol{p} = (p_1,\dots,p_N)^{\mathsf{T}}$ lie in the state space $\boldsymbol{X}_N = \{\,\boldsymbol{p} \in [0,1]^N \mid \boldsymbol{p}\cdot\boldsymbol{e} = 1\,\}$, where $\boldsymbol{e} = (1,\dots,1)^{\mathsf{T}}$. The evolution is given in terms of the linear ODE

$$\dot{\boldsymbol{p}} = A\boldsymbol{p}, \quad \text{where } A \in \mathbb{R}^{N\times N} \text{ with } A_{ij} \geq 0 \text{ for } i \neq j \text{ and } A^{\mathsf{T}}\boldsymbol{e} = 0.$$

Here $A_{ij} \geq 0$ denotes the transition rate from j to i, and $A^{\mathsf{T}}\boldsymbol{e} = 0$ guarantees that $\boldsymbol{p}$ stays in $\boldsymbol{X}_N$.

We assume that the Markov chain $\dot{\boldsymbol{p}} = A\boldsymbol{p}$ is irreducible, which means that the kernel of A is one-dimensional such that there is a unique steady state $\boldsymbol{w} = (w_1,\dots,w_N)^{\mathsf{T}} \in \boldsymbol{X}_N$. An irreducible Markov chain is reversible (or is said to satisfy the "condition of detailed balance"), if

$$A_{nm}w_m = A_{mn}w_n \quad \text{for all } n,m \in \{1,\dots,N\}.$$

For such Markov chains $\dot{\boldsymbol{p}} = A\boldsymbol{p}$ it was shown in [4, 17, 20] that they can be understood as the gradient system $(\boldsymbol{X}_N, \mathscr{F}_{\mathrm{Mv}}, \mathbb{K}_{\mathrm{Mv}})$ with

$$\mathscr{F}_{\mathrm{Mv}}(\boldsymbol{p}) = \sum_{n=1}^N p_n\log(p_n/w_n) \quad \text{and}$$

$$\mathbb{K}_{\mathrm{Mv}}(\boldsymbol{p}) = \frac{1}{2}\sum_{n,m=1}^N A_{nm}w_m\Lambda\big(\tfrac{p_n}{w_n},\tfrac{p_m}{w_m}\big)\,(\boldsymbol{e}^n-\boldsymbol{e}^m)\otimes(\boldsymbol{e}^n-\boldsymbol{e}^m) \in \mathbb{R}^{N\times N}_{\geq 0},$$

where Λ is defined in (21.29) and $\boldsymbol{e}^n$ denotes the n-th unit vector in $\mathbb{R}^N$. Note that $\mathbb{K}_{\mathrm{Mv}}$ again is homogeneous of degree 1, namely $\mathbb{K}_{\mathrm{Mv}}(\lambda\boldsymbol{p}) = \lambda\mathbb{K}_{\mathrm{Mv}}(\boldsymbol{p})$.

In fact, one can see the Markov chain as the restriction of the quantum mechanical density functional theory to the case that ρ is a diagonal matrix, i.e. $\rho = \mathrm{diag}(\boldsymbol{p}) \in \mathfrak{R}$. Thus, we restrict the non-commutative operator theory in $\mathrm{L}^2_{\mathrm{S}}(\mathbb{C}^N)$ to the commutative case in $\boldsymbol{X}_N$. Note that the dual dissipation potential Ψ^*_{Mv} can be rewritten using the canonical correlation operator and commutators as follows:

$$\Psi^*_{\mathrm{Mv}}(\boldsymbol{p},\boldsymbol{\pi}) = \tfrac{1}{2}\boldsymbol{\pi}\cdot\mathbb{K}_{\mathrm{Mv}}(\boldsymbol{p})\boldsymbol{\pi} = \tfrac{1}{2}\sum_{n,m=1}^N \| [Q_{nm},\mathrm{diag}(\boldsymbol{\pi})] \|^2_{\mathscr{C}_{\hat\rho(\boldsymbol{p})}},$$

where $\hat\rho(\boldsymbol{p}) = \mathrm{diag}(p_n/w_n)$ and $Q_{nm} = (A_{nm}w_m)^{1/2}\frac{1}{2}(\boldsymbol{e}^n\otimes\boldsymbol{e}^m + \boldsymbol{e}^m\otimes\boldsymbol{e}^n)$. This form shows clearly the analogy between reversible Markov chains and dissipative quantum mechanics.

Open problem 6. *It would be interesting to find sets of commutators $(Q^n)_{n=1,..,N}$ such that the distance $d_{\mathbb{K}_{\mathrm{qm}}}$ can be characterized in more detail. In particular, one would be interested in an explicit characterization like for the Wasserstein distance $d_{\mathbb{K}_{\mathrm{FP}}}$. A particularly promising situation is the case discussed in [3].*

It would be interesting to see whether the optimal-transport formulation of [34] of the Schrödinger equation can be used to define suitable many-particle versions and to study their couplings to dissipative macroscopic systems.

Acknowledgements The author is grateful for helpful and stimulation discussions with Hans-Christian Öttinger and for Hagen Neidhardt's help in proving Proposition 21.2. He also thanks P.S. Krishnaprasad for discussion concerning $\mathscr{C}_\rho$ as a non-commutative generalization of the Fisher information. The author thanks an unknown referee for many helpful remarks. The research was partially supported by the European Research Council under *AnaMultiScale* ERC-2010-AdG 267802.

References

1. Abraham, R., Marsden, J.E.: Foundations of Mechanics. Benjamin/Cummings Publishing Co., Advanced Book Program, Reading, MA (1978). Second edition, revised and enlarged, With the assistance of Tudor Raţiu and Richard Cushman
2. Bhatia, R., Holbrook, J.A.R.: On the Clarkson-McCarthy inequalities. Math. Ann. **281**(1), 7–12 (1988)
3. Carlen, E.A., Maas, J.: An analog of the 2-Wasserstein metric in non-commutative probability under which the Fermionic Fokker–Planck equation is gradient flow for the entropy. arXiv 1203.5377, to appear in Commun. Math. Physics (2013)
4. Erbar, M., Maas, J.: Ricci curvature of finite Markov chains via convexity of the entropy. arXiv: 1111.2687 (2011)
5. Gohberg, I.C., Krein, M.G.: Introduction to the Theory of Linear Nonself-adjoint Operators. AMS Transl. Mathem. Monographs (1969)
6. Grabert, H.: Nonlinear relaxation and fluctuations of damped quantum systems. Z. Phys. B **49**(2), 161–172 (1982)
7. Grmela, M.: Why GENERIC? J. Non-Newtonian Fluid Mech. **165**, 980–986 (2010)
8. Grmela, M., Öttinger, H.C.: Dynamics and thermodynamics of complex fluids. I. Development of a general formalism. Phys. Rev. E (3) **56**(6), 6620–6632 (1997)

9. Hirsch, M.W., Smale, S.: Differential Equations, Dynamical Systems, and Linear Algebra. Academic Press, New York/London (1974)
10. Jelić, A., Hütter, M., Öttinger, H.C.: Dissipative electromagnetism from a nonequilibrium thermodynamics perspective. Phys. Rev. E **74**, 041126, 8 pp (2006)
11. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker-Planck equation. SIAM J. Math. Anal. **29**(1), 1–17 (1998)
12. Kubo, R., Toda, M., Hashitsume, N.: Nonequilibrium Statistical Mechanics, 2nd edn. Springer, New York (1991)
13. Kubo, R.: The fluctuation-dissipation theorem. Part I. Rep. Prog. Phys. **29**, 255–284 (1966)
14. Liero, M., Mielke, A.: Gradient structures and geodesic convexity for reaction-diffusion systems. Phil. Trans. Royal Soc. A (2013) To appear. WIAS preprint 1701 (April 2012)
15. Lindblad, G.: On the generators of quantum dynamical semigroups. Commun. Math. Phys. **48**(2), 119–130 (1976)
16. Lindblad, G.: Nonequilibrium Entropy and Irreversibility, vol. 5. D. Reidel Publishing, Dordrecht (1983)
17. Maas, J.: Gradient flows of the entropy for finite Markov chains. J. Funct. Anal. **261**, 2250–2292 (2011)
18. Marsden, J.E., Ratiu, T.S.: Introduction to Mechanics and Symmetry, vol. 17 of Texts in Applied Mathematics, 2nd edn. Springer, New York (1999)
19. Mielke, A.: Formulation of thermoelastic dissipative material behavior using GENERIC. Contin. Mech. Thermodyn. **23**(3), 233–256 (2011)
20. Mielke, A.: Geodesic convexity of the relative entropy in reversible Markov chains. Calc. Var. Part. Diff. Eqns. (2012). Online 10.1007/s00526-012-0538-8
21. Mielke, A., Neidhardt, H., Racec, P.: Coupling simple heat baths to quantum systems using GENERIC. In preparation (2013)
22. Mielke, A., Thomas, M.: GENERIC – a powerful tool for thermomechanical modeling. In preparation (2012)
23. Michor, P.W., Petz, D., Andai, A.: On the curvature of a certain Riemannian space of matrices. Infin. Dimens. Anal. Quantum Probab. Relat. Top. **3**(2), 199–212 (2000)
24. Öttinger, H.C., Grmela, M.: Dynamics and thermodynamics of complex fluids. II. Illustrations of a general formalism. Phys. Rev. E (3) **56**(6), 6633–6655 (1997)
25. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. Commun. Partial Differ. Equ. **26**, 101–174 (2001)
26. Öttinger, H.C.: Beyond Equilibrium Thermodynamics. Wiley, New Jersey (2005)
27. Öttinger, H.C.: The nonlinear thermodynamic quantum master equation. Phys. Rev. A **82**, 052119(11) (2010)
28. Öttinger, H.C.: The geometry and thermodynamics of dissipative quantum systems. Europhys. Lett. **94**, 10006(6) (2011)
29. Petz, D., Sudár, C.: Extending the Fisher metric to density matrices. In: Barndorff-Nielsen, O., Vendel Jensen, E. (eds.), Geometry in Present Day Science, pp. 21–34. World Scientific, Singapore (1999)
30. Petz, D.: Geometry of canonical correlation on the state space of a quantum system. J. Math. Phys. **35**(2), 780–795 (1994)
31. Streater, R.F.: Information geormetry and reduced quantum description. Rep. Math. Phys. **38**, 419–436 (1996)
32. Streater, R.F.: The analytic quantum information manifold. Canad. Math. Soc. Conf. Proc. **29**, 603–611 (2000)
33. Tanimura, Y.: Stochastic Liouville, Langevin, Fokker–Planck, and master equation approaches to quantum dissipative systems. J. Phys. Soc. Jpn. **75**(8), 082001–39 (2006)
34. von Renesse, M.-K.: An optimal transport view of Schrödinger equation. Canad. Math. Bull. **55**, 858–869 (2012)
35. Weiss, U.: Quantum Dissipative Systems, 2nd edn. World Scientific, Singapore (1999)

Chapter 22
Modelling of Thin Martensitic Films with Nonpolynomial Stored Energies

Martin Kružík and Johannes Zimmer

Dedicated to Jürgen Scheurle on the occasion of his 60th birthday

Abstract A study of the thin film limit of martensitic materials is presented, with the film height tending to zero. The behaviour of the material is modelled by a stored elastic energy which grows to infinity if the normal to the deformed film tends to zero. We show that the macroscopic behaviour of the material can be described by gradient Young measures if Dirichlet boundary conditions are prescribed at the boundary of the film. In this situation, we also formulate a rate-independent problem describing evolution of the material. A second approach, perhaps useful in case of non-Dirichlet loading, is presented as well, relying on suitable generalised Young measures.

22.1 Introduction

In this article, we consider the thin film limit of a model for shape-memory alloys. Shape-memory alloys have been the focus of many investigations in the last decade. This interest can partially be attributed to the shape-memory effect itself (see Sect. 22.1.1), but even more the nonconvexity of the Helmholtz energy density

M. Kružík
Institute of Information Theory and Automation of the ASCR, Pod vodárenskou věží 4, 182 08 Prague, Czech Republic, and Faculty of Civil Engineering, Czech Technical University, Thákurova 7, 166 29 Prague, Czech Republic

J. Zimmer (✉)
Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, United Kingdom
e-mail: zimmer@maths.bath.ac.uk

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_22,

due to the co-existence of several variants, which poses a significant mathematical challenge.

Here, our motivation is to address one typical difficulty of modelling shape memory alloys. Namely, a common framework for such models is three-dimensional elasticity, and more specifically hyperelasticity, which means that the first Piola-Kirchhoff stress tensor has a potential W. Static equilibria are minimisers of the elastic energy; one is thus led to solve

$$\text{minimise } J(y) := \int_\Omega W(\nabla y(x))\,\mathrm{d}x\,, \tag{22.1}$$

where $\Omega \subset \mathbb{R}^n$ denotes the reference configuration of the material, ∇ is the gradient operator, $y \in W^{1,p}(\Omega;\mathbb{R}^n)$ is the deformation, with $1 < p < +\infty$, $y = y_0$ on $\partial\Omega$, and $W\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ is the stored energy density.

A central point of interest of this paper is to incorporate the important physical assumption

$$W(F) \to +\infty \text{ whenever } \det(F) \searrow 0 \tag{22.2}$$

(this is related to the physical constraint that an elastic deformation of a body has to be orientation-preserving, which means $\det(\nabla y(x)) > 0$ almost everywhere).

One class of materials where this constraint can be included is that of *polyconvex* W, i.e., $W(F)$ can be written as a convex function of all minors of F. The existence of minimisers to (22.1) was proved by J. M. Ball in his pioneering paper [4]. The existence theory for polyconvex energy densities can deal with the growth behaviour (22.2). We refer, e.g., to [12, 14] for various results in this direction.

However, materials cannot be modelled by polyconvex stored energies. Prominent examples are materials with microstructure, such as shape-memory materials [6, 32], and we see this analysis as a prototype of modelling materials whose stored energy is not polyconvex (or quasiconvex, see below). We develop a framework in which the constraint (22.2) can be included, even in the presence of oscillations and concentrations of minimising functions. Other than the theory for polyconvex functions, there are few results. For a limit leading to a one-dimensional equation (where it is not a significant constraint to be a gradient), Freddi and Paroni develop a comprehensive Young measure approach [20], building on earlier work by Acerbi, Buttazzo and Percivale [1]. The requirement (22.2) also appears in the relaxation result by Ben Belgacem [5], which is also inspired by [1]. For the full vectorial case, we refer to recent results in this direction by one of the authors and coworkers [9]. Here, we pursue a different line of thought, by considering a thin-film equivalent of (22.2) (see (22.10) below). We exploit the fact that the related quasiconvex envelope has polynomial growth; this allows us to study both the static case and a rate-independent evolution.

22.1.1 Shape Memory Alloys

Some materials can, after a deformation, recover their original shape upon heating, and this is called the shape memory effect. We summarise key properties of this effect here; see [28] for a more extensive discussion. It is based on the ability of the shape-memory alloy to rearrange atoms in different crystallographic configurations (in particular, with different symmetry groups). Such materials have a high-temperature phase called austenite and a low-temperature phase called martensite. Since austenite is more symmetric, the martensitic phase exists in several variants, with the number of variants M, say, being the quotient of the order of the austenitic symmetry group and the order of the martensitic group. So for a cubic high-symmetry phase, $M = 3$, 6, 12, or 4 for the tetragonal, orthorhombic, monoclinic, respectively, triclinic martensites. We denote the stress-free strains of the variants U_ℓ, $\ell = 1, 2, \ldots, M$, and U_0 stands for the stress-free strain of the austenite. Since the martensitic phase exists in several symmetry related variants, it can form a microstructure by mixing those variants (possibly also with the austenitic variant) on a fine scale. Examples of these coherent combinations are twins of two variants, which is often called a laminate. This ability to form microstructures is one reason why shape memory alloys, as, for example, Ni-Ti, Cu-Al-Ni or In-Th, have various technological applications.

22.1.2 Variational Models for Shape Memory Alloys

Variational models for microstructures assume that formed structure has some optimality property. The reason for the formation of microstructures is that no exact optimum can be achieved and optimising sequences have to develop finer and finer oscillations. A typical example is a microstructure in a shape memory alloy.

We confine ourselves to the case of negligible hysteretic behaviour. This leads to a multidimensional vectorial variational problem, whose relaxation (i.e., suitable extension) is not yet satisfactorily understood. We study microstructures on mesoscopical level, which means that we do not only take care of some macroscopic effective response of the material but also provide some information on optimising sequences. In the last decade, similar mesoscopical models equipped with suitable dissipative potentials have been developed to treat materials with significant hysteresis; see [22, 30]. For a review of various mathematical problems related to martensitic crystals, we refer the reader to [32].

For shape memory alloys, W is not quasiconvex [29]. We recall that quasiconvexity means that for all $\varphi \in W_0^{1,\infty}(\Omega; \mathbb{R}^n)$ and all $F \in \mathbb{R}^{n\times n}$

$$|\Omega|\, W(F) \le \int_\Omega W(F + \nabla\varphi(x))\, \mathrm{d}x\ ; \tag{22.3}$$

we introduce for later use the notation $Qv\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ for the quasiconvex envelope of $v\colon \mathbb{R}^{m\times n} \to \mathbb{R}$. That is,

$$Qv(F)\,|\Omega| := \inf_{\varphi \in W_0^{1,\infty}(\Omega;\mathbb{R}^m)} \int_\Omega v(F + \nabla\varphi(x))\,\mathrm{d}x\ .$$

For energies which are not quasiconvex, minimisers to J in (22.1) do not necessarily exist. However, if we give up (22.2) and suppose that W has polynomial growth at infinity, so that for $c, C > 0$

$$c\,(-1 + |F|^p) \le W(F) \le C\,(1 + |F|^p)\ , \tag{22.4}$$

the existence of a solution to (22.1) is guaranteed if W is quasiconvex. Here and below, $|F| := \sqrt{\mathrm{tr}(F^\top F)}$ denotes the Frobenius norm of the matrix F.

Yet, quasiconvexity is a complicated property and difficult to verify in most concrete cases. Moreover, as mentioned above, the stored energy densities of materials with microstructure are not quasiconvex. As a result, solutions to (22.1) might not exist. Various relaxation techniques were developed [14, 32, 33] to overcome this drawback. One is to extend the notion of solutions from Sobolev mappings to parametrised measures called Young measures [3, 33–35]. The idea is to describe limit behaviour of $\{J(y_k)\}_{k\in\mathbb{N}}$ along a minimising sequence $\{y_k\}_{k\in\mathbb{N}}$. Nevertheless, the *growth condition* (22.4) *is still a key ingredient* in these considerations. We sketch in this article a new approach to deal with more general growth conditions, allowing to incorporate (22.2).

22.2 Thin Films

22.2.1 Static Problems

Bhattacharya and James [7] considered the following problem of a thin film limit. Let $\omega \subset \mathbb{R}^2$ be an open, bounded domain with Lipschitz boundary. We write $I := (0,1)$ and define $\Omega_\epsilon := \omega \times \epsilon I$ as the reference state for the space occupied by a specimen. Let $\{e_1, e_2, e_3\}$ be an orthonormal basis in $\mathbb{R}^3$; we suppose that e_3 is perpendicular to the plane of the film, whereas e_1, e_2 lie in the film plane.

The plane gradient $\nabla_{1,2}$ of a (weakly differentiable) map $y\colon \omega \to \mathbb{R}^3$ denoting the deformation is defined as

$$\nabla_{1,2} y = y_{,1} \otimes e_1 + y_{,2} \otimes e_2\ ,$$

where $y_{,i}$ denotes the vector of derivatives of y with respect to x_i, $i = 1, 2$. Moreover, having a matrix $F \in \mathbb{R}^{3\times 3}$, we write $F := (f_1|f_2|f_3)$ if $F = f_1 \otimes e_1 + f_2 \otimes e_2 + f_3 \otimes e_3$, where $f_i \in \mathbb{R}^3$ for $i = 1, 2, 3$.

Bhattacharya and James study the problem

$$\text{minimise } J_\epsilon^\kappa(y) = \frac{1}{\epsilon}\int_{\Omega_\epsilon} \left[W(\nabla y(x)) + \kappa\left|\nabla^2 y(x)\right|^2\right]\,\mathrm{d}x\ , \tag{22.5}$$

where $\kappa > 0$ is a constant describing the surface energy and

$$y \in \{u \in W^{2,2}(\Omega_\epsilon; \mathbb{R}^3) \,\big|\, y(x) = Ax \text{ if } x \in \partial\Omega_\epsilon\} \,,$$

with $A \in \mathbb{R}^{3\times 3}$ fixed. It is shown that (up to a subsequence), minimisers y^ϵ, say, of J_ϵ^κ satisfy for $\epsilon \to 0$

$$\begin{aligned} \nabla_{1,2}^2 y^\epsilon &\to \nabla_{1,2}^2 \bar{y} \text{ in } L^2(\Omega_1) \,, \\ \frac{1}{\epsilon} \nabla y_{,3}^\epsilon &\to \nabla_{,3} \bar{b} \text{ in } L^2(\Omega_1) \,, \\ \frac{1}{\epsilon^2} y_{,33}^\epsilon &\to 0 \text{ in } L^2(\Omega_1) \,. \end{aligned}$$

Moreover, $(\bar{y}, \bar{b}) \in W^{2,2}(\omega; \mathbb{R}^3) \times W^{1,2}(\omega; \mathbb{R}^3)$ minimise the energy

$$J_0^\kappa(y, b) := \int_\omega \left[W\left(y_{,1}(x) | y_{,2}(x) | b(x)\right) + \kappa \left(\left| \nabla_{1,2}^2 y(x) \right|^2 + |\nabla_{1,2} b|^2 \right) \right] \mathrm{d}S \tag{22.6}$$

subject to the boundary conditions $y(x_1, x_2) = (a_1|a_2)x$ and $b(x_1, x_2) = a_3$ if $(x_1, x_2) \in \partial\omega$. The coefficients $(a_1|a_2|a_3) \in \mathbb{R}^{3\times 3}$ are fixed. Physically, $y \colon \omega \to \mathbb{R}^3$ describes the average deformation of the film, while $b \colon \omega \to \mathbb{R}^3$ describes the shear of the cross-section of the film. Since κ is small, we may consider the model without the surface energy, i.e., the elastic energy stored in the film is now

$$J_0(y, b) := \int_\omega W\left(y_{,1}(x) | y_{,2}(x) | b(x)\right) \mathrm{d}S \,. \tag{22.7}$$

The functional $\bar{J}_0$ is nonconvex and minimiser does not have to exist in the set $W^{1,2}(\omega; \mathbb{R}^3) \times L^2(\omega; \mathbb{R}^3)$ equipped with affine boundary conditions.

There is a central difference to the analogous model for the bulk material. Namely, let us consider the situation where $\omega = \omega_1 \cup \omega_2 \cup L$, with ω_1 and ω_2 being disjoint subsets of ω, and L being a line interface between them, and that

$$(y_{,1}|y_{,2}|b) = \begin{cases} R_i F_i & \text{in } \omega_1 \\ R_j F_j & \text{in } \omega_2 \,, \end{cases}$$

where $R_i, R_j \in \mathrm{SO}(3)$ and F_i, F_j are zero energy deformation gradients. One further requires that y is continuous in ω, while $y_{,1}$, $y_{,2}$ as well as b may suffer jumps across the interface L. It is shown in [7] that in order to satisfy these requirements, the following *thin-film twinning equation* must be satisfied

$$R_i F_i - R_j F_j = a \otimes n + c \otimes e_3 \,, \tag{22.8}$$

where $a, n \in \mathbb{R}^3$, $n \cdot e_3 = 0$ and $c \in \mathbb{R}^3$ denotes the jump of b across the interface. The vector n is normal to the line interface. Thus, we say that martensitic variants i and j can form a *linear thin-film interface* if there are rotations R_i, R_j and vectors a, n, c as above that (22.8) holds. We note that this condition is much weaker than the bulk situation, where $\operatorname{rank}(R_iF_i - R_jF_j) = 1$ has to hold. Namely, it is a necessary and sufficient condition that one can construct a piecewise affine but continuous map whose gradient only takes values R_iF_i and R_jF_j, $i \neq j$. As a consequence, there are interfaces between martensitic variants in the thin film which cannot exist in the bulk material.

Since the surface energy of the film is not considered here, the model includes only b, but not its gradient. Therefore we can eliminate b from the theory by setting

$$\bar{W}(f_1|f_2) := \min_{b\in\mathbb{R}^3} W(f_1|f_2|b)\,. \tag{22.9}$$

The continuity, coercivity and boundedness of W in the form of (22.4) ensure that a minimum exists. Hence, we can rewrite (22.7) as

$$J(y) := \int_\omega \bar{W}\left(y_{,1}(x)|y_{,2}(x)\right)\,\mathrm{d}S\,,$$

because then the minima of J and J_0 are the same.

If we want to include a condition analogous to (22.2) in the thin-film model, we immediately see that f_1, f_2 in (22.9) should not be parallel, since otherwise $\det(|f_1|f_2|b) = 0$. Namely, $f_1 \times f_2$ is the normal vector to the thin-film surface in the deformed configuration $y(\omega)$ and $|f_1 \times f_2|$ measures area changes. More precisely, if $y\colon \omega \to \mathbb{R}^3$ is invertible, then for $O \subset \omega$ measurable

$$\operatorname{meas}(y(O)) = \int_O |y_{,1} \times y_{,2}|\,\mathrm{d}S\,.$$

Thus, taking $r > 1$, we define the following modified thin-film energy density $\hat{W}\colon \mathbb{R}^{3\times 2} \to \mathbb{R} \cup \{+\infty\}$

$$\hat{W}(f_1|f_2) := \bar{W}(f_1|f_2) + \frac{1}{|f_1 \times f_2|^r}\,. \tag{22.10}$$

Consider $y_0\colon \bar{\omega} \to \mathbb{R}^3$, a continuous and piecewise affine mapping, and set

$$W^{1,p}_{y_0}(\omega;\mathbb{R}^3) := \{z \in W^{1,p}(\omega;\mathbb{R}^3);\ z = y_0 \text{ on } \partial\omega\}\,.$$

Moreover, let us define $I\colon W^{1,p}_{y_0}(\omega;\mathbb{R}^3) \to \mathbb{R} \cup \{+\infty\}$,

$$I(y) := \int_\omega \hat{W}\left(y_{,1}(x)|y_{,2}(x)\right)\,\mathrm{d}S$$

and $I_Q : W^{1,p}_{y_0}(\omega;\mathbb{R}^3) \to \mathbb{R} \cup \{+\infty\}$,

$$I_Q(y) := \int_\omega Q\hat{W}\left(y_{,1}(x)|y_{,2}(x)\right)\,\mathrm{d}S\,.$$

We immediately see that $\hat{W}$ does not satisfy polynomial growth assumptions and standard coercivity conditions. Nevertheless, we have the following relaxation result due to Anza Hafsa and Mandallena [2, Th. 1.2].

Proposition 22.1. *Assume that* (22.10) *holds, and let* $+\infty > p > 1$. *Then* $\inf\{I(u);\ u \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)\} = \min\{I_Q(u);\ u \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)\}$. *Moreover, if* $\{u_k\} \subset W^{1,p}_{y_0}(\omega;\mathbb{R}^3)$ *is a minimising sequence for* I *and* $u_k \rightharpoonup u$ *in* $W^{1,p}(\omega;\mathbb{R}^3)$, *then* u *is a minimiser to* I_Q. *On the other hand, if* $\bar{u} \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)$ *is a minimiser to* I_Q, *then there is a minimising sequence* $\{\bar{u}_k\} \subset W^{1,p}_{y_0}(\omega;\mathbb{R}^3)$ *of* I *which weakly converges to* $\bar{u}$ *in* $W^{1,p}(\omega;\mathbb{R}^3)$.

The proof of the proposition relies on the fact that $Q\hat{W}$ has a polynomial growth at infinity, i.e., for all $F \in \mathbb{R}^{3\times 2}$, $\left|Q\hat{W}(F)\right| \le C(1+|F|^p)$, with some $C > 0$.

In general, it is not possible to determine the quasiconvex envelope in closed form. There is, however, its representation in terms of gradient Young measures. While finding such a representation is an equally difficult problem, there are known subsets and supsets of gradient Young measures which can be exploited efficiently in numerical calculations [8, 26].

Let us denote by $\mathscr{G}^{p,r}_{y_0}(\omega;\mathbb{R}^{3\times 2})$ the set of gradient Young measures $\mu = \{\mu_x\}_{x\in\omega}$ generated by sequences of gradients of mappings from $W^{1,p}_{y_0}(\omega;\mathbb{R}^3)\}$ such that for $F = (f_1|f_2) \in \mathbb{R}^{3\times 2}$,

$$\int_\omega \int_{\mathbb{R}^{3\times 2}} (|F|^p + |f_1 \times f_2|^{-r})\mu_x(\mathrm{d}F)\,\mathrm{d}x < +\infty\,.$$

Then we can define the integral functional $\mathscr{J} : \mathscr{G}^{p,r}(\omega;\mathbb{R}^{3\times 2}) \to \mathbb{R}$ as

$$\mathscr{J}(\nu) = \int_\omega \int_{\mathbb{R}^{3\times 2}} \hat{W}(F)\nu_x(\mathrm{d}F)\,\mathrm{d}x\,.$$

Proposition 22.2. *Let* (22.10) *hold and let* $p > 1$ *be finite. Then*

$$\begin{aligned}\min\{I_Q(u);\ u \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)\} &= \inf\{I(u);\ u \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)\}\\ &= \min\{\mathscr{J}(\mu);\ \mu \in \mathscr{G}^{p,r}_{y_0}(\omega;\mathbb{R}^{3\times 2})\}\,.\end{aligned}$$

Moreover, if ν *minimises* $\mathscr{J}$ *and* $\nabla u(x) = \int_{\mathbb{R}^{3\times 2}} \hat{F}\nu_x(\mathrm{d}F)$ *for almost all* $x \in \omega$ *and some* $u \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)$, *then* u *minimises* I_Q. *On the other hand, if* $y \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)$ *minimises* I_Q *and* $Q\hat{W}(\nabla y(x)) = \int_{\mathbb{R}^{3\times 2}} \hat{W}(F)\mu_x(\mathrm{d}F)$ *for some* $\mu \in \mathscr{G}^{p,r}_{y_0}(\omega;\mathbb{R}^{3\times 2})$, *then* μ *minimises* $\mathscr{J}$.

Proof. We use the fact that if $\{y_k\} \subset L^p(\Omega;\mathbb{R}^{m\times n})$ and ν is the associated Young measure, then for every normal integrand $\psi\colon \Omega \times \mathbb{R}^{m\times n} \to (-\infty;\infty]$ bounded from below it holds that [17, Theorem 8.2]

$$\liminf_{k\to\infty} \int_\Omega \psi(x, y_k(x))\,\mathrm{d}x \geq \int_\Omega \int_{\mathbb{R}^{m\times n}} \psi(x,s)\nu_x(\mathrm{d}s)\,\mathrm{d}x\,. \tag{22.11}$$

The first equality in the proposition then follows from Proposition 22.1. Indeed, combining Proposition 22.1 with (22.11), we obtain for a minimising sequence $\{u_k\} \subset W^{1,p}_{y_0}(\omega;\mathbb{R}^3)$ of I converging weakly to u, a minimiser of I_Q, and generating a Young measure ν such that

$$\int_\Omega Q\hat{W}(\nabla u(x))\,\mathrm{d}x = \lim_{k\to\infty} I(u_k) \geq \int_\omega \int_{\mathbb{R}^{3\times 2}} \hat{W}(F)\nu_x(\mathrm{d}F)\,\mathrm{d}x\,. \tag{22.12}$$

At the same time, $\nabla u(x) = \int_{\mathbb{R}^{3\times 2}} F\nu_x(\mathrm{d}F)$ for almost all $x \in \Omega$ and [23]

$$Q\hat{W}(\nabla u(x)) \leq \int_{\mathbb{R}^{3\times 2}} Q\hat{W}(F)\nu_x(\mathrm{d}F)\,\mathrm{d}x \leq \int_{\mathbb{R}^{3\times 2}} \hat{W}(F)\nu_x(\mathrm{d}F)\,. \tag{22.13}$$

By combining (22.12) and (22.13), we get that for almost all $x \in \Omega$

$$Q\hat{W}(\nabla u(x)) = \int_{\mathbb{R}^{3\times 2}} \hat{W}(F)\nu_x(\mathrm{d}F) \tag{22.14}$$

and that ν_x is supported on the set $\{F \in \mathbb{R}^{3\times 2};\ Q\hat{W}(F) = \hat{W}(F)\}$ for a.a. $x \in \Omega$. Thus we showed that $\min I_Q \geq \inf \mathscr{J}$. Assume that there is $\mu \in \mathscr{G}^{p,r}_{y_0}(\omega;\mathbb{R}^{3\times 2})$ such that $\mathscr{J}(\mu) < \min I_Q$. Then there is $\{y_k\} \in W^{1,p}_{y_0}(\omega;\mathbb{R}^3)$ such that $\{\nabla y_k\}$ generates μ and $y_k \rightharpoonup y$. Since $\hat{W} \geq Q\hat{W}$, it follows that

$$\begin{aligned}\lim_{k\to\infty} I(y_k) \geq \mathscr{J}(\mu) &= \int_\omega \int_{\mathbb{R}^{3\times 2}} \hat{W}(F)\mu_x(\mathrm{d}F) \geq \int_\omega \int_{\mathbb{R}^{3\times 2}} Q\hat{W}(F)\mu_x(\mathrm{d}F)\\ &\geq \int_\omega Q\hat{W}(\nabla y(x))\,\mathrm{d}x \geq \min I_Q\,,\end{aligned} \tag{22.15}$$

hence $\mathscr{J}(\mu) \geq \min I_Q$ in contradiction to our assumption that $\mathscr{J}(\mu) < \min I_Q$. The second but last inequality in (22.15) follows from Jensen's inequality (for quasiconvex functions, as in the characterisation of gradient Young measures given by Kinderlehrer and Pedregal [23]) since $\left|Q\hat{W}(F)\right| \leq C(1+|F|^p)$, as proved in [2]. □

22.2.2 *Evolutionary Problems*

Changes of external conditions typically lead to evolution of material deformation and may initiate phase transformations. This phenomenon is usually connected with energy dissipation. Hence, we enrich our static relaxed model, i.e., the one defined on Young measures ν by a suitable dissipation mechanism and time-dependent loading via Dirichlet boundary conditions.

22.2.2.1 Dissipation Related to Phase Transitions

The dissipation mechanism we describe to model phase transition is a two-dimensional version of a common one, for example described in [28]. For completeness, we give a summary, following the presentation in [28]. In order to describe dissipation due to transformations we adopt, as, e.g., [30], the standpoint that the amount of dissipated energy associated with a particular phase transition between austenite and a martensitic variant or between two martensitic variants can be described by a specific energy (of the dimension $\mathrm{J/m^2}$). For an explicit definition of the transformation dissipation, we need to identify the particular phases or phase variants. To do so, we define a continuous mapping $\mathscr{L}\colon \mathbb{R}^{3\times 2} \to \triangle$, where

$$\triangle := \left\{\zeta \in \mathbb{R}^{1+M} \;\middle|\; \zeta_\ell \geq 0 \text{ for } \ell = 1, \ldots, M+1, \text{ and } \sum_{\ell=1}^{M+1} \zeta_\ell = 1\right\}$$

is a simplex with $M+1$ vertices, with M being the number of martensitic variants. Here $\mathscr{L}$ is related to the material itself and thus has to be frame indifferent. We assume, besides $\zeta_\ell \geq 0$ and $\sum_{\ell=1}^{M+1} \zeta_\ell = 1$, that the coordinate ζ_ℓ of $\mathscr{L}(F)$ takes the value 1 if there is $b \in \mathbb{R}^3$ such that $(F|b)$ is in the ℓth (phase) variant, that is, $(F|b)$ is in a vicinity of the ℓth well $\mathrm{SO}(n)U_\ell$ of W, which can be identified by the stretch tensor $(F|b)^\top(F|b)$ being close to $U_\ell^\top U_\ell$. Here, $U_\ell^\top U_\ell$ denotes the right Cauchy–Green strain tensors of the stress-free strain states. They represent particular martensitic variants and the austenite. If $\mathscr{L}(F)$ is not in any vertex of $\triangle$, then it means that F is in the spinodal region where no definite phase or variant is specified. We assume, however, that the wells are sufficiently deep and the phases and variants are geometrically sufficiently far from each other that the tendency for minimisation of the stored energy will essentially prevent F to range into the spinodal region. Thus, the concrete form of $\mathscr{L}$ is not important as long as $\mathscr{L}$ enjoys the properties listed above. We remark that $\mathscr{L}$ plays the rôle of what is often called a vector of *order parameters* or a vector-valued *internal variable*.

For two states q_1 and q_2, with $q_j = (y_j, \nu_j, \lambda_j)$ (deformation, Young measure, and volume fraction) for $j = 1, 2$, we now define the dissipation due to martensitic transformation which "measures" changes in the volume fraction $\lambda \in L^\infty\left(\Omega; \mathbb{R}^{M+1}\right)$. Although λ_j is given by ν_j, it is convenient to consider them here

as a pair of independent variables and put their relationship as a constraint to the set of admissible states. This dissipation is given by

$$\mathscr{D}(q_1, q_2) := \int_\omega |\lambda_1(x) - \lambda_2(x)|_{\mathbb{R}^{M+1}} \mathrm{d}x \,, \tag{22.16}$$

where

$$\lambda_j(x) := \int_{\mathbb{R}^{3\times 2}} \mathscr{L}(F)\nu_{j,x}(\mathrm{d}F) \tag{22.17}$$

and $|\cdot|_{\mathbb{R}^{M+1}}$ is a norm on $\mathbb{R}^{M+1}$. As $\lambda^j(x)$ represents the volume fraction of the jth phase in the material point x we inevitably get $\sum_{j=1}^{M+1} \lambda^j = 1$ and we call λ the vector-valued volume fraction as it gives us relative portions of variants at almost every $x \in \Omega$. In what follows, we will assume that the norm on $\mathbb{R}^{M+1}$ defining the dissipation in (22.16) is given as

$$|X|_{\mathbb{R}^{M+1}} := \sum_{i=1}^{M+1} c^i \left|X^i\right| \,, \; X = (X^1, \ldots, X^{M+1}) \tag{22.18}$$

where $|\cdot|$ is the absolute value and $c^i > 0$ for all i. The physical meaning of c^i is the specific energy dissipated if X^i changes from zero to one (or vice versa).

22.2.2.2 Energetic Solution

Combining the previous considerations, we arrive at the energy functional i of the form

$$\mathrm{i}(t, q) := \int_\omega \int_{\mathbb{R}^{3\times 2}} \hat{W}(F + \nabla y_0(t, x))\nu_x(\mathrm{d}F)\mathrm{d}x + \varepsilon \left\|\nabla\lambda\right\|_{L^2\left(\omega;\mathbb{R}^{(1+M)\times 2}\right)} \,, \tag{22.19}$$

where the term $\nabla\lambda$ is included to regularise the problem. It penalises spatial jumps of the volume fraction λ and introduces a length scale to the problem, depending on the parameter $\varepsilon > 0$. In particular, it allows us to pass to the limit in the dissipation term. In order to define an admissible set where we look for a solution triple $q = (y, \nu, \lambda)$, we put

$$y \in W^{1,p}_{y_0}(\omega; \mathbb{R}^3) \tag{22.20}$$

Here $y_0(t) \in W^{1,p}(\omega; \mathbb{R}^3)$ with piecewise affine boundary conditions and $t \in [0; \mathfrak{T}]$ ranges within the process time interval, with $\mathfrak{T} > 0$ being the final time.

Then we look for $q \in \mathscr{Q} := W^{1,p}(\omega; \mathbb{R}^m) \times \mathscr{G}^{p,r}_0(\omega; \mathbb{R}^{m\times n}) \times W^{1,2}(\omega; \mathbb{R}^{M+1})$ and restrict the space further by imposing the *admissibility condition*

$$\mathbb{Q} := \left\{ q \in \mathscr{Q} \;\middle|\; \lambda = \mathscr{L} \bullet \nu \text{ and } \nabla y = \mathbb{I} \bullet \nu \right\} , \tag{22.21}$$

where, for almost all $x \in \Omega$, $[\mathscr{L} \bullet \nu](x) := \int_{\mathbb{R}^{m\times n}} \mathscr{L}(F)\nu_x(\mathrm{d}F)$.

Following [18], we assume that there are constants $C_0, C_1 > 0$ such that

$$|\partial_t \mathrm{i}(t,q)| \le C_0(C_1 + \mathrm{i}(t,q)) . \tag{22.22}$$

We also assume uniform continuity of $t \mapsto \partial_t \mathrm{i}(t,q)$ in the sense that there is $\omega\colon [0,\mathfrak{T}] \to [0,+\infty)$ nondecreasing such that for all $t_1, t_2 \in [0,\mathfrak{T}]$

$$|\partial_t \mathrm{i}(t_1,q) - \partial_t \mathrm{i}(t_2,q)| \le \omega(|t_1 - t_2|) . \tag{22.23}$$

Finally, we suppose that $q \mapsto \partial_t \mathrm{i}(t,q)$ is weakly continuous for all $t \in [0,\mathfrak{T}]$.

We seek to analyse the time evolution of a process $q(t) \in \mathbb{Q}$ during the time interval $[0,\mathfrak{T}]$. The following two properties are key ingredients of the so-called energetic solution [31].

(i) *Stability inequality*: for every $t \in [0,\mathfrak{T}]$ and every $\tilde{q} \in \mathbb{Q}$, it holds that

$$\mathscr{I}(t,q(t)) \le \mathscr{I}(t,\tilde{q}) + \mathscr{D}(q(t),\tilde{q}) . \tag{22.24}$$

(ii) *Energy balance*: For every $0 \le t \le \mathfrak{T}$,

$$\mathscr{I}(t,q(t)) + \mathrm{diss}(\mathscr{D},q;[0,t]) = \mathscr{I}(0,q(0)) + \int_0^t \partial_t \mathscr{I}(\xi,q(\xi))\,\mathrm{d}\xi , \tag{22.25}$$

where

$$\mathrm{diss}(\mathscr{D},q;[s,t]) := \sup\left\{ \sum_{j=1}^{N} \mathscr{D}\left(q\left(t_{j-1}\right), q\left(t_j\right)\right) \;\middle|\; \{t_j\}_{j=0}^{N} \text{ is a partition of } [s,t] \right\}$$

is the *variation* of the dissipation.

Definition 22.1. The mapping $q\colon [0,\mathfrak{T}] \to \mathbb{Q}$ is an *energetic solution* to the problem $(\mathscr{I},\mathscr{D})$ with the energy functional $\mathscr{I}$ as in (22.19) and the dissipation $\mathscr{D}$ if the stability inequality (22.24) and energy balance (22.25) are satisfied for every $t \in [0,\mathfrak{T}]$.

Further, we define the set of stable states at time $t \in [0,\mathfrak{T}]$ as

$$\mathbb{S}(t) := \{ q \in \mathbb{Q};\ \forall \tilde{q} \in \mathbb{Q} : \mathscr{I}(t,q) \le \mathscr{I}(t,\tilde{q}) + \mathscr{D}(q,\tilde{q}) \} .$$

In particular, we will always assume that the initial condition is stable, i.e., $q_0 \in \mathbb{S}(0)$. The following theorem regarding the existence of an energetic solution can be proved using a general strategy described in [18].

Theorem 22.1. *Let $p > 2$, and let assumptions (22.4), (22.22), and (22.23) hold. Then there is a process $q\colon [0,\mathfrak{T}] \to \mathbb{Q}$ with $q(t) = (y(t), \nu(t), \lambda(t))$ such that q is an energetic solution according to Definition 22.1 for a given stable initial condition $q_0 \in \mathbb{Q}$.*

Proof. The proof of this theorem follows a now well-established route and we thus omit any details [30]. The argument proceeds via semidiscretisations in time for decreasing time steps, by a limit passage in the stability inequality (22.24) and in the energy equality (22.25); cf. also [18] for a general strategy how to prove existence of energetic solutions. □

22.3 Problems Involving Concentration

The previous result relies on the specific form (22.10) of the thin film energy and on applied Dirichlet boundary conditions. An advantage of this approach is that the analysis remains in the realm of Young measures, and established tools from analysis can be applied to the quasiconvexified problem. This is just possible because minimizing sequences $\{y_k\}$ to $\mathscr{I} J$ are such that $\{W(\nabla y_k\}_{k\in\mathbb{N}}$ is weakly relatively compact in $L^1(\omega)$. Sometimes it is desirable to study problems where concentration effects may appear as well; then Young measures prove to be insufficient and DiPerna–Majda measures are an appropriate tool. This might perhaps happen for energies satisfying (22.26) if we require additionally that $\det \nabla y > 0$ in Ω. We sketch a corresponding framework and give a simple application. It is worth pointing out that while in spirit the approach is the same as the one taken in Sect. 22.2.1, we there start with a specific two-dimensional energy. Here, we consider a class of three-dimensional energy densities satisfying some growth conditions, and then pass to a two-dimensional setting by considering scaled versions.

Our goal is to tailor the relaxation to functions satisfying (22.2) in the situation of a thin film. The key new idea is that we allow W to depend on the inverse of its argument. Specifically, we suppose that W is continuous on regular matrices and that there exist positive constants $c, C > 0$ such that

$$c\left(-1 + |F|^p + \left|F^{-1}\right|^p\right) \le W(F) \le C\left(1 + |F|^p + \left|F^{-1}\right|^p\right). \tag{22.26}$$

We point out that (22.26) implies (22.2) and that W has polynomial growth in $|F|$ and $\left|F^{-1}\right|$ at infinity. Hence, in our setting we have an L^p bound not only on the deformation gradient but also on its inverse. Different and even negative powers of F (called the Seth–Hill family of strain measures) are frequently used to describe deformation strain, see, e.g., the survey [13]. Notice that if $y\colon \Omega \to \mathbb{R}^3$ is a deformation map and its inverse, $y^{-1}\colon y(\Omega) \to \mathbb{R}^3$ is smooth, then $(\nabla y(x))^{-1} = \nabla y^{-1}(y(x))$. Hence, exchanging the role of the reference and

the deformed configuration, the growth condition on F^{-1} just expresses that the gradient of the inverse deformation has the same integrability as the gradient of the original deformation.

A simple example of a function satisfying (22.26) is, e.g., a stored energy density describing martensitic materials:

$$W(F) := \min_{i=1,\dots,M} \left(\left| F^\top F - F_i^\top F_i \right|^2 + \left| F^{-1} F^{-\top} - F_i^{-1} F_i^{-\top} \right|^2 \right) ,$$

where $F_i \in \mathbb{R}^{3\times 3}$, $i = 1, \dots, M$, are positions of the minima of the multiwell energy. Due to the lack of convexity of W, the existence of a minimiser is typically not guaranteed in the Sobolev space $W^{1,p}(\Omega; \mathbb{R}^3)$; we only trace the behaviour of minimising sequences of J. We describe the necessary tools in the next subsection.

22.3.1 DiPerna–Majda Measures

Prior to developing the new framework, we sketch the established theory, which is described in greater detail in, for example, [27]. Unless stated otherwise, Ω is an open domain in $\mathbb{R}^n$. The definition of DiPerna–Majda measures involves a compactification; so let us take a separable completely regular algebra $\mathscr{R}$ of continuous bounded functions $\mathbb{R}^{n\times n} \to \mathbb{R}$. We recall that an algebra is *completely regular* if it contains the constants, separates points from closed subsets and is closed with respect to the maximum (Chebyshev) norm. It is known [16, Sect. 3.12.21] that there is a one-to-one correspondence $\mathscr{R} \mapsto \beta_{\mathscr{R}}\mathbb{R}^{n\times n}$ between such subalgebras of bounded continuous functions and metrisable compactifications of $\mathbb{R}^{n\times n}$; by a compactification we mean here a compact set, denoted by $\beta_{\mathscr{R}}\mathbb{R}^{n\times n}$, into which $\mathbb{R}^{n\times n}$ is homeomorphically and densely embedded. For simplicity, we shall not distinguish between $\mathbb{R}^{n\times n}$ and its image in $\beta_{\mathscr{R}}\mathbb{R}^{n\times n}$. Similarly, we do not distinguish between elements of $\mathscr{R}$ and their unique continuous extensions on $\beta_{\mathscr{R}}\mathbb{R}^{n\times n}$. The reader can, for instance, think about a one point compactification corresponding to the algebra of functions which have limits if the norm of its argument diverges to infinity. Another example is a compactification by the sphere generated by functions which have limits along rays arising from the origin.

Let σ be a positive Radon measure on $\bar{\Omega}$, $\sigma \in M(\bar{\Omega})$. We consider a map $\hat{\nu}$ mapping $x \in \bar{\Omega}$ to a Radon measure $\nu_x \in M(\beta_{\mathscr{R}}\mathbb{R}^{n\times n})$. We recall that such a map $\hat{\nu}\colon x \mapsto \hat{\nu}_x$ is weakly* σ-measurable if for any $v_0 \in C_0(\mathbb{R}^{n\times n})$, the mapping $\bar{\Omega} \to \mathbb{R}, x \mapsto \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} v_0(s)\hat{\nu}_x(\mathrm{d}s)$ is σ-measurable in the usual sense; the space of weakly* measurable functions is denoted $L^\infty_{\mathrm{w}}(\bar{\Omega}, \sigma; M(\beta_{\mathscr{R}}\mathbb{R}^{n\times n}))$. If additionally $\hat{\nu}_x$ is a probability measure, $\nu_x \in \mathrm{Prob}(\beta_{\mathscr{R}}\mathbb{R}^{n\times n})$ for σ-a.a. $x \in \bar{\Omega}$, then the collection $\{\hat{\nu}_x\}_{x\in\bar{\Omega}}$ is a Young measure on $(\bar{\Omega}, \sigma)$ [36]; see also [3, 33–35]. Young measures can record oscillations in minimising sequences but not concentration effects; an extension developed by DiPerna and Majda is capable of describing concentration effects as well.

Specifically, DiPerna and Majda [15] have shown that for a bounded sequence $\{y_k\}_{k\in\mathbb{N}}$ in $L^p(\Omega;\mathbb{R}^{n\times n})$ with $1 \le p < +\infty$, there exists a subsequence (not relabelled), a positive Radon measure $\sigma \in M(\bar\Omega)$ and a Young measure $\hat\nu\colon x \to \hat\nu_x$ on $(\bar\Omega,\sigma)$ such that $(\sigma,\hat\nu)$ is attainable by $\{y_k\}_{k\in\mathbb{N}}$ in the sense that for every $g \in C(\bar\Omega)$ and for every $v_0 \in \mathscr{R}$

$$\lim_{k\to\infty}\int_\Omega g(x)v(y_k(x))\mathrm{d}x = \int_{\bar\Omega}\int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} g(x)v_0(s)\hat\nu_x(\mathrm{d}s)\sigma(\ \mathrm{textrmd}x)\ , \tag{22.27}$$

where

$$v \in \mathcal{Y}^p_{\mathscr{R}}(\mathbb{R}^{m\times n}) := \left\{v_0(1+|\cdot|^p)\ \middle|\ v_0\in\mathscr{R}\right\}\ . \tag{22.28}$$

We remark that it is easy to see that (22.27) can be also written in the form

$$\lim_{k\to\infty}\int_\Omega h(x,y_k(x))\mathrm{d}x = \int_{\bar\Omega}\int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} h_0(x,s)\hat\nu_x(\mathrm{d}s)\sigma(\mathrm{d}x)\ , \tag{22.29}$$

where $h(x,s) := h_0(x,s)(1+|s|^p)$ and $h_0 \in C(\bar\Omega\otimes\beta_{\mathscr{R}}\mathbb{R}^{n\times n})$.

In particular, setting $v_0 = 1 \in \mathscr{R}$ in (22.27), we can see that

$$\lim_{k\to\infty}(1+|y_k|^p) \ =\ \sigma \qquad \text{weakly* in } M(\bar\Omega)\ . \tag{22.30}$$

We say that $\{y_k\}_{\in\mathbb{N}}$ *generates* $(\sigma,\hat\nu)$ if (22.27) holds. Let us write $\mathscr{DM}^p_{\mathscr{R}}(\Omega;\mathbb{R}^{m\times n})$ for the set of all *DiPerna–Majda measures*, that is, the set of all pairs $(\sigma,\hat\nu) \in M(\bar\Omega)\times L^\infty_{\mathrm{w}}(\bar\Omega,\sigma;M(\beta_{\mathscr{R}}\mathbb{R}^{n\times n}))$ attainable by sequences from $L^p(\Omega;\mathbb{R}^{n\times n})$. Note that, taking $v_0 = 1$ in (22.27), one can see that these sequences must inevitably be bounded in $L^p(\Omega;\mathbb{R}^{n\times n})$. The explicit description of the elements from $\mathscr{DM}^p_{\mathscr{R}}(\Omega;\mathbb{R}^{m\times n})$ for unconstrained sequences is given in [24, Theorem 2] or in [25].

Here the energy depends on the deformation gradient and its inverse. We first ignore the latter dependence and return to this central point at the end of this section. We thus consider the subset of $\mathscr{DM}^p_{\mathscr{R}}(\Omega;\mathbb{R}^{m\times n})$ which are generated by $\{\nabla y_k\}_{k\in\mathbb{N}}$ for some bounded $\{y_k\}_{k\in\mathbb{N}} \subset W^{1,p}(\Omega;\mathbb{R}^m)$; this subset is here denoted as $\mathscr{GDM}^p_{\mathscr{R}}(\Omega;\mathbb{R}^{m\times n})$. Elements of $\mathscr{GDM}^p_{\mathscr{R}}(\Omega;\mathbb{R}^{m\times n})$ generated by gradients of mappings with the same trace on $\partial\Omega$ are characterised in the following theorem, which is proved in [21]. To formulate the statement, we introduce the notation $d_\sigma \in L^1(\Omega)$ for the absolutely continuous part of σ in the Lebesgue decomposition of σ, with respect to the Lebesgue measure. We recall that Qv denotes the quasiconvex envelope of a function v.

Theorem 22.2. *Let $\Omega \subset \mathbb{R}^n$ be a bounded Lipschitz domain , $1 < p < +\infty$ and $(\sigma,\hat\nu) \in \mathscr{DM}^p_{\mathscr{R}}(\Omega;\mathbb{R}^{m\times n})$. Then there is a bounded sequence $\{y_k\}_{k\in\mathbb{N}} \subset W^{1,p}(\Omega;\mathbb{R}^m)$ such that $y_k - y_j \in W^{1,p}_0(\Omega;\mathbb{R}^m)$ for any $j,k\in\mathbb{N}$ (i.e., all have*

the same trace) and $\{\nabla y_k\}_{k\in\mathbb{N}}$ *generates* $(\sigma,\hat{\nu})$ *if and only if the following three conditions hold:*

1. *There exists* $y \in W^{1,p}(\Omega;\mathbb{R}^m)$ *such that for Lebesgue-almost every* $x \in \Omega$

$$\nabla y(x) = d_\sigma(x) \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} \frac{s}{1+|s|^p}\hat{\nu}_x(\mathrm{d}s)\,. \tag{22.31}$$

2. *For all* $v \in \Upsilon^p_{\mathscr{R}}(\mathbb{R}^{m\times n})$ *as defined in* (22.28)*, it holds that Lebesgue-almost everywhere*

$$Qv(\nabla y(x)) \le d_\sigma(x) \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} \frac{v(s)}{1+|s|^p}\hat{\nu}_x(\mathrm{d}s)\,. \tag{22.32}$$

3. *For* σ*-almost all* $x \in \bar{\Omega}$ *and all* $v \in \Upsilon^p_{\mathscr{R}}(\mathbb{R}^{m\times n})$ *with* $Qv > -\infty$ *it holds that*

$$0 \le \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}\setminus\mathbb{R}^{n\times n}} \frac{v(s)}{1+|s|^p}\hat{\nu}_x(\mathrm{d}s)\,. \tag{22.33}$$

22.3.2 DiPerna–Majda Measures Depending on the Inverse

We now consider the case where the energy depends on the deformation gradient and its inverse. An existence result for the DiPerna–Majda measures generated by functions with this dependence is required; we state the equivalent to (22.27) for this case. In what follows, $\mathbb{R}^{n\times n}_{\mathrm{inv}}$ denotes the set of invertible matrices. The proof of the following theorem is exactly the same as the proof of [33, Theorem 3.2.12].

Theorem 22.3. *Let* $\Omega \subset \mathbb{R}^n$ *be open and bounded, and let* $\{Y_k\}_{k\in\mathbb{N}}$, $\{Y_k^{-1}\}_{k\in\mathbb{N}} \subset L^p(\Omega;\mathbb{R}^{n\times n})$ *be bounded, for some* p *with* $1 \le p < +\infty$*. Then there are a subsequence of* $\{Y_k\}_{k\in\mathbb{N}}$ *(not relabeled),* $\pi \in M(\bar{\Omega})$ *and* $\hat{\mu} \in L^\infty_{\mathrm{w}}(\bar{\Omega},\pi;M(\beta_{\mathscr{R}}\mathbb{R}^{n\times n}))$ *such that for every* $g \in C(\bar{\Omega})$ *and every* $v(s) = v_1(s)(1+|s|^p+\left|s^{-1}\right|^p)$ *with* $v_1 : \mathbb{R}^{n\times n}_{\mathrm{inv}} \to \mathbb{R}$ *which can be continuously extended to* $v_0 \in \mathscr{R}$*, it holds that*

$$\lim_{k\to\infty}\int_\Omega g(x)v(Y_k(x))\,\mathrm{d}x = \int_{\bar{\Omega}}\int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} g(x)v_0(s)\hat{\mu}_x(\mathrm{d}s)\,\pi(\mathrm{d}x)\,. \tag{22.34}$$

Moreover, $\pi = w^* - \lim_{k\to\infty} 1+|Y_k|^p+\left|Y_k^{-1}\right|^p$ *in* $M(\bar{\Omega})$.

We will denote the set of pairs $(\pi,\hat{\mu})$ defined in Theorem 22.3 by $\mathscr{DM}^{p,-p}_{\mathscr{R}}(\Omega;\mathbb{R}^{n\times n})$ and its subset generated by gradients of functions from $W^{1,p}(\Omega;\mathbb{R}^n)$, i.e., if $Y_k := \nabla y_k$, by $\mathscr{GDM}^{p,-p}_{\mathscr{R}}(\Omega;\mathbb{R}^{n\times n})$.

The classic DiPerna–Majda measures are fully characterised, as stated in Theorem 22.2. For measures depending on the inverse as defined in Theorem 22.3, there

is no full characterisation available; this is since the nonlinearity introduced by the inverse rules out the application of existing tools for the characterisation of DiPerna–Majda measures. We now give a partial characterisation. Taking $v(s) := 1 + |s|^p$, we have $v_0(s) = 1$ in (22.27); we claim that

$$v_0(s) := \begin{cases} \frac{1+|s|^p}{1+|s|^p+|s^{-1}|^p} & \text{if } s \in \mathbb{R}^{n\times n}_{\mathrm{inv}} \\ 0 & \text{otherwise} \end{cases} \tag{22.35}$$

in (22.34). To see this, we set

$$v_1(s) = \frac{1+|s|^p}{1+|s|^p+|s^{-1}|^p}$$

and notice that

$$\lim_{\substack{|s^{-1}|\to\infty \\ |s| \text{ bounded}}} v_1(s) = 0$$

exists; thus we can first extend v_1 by continuity to the set of non-invertible matrices and obtain v_0 as in (22.35). Comparing (22.27) and (22.34), we then find for π-almost all $x \in \bar{\Omega}$

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\pi}(x) = \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} \frac{1+|s|^p}{1+|s|^p+|s^{-1}|^p}\hat{\mu}_x(\mathrm{d}s)\,. \tag{22.36}$$

Using this and choosing arbitrary $v_0 \in \mathscr{R}$ we obtain

$$\begin{aligned} &\int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} v_0(s)\hat{\nu}_x(\mathrm{d}s) \\ &= \left(\int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} \frac{1+|s|^p}{1+|s|^p+|s^{-1}|^p}\hat{\mu}_x(\mathrm{d}s)\right)^{-1} \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}} \frac{v_0(s)(1+|s|^p)}{1+|s|^p+|s^{-1}|^p}\hat{\mu}_x(\mathrm{d}s)\,. \end{aligned} \tag{22.37}$$

To ensure that the DiPerna–Majda measure is generated by a sequence of gradients (and their inverses), we use this characterisation in terms of $(\sigma, \hat{\nu})$. Namely, we require that the DiPerna–Majda measure $(\sigma, \hat{\nu})$ defined by (22.36), (22.37) must belong to $\mathscr{GDM}^p_{\mathscr{R}}(\Omega;\mathbb{R}^{m\times n})$; that is, the conditions in Theorem 22.2 must be fulfilled. We remark that then the gradient of the macroscopic deformation ∇y can be expressed for almost all $x \in \Omega$ as

$$\nabla y(x) = d_\pi(x) \int_{\beta_{\mathscr{F}}\mathbb{R}^{n\times n}_{\mathrm{inv}}} \frac{s}{1+|s|^p+|s^{-1}|^p}\hat{\mu}_x(\mathrm{d}s)\,, \tag{22.38}$$

where d_π is the density of the absolutely continuous part of π with respect to the Lebesgue measure.

Then the relaxation of J reads

$$\text{minimise} \int_{\bar{\Omega}} \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}_{\text{inv}}} \frac{W(s)}{1+|s|^p+|s^{-1}|^p} \hat{\mu}_x(\mathrm{d}s)\,\pi(\mathrm{d}x)\,, \tag{22.39}$$

where $(\pi, \hat{\mu}) \in \mathscr{GDM}^{p;-p}_{\mathscr{R}}(\Omega;\mathbb{R}^{n\times n})$, where $\hat{\mu}_x$ is supported on the set of matrices with positive determinant for π-almost all $x \in \bar{\Omega}$.

22.3.3 Application to a Thin Film Model

We now describe a setting for the analysis of thin martensitic films. A similar framework has been analysed by Bocea [10]; while the description of the thin film bears many resemblances, there are two crucial differences. We allow the energy to depend on the inverse, with the growth condition as in (22.26). This has the benefit that the important physical constraint (22.2) is satisfied. However, a price to pay is we cannot apply a decomposition lemma as used by Bocea [10]; it is an open problem whether a decomposition lemma holds for measures generated by gradients and their inverses. We recall the classical decomposition lemma, which can be found in [19].

Lemma 22.1. *Let $1 < p < +\infty$ and $\Omega \subset \mathbb{R}^n$ be an open bounded set and let $\{u_k\}_{k\in\mathbb{N}} \subset W^{1,p}(\Omega;\mathbb{R}^m)$ be bounded. Then there is a subsequence $\{u_j\}_{j\in\mathbb{N}}$ and a sequence $\{z_j\}_{j\in\mathbb{N}} \subset W^{1,p}(\Omega;\mathbb{R}^m)$ such that*

$$\lim_{j\to\infty} |\{x\in\Omega;\ z_j(x) \neq u_j(x) \text{ or } \nabla z_j(x) \neq \nabla u_j(x)\}| = 0 \tag{22.40}$$

and $\{|\nabla z_j|^p\}_{j\in\mathbb{N}}$ is relatively weakly compact in $L^1(\Omega)$. In particular, $\{\nabla u_j\}$ and $\{\nabla z_j\}$ generate the same Young measure.

It is, however, not known if an analogous lemma hold when a bound on the gradient inverse is also required. Thus we cannot rule out concentration effects and have to resort to a variant of DiPerna–Majda measures as described in Sect. 22.3.2. (The special case discussed in Sect. 22.2.1, with the specific energy given in (22.10), is an example where we have shown that no concentrations occur, so there the framework of Young measures is suitable).

We recall that $\omega \subset \mathbb{R}^2$ is an open, bounded domain with Lipschitz boundary, and that we set $I := (0,1)$ and $\Omega_\epsilon := \omega \times \epsilon I$ as the reference state for the space occupied by a specimen. Further, $y_\epsilon \colon \Omega_\epsilon \to \mathbb{R}^3$ is the deformation and $u_\epsilon \colon \Omega_\epsilon \to \mathbb{R}^3$ is the displacement, which are related to each other via the identity $y_\epsilon(x^\epsilon) = x^\epsilon + u_\epsilon(x^\epsilon)$, where $x^\epsilon \in \Omega_\epsilon$. Hence the deformation gradient is $F_\epsilon := \nabla y_\epsilon = \mathbb{I} + \nabla u_\epsilon$. Here, $\mathbb{I} \in \mathbb{R}^{3\times 3}$ is the identity matrix.

The total stored energy in the bulk occupying, in its reference configuration, the domain Ω_ϵ, is then

$$V(y_\epsilon) := \int_{\Omega_\epsilon} W(\nabla y_\epsilon(x^\epsilon))\,\mathrm{d}x^\epsilon\,. \tag{22.41}$$

Here the bulk free energy density W is a function $W\colon \mathbb{R}^{n\times n}_{\mathrm{inv}+} \to \mathbb{R}$, taking the deformation gradient as its argument. In reality, W also depends on the temperature, but we restrict here the analysis to the isothermal case of a temperature below the critical temperature, since the difficulty of non-convexity appears here as isolated as possible from other effects. Then several martensitic variants coexist, and this is what we want to capture.

The usual symmetry requirements will be made, namely frame indifference

$$W(QF) = W(F) \text{ for every } Q \in \mathrm{SO}(3) \text{ and } F \in \mathbb{R}^{n\times n}_{\mathrm{inv}+} \tag{22.42}$$

and crystalline symmetry

$$W(FP) = W(F) \text{ for every } P \in \mathscr{P} \text{ and } F \in \mathbb{R}^{n\times n}_{\mathrm{inv}+}\,, \tag{22.43}$$

here $\mathscr{P}$ denotes the point group of the austenitic phase. The set of minimisers of W is given by the martensitic variants $U_1, \dots, U_M$. The frame indifference (22.42) then implies that W is minimised on $\cup_{j=1}^{M} \mathrm{SO}(3)U_j$, the union of the orbits of U_j under the operation of $\mathrm{SO}(3)$ from the left.

It is convenient to consider $\Omega := \omega \times I$ which originates from Ω_ϵ via dilatation by $\frac{1}{\epsilon}$ in the direction of the third component. So, if the coordinates in Ω_ϵ are $(x_1^\epsilon, x_2^\epsilon, x_3^\epsilon)$, then the coordinates in Ω are (x_1, x_2, x_3) with

$$x_1 = x_1^\epsilon, \qquad x_2 = x_2^\epsilon, \qquad x_3 = \frac{1}{\epsilon} x_3^\epsilon\,.$$

The deformation $y_\epsilon\colon \Omega_\epsilon \to \mathbb{R}^3$ then gives in a natural way rise to the rescaled deformation $y\colon \Omega \to \mathbb{R}^3$ via $y(x) = y_\epsilon(x^\epsilon(x))$. A rescaling of the energy (22.41) by a factor $\frac{1}{\epsilon}$ yields then

$$V_\epsilon(y) := \int_\Omega W\left(\nabla_{1,2} y(x) \,\middle|\, \tfrac{1}{\epsilon}\nabla_3 y(x)\right)\,\mathrm{d}x\,; \tag{22.44}$$

we recall that $\nabla_{1,2}$ is the 3×2 matrix of partial derivatives $y_{j,k} = \frac{\partial y_j}{\partial x_k}$ with $j \in \{1,2,3\}$ and $k \in \{1,2\}$ and $\nabla_3 y = y_{,3}$ is the (column) vector containing the derivatives of y with respect to x_3 and $(F|f)$ denotes the 3×3 matrix formed of the 3×2 matrix F as the first two columns and the vector f as the third column.

We assume that W satisfies (22.26), that is,

$$c(-1 + |F|^p + \left|F^{-1}\right|^p) \le W(F) \le C(1 + |F|^p + \left|F^{-1}\right|^p)$$

for some $p \in \mathbb{N}$. This is a central difference to the work in [10], where the growth condition is in terms of F alone rather than both F and F^{-1}.

Definition 22.2 (Scaled DiPerna–Majda measures). Let $A = (a_1|a_2|a_3) \in \mathbb{R}^{3\times 3}$ be given. Let $\{y_k\}_{k\in\mathbb{N}}$ be a sequence of functions with affine boundary data in the sense that $y_k(x) = A^{\epsilon_k}x := (a_1|a_2|\epsilon_k a_3)x$ for $x \in \partial\omega \times I$. Suppose further that $\left(\nabla_{1,2}y_k(x) \,\middle|\, \frac{1}{\epsilon}\nabla_3 y_k(x)\right)$ and $\left(\nabla_{1,2}y_k(x) \,\middle|\, \frac{1}{\epsilon}\nabla_3 y_k(x)\right)^{-1}$ are both uniformly bounded in $L^p(\Omega;\mathbb{R}^{n\times n})$, for some p with $1 \le p < +\infty$. Then there is a subsequence which generates a measure $(\pi, \hat{\mu}) \in \mathscr{DM}^{p,-p}_{\mathscr{R}}(\Omega;\mathbb{R}^{n\times n})$; this measure is called a *scaled DiPerna–Majda measure*.

The existence of a scaled DiPerna–Majda measure follows directly from Theorem 22.3 .

As an application of the theory developed, we formulate the following result. A related result was given in [10].

Proposition 22.3. *Let W satisfy the growth condition (22.26). Let $\{\epsilon_k\}_{k\in\mathbb{N}}$ be a sequence of real numbers with $\epsilon_k \to 0$ as $k \to \infty$ and $\{y_k\}_{k\in\mathbb{N}}$ be a sequence of functions with affine boundary data, $y_k(x) = A^{\epsilon_k}x := (a_1|a_2|\epsilon_k a_3)x$ for $x \in \partial\omega \times I$. Suppose that $\{y_k\}_{k\in\mathbb{N}}$ is a minimising sequence for (22.44), in the sense that*

$$V_\epsilon(y_k) = \int_\Omega W\left(\nabla_{1,2}y_k(x) \,\middle|\, \tfrac{1}{\epsilon_k}\nabla_3 y_k(x)\right) \mathrm{d}x < I_k + \epsilon_k \,,$$

where

$$I_k := \inf\left\{V_{\epsilon_k}(y) \,\middle|\, y \in W^{1,p}(\Omega,\mathbb{R}^3), y(x) = A^{\epsilon_k}x \text{ for } x \in \partial\omega \times I\right\} \,.$$

Then a subsequence of $\{y_k\}_{k\in\mathbb{N}}$ (not relabeled) generates a scaled DiPerna–Majda measure $(\pi, \hat{\mu}) \in \mathscr{DM}^{p,-p}_{\mathscr{R}}(\Omega;\mathbb{R}^{n\times n})$, and $(\pi, \hat{\mu})$ $(\pi, \hat{\mu})$ minimises the effective film energy

$$\int_{\bar{\Omega}} \int_{\beta_{\mathscr{R}}\mathbb{R}^{n\times n}_{\mathrm{inv}}} \frac{W(s)}{1 + |s|^p + |s^{-1}|^p} \hat{\mu}_x(\mathrm{d}s)\, \pi(\mathrm{d}x) \,, \tag{22.45}$$

among all scaled gradient measures in $\mathscr{DM}^{p,-p}_{\mathscr{R}}(\Omega;\mathbb{R}^{n\times n})$.

Proof. The growth assumption (22.26) yields that

$$\left(\nabla_{1,2}y_k(x) \,\middle|\, \tfrac{1}{\epsilon}\nabla_3 y_k(x)\right) \quad \text{and} \quad \left(\nabla_{1,2}y_k(x) \,\middle|\, \tfrac{1}{\epsilon}\nabla_3 y_k(x)\right)^{-1}$$

are both uniformly bounded in $L^p(\Omega;\mathbb{R}^{n\times n})$. The existence of a scaled DiPerna–Majda measure follows then from Theorem 22.3, and it is a standard fact that minimising sequences $\{y_k\}_{k\in\mathbb{N}}$ generate a minimiser of the relaxed functional, which is here (22.45). □

22.4 Open Problems

The analysis of problems satisfying the physically natural growth assumption (22.2) is currently not well developed; we highlight some avenues of possible future research.

For the full vectorial case (i.e., $m, n > 1$), in [9], a relaxation theory in terms of Young measures is given. One challenge is the characterisation of the measures involved; a second one is the inclusion of possible concentrations. This would lead to a description of DiPerna–Majda measures depending on the inverse, as introduced in Sect. 22.3. There again, the characterisation of the measures obtained is a mathematical problem in its own right. A further problem is that at present we do not know under which conditions a decomposition lemma holds (see Sect. 22.3.3).

From the point of view of applications, the existence theory for models with energies satisfying the growth assumption (22.2) is a natural source of problems. In addition, the analysis of limits is open; for example, which thin-film energies of type (22.2) can be obtained from three-dimensional equations, in the limit of vanishing film thickness? Similarly as in bulk materials positivity of the determinant plays a crucial role, not only the length of the normal but also its orientation is important in thin films. We refer to [11] for a recent result in this direction using the notion of surface polyconvexity.

Acknowledgements The work of MK was partially supported by grants P201/10/0357, P201/12/0671, P105/11/0411 (GA ČR), the work of JZ was partially supported by EPSRC (EP/H05023X/1). Both authors gratefully acknowledge funding by the Royal Society (JP080789).

References

1. Acerbi, E., Buttazzo, G., Percivale, D.: A variational definition of the strain energy for an elastic string. J. Elasticity **25**(2), 137–148 (1991)
2. Anza Hafsa, O., Mandallena, J.-P.: Relaxation theorems in nonlinear elasticity. Ann. Inst. H. Poincaré Anal. Non Linéaire **25**(1), 135–148 (2008)
3. Ball, J.M.: A version of the fundamental theorem for Young measures. In: Rascle, M., Serre, D., Slemrod, M. (eds.), PDEs and Continuum Models of Phase Transitions (Nice, 1988), pp. 207–215. Springer, Berlin (1989)
4. Ball, J.M.: Convexity conditions and existence theorems in nonlinear elasticity. Arch. Ration. Mech. Anal. **63**(4), 337–403 (1976/1977)
5. Ben Belgacem, H.: Relaxation of singular functionals defined on Sobolev spaces. ESAIM Control Optim. Calc. Var. **5**, 71–85 (electronic) (2000)
6. Ball, J.M., James, R.D.: Fine phase mixtures as minimizers of energy. Arch. Ration. Mech. Anal. **100**(1), 13–52 (1987)
7. Bhattacharya, K., James, R.D.: A theory of thin films of martensitic materials with applications to microactuators. J. Mech. Phys. Solids **47**(3), 531–576 (1999)
8. Bartels, S., Kružík, M.: An efficient approach to the numerical solution of rate-independent problems with nonconvex energies. Multiscale Model. Simulat. **9**(3), 1276–1300 (2011)
9. Benešová, B., Kružík, M., Pathó, G.: Young measures supported on invertible matrices. http://arxiv.org/abs/1103.2859, 2011

10. Bocea, M.: A justification of the theory of martensitic thin films in the absence of an interfacial energy. J. Math. Anal. Appl. **342**(1), 485–496 (2008)
11. Ciarlet, P.G., Gogu, R., Mardare, C.: A notion of polyconvex function on a surface suggested by nonlinear shell theory. C. R. Math. Acad. Sci. Paris **349**(21–22), 1207–1211 (2011)
12. Ciarlet, P.G.: Mathematical elasticity. Vol. I, vol. 20 of Studies in Mathematics and Its Applications. North-Holland, Amsterdam (1988). Three-dimensional elasticity
13. Curnier, A., Rakotomanana, L.: Generalized strain and stress measures: critical survey and new results. Engrg. Trans. **39**(3–4), 461–538 (1992), 1991
14. Dacorogna, B.: Direct Methods in the Calculus of Variations, vol. 78 of Applied Mathematical Sciences. Springer, Berlin (1989)
15. DiPerna, R.J., Majda, A.J.: Oscillations and concentrations in weak solutions of the incompressible fluid equations. Commun. Math. Phys. **108**(4), 667–689 (1987)
16. Engelking, R.: General Topology. PWN—Polish Scientific Publishers, Warsaw (1977). Translated from the Polish by the author, Monografie Matematyczne, Tom 60. [Mathematical Monographs, Vol. 60]
17. Fonseca, I., Leoni, G.: Modern methods in the calculus of variations: L^p spaces. Springer Monographs in Mathematics. Springer, New York (2007)
18. Francfort, G., Mielke, A.: Existence results for a class of rate-independent material models with nonconvex elastic energies. J. Reine Angew. Math. **595**, 55–91 (2006)
19. Fonseca, I., Müller, S., Pedregal, P.: Analysis of concentration and oscillation effects generated by gradients. SIAM J. Math. Anal. **29**(3), 736–756 (1998)
20. Freddi, L., Paroni, R.: A 3D–1D Young measure theory of an elastic string. Asymptot. Anal. **39**(1), 61–89 (2004)
21. Kałamajska, A., Kružík, M.: Oscillations and concentrations in sequences of gradients. ESAIM Control Optim. Calc. Var. **14**(1), 71–104 (2008)
22. Kružík, M., Mielke, A., Roubíček, T.: Modelling of microstructure and its evolution in shape-memory-alloy single-crystals, in particular in CuAlNi. Meccanica **40**(4–6), 389–418 (2005)
23. Kinderlehrer, D., Pedregal, P.: Gradient Young measures generated by sequences in Sobolev spaces. J. Geom. Anal. **4**(1), 59–90 (1994)
24. Kružík, M., Roubíček, T.: On the measures of DiPerna and Majda. Math. Bohem. **122**(4), 383–399 (1997)
25. Kružík, M., Roubíček, T.: Optimization problems with concentration and oscillation effects: Relaxation theory and numerical approximation. Numer. Funct. Anal. Optim. **20**(5–6), 511–530 (1999)
26. Kružík, M.: Numerical approach to double well problems. SIAM J. Numer. Anal. **35**(5), 1833–1849 (1998)
27. Kružík, M., Zimmer, J.: Evolutionary problems in nonreflexive spaces. ESAIM Control Optim. Calc. Var. **16**(1), 1–22 (2010)
28. Kružík, M., Zimmer, J.: A model of shape memory alloys taking into account plasticity. IMA J. Appl. Math. **76**(1), 193–216 (2011)
29. Morrey, C.B., Jr.: Multiple integrals in the calculus of variations. Classics in Mathematics. Springer, Berlin (2008). Reprint of the 1966 edition [MR0202511]
30. Mielke, A., Roubíček, T.: A rate-independent model for inelastic behavior of shape-memory alloys. Multiscale Model. Simulat. **1**(4), 571–597 (electronic) (2003)
31. Mielke, A., Theil, F., Levitas, V.I.: A variational formulation of rate-independent phase transformations using an extremum principle. Arch. Ration. Mech. Anal. **162**(2), 137–177 (2002)
32. Müller, S.: Variational models for microstructure and phase transitions. In: Calculus of Variations and Geometric Evolution Problems (Cetraro, 1996), vol. 1713 of Lecture Notes in Math., pp. 85–210. Springer, Berlin (1999)
33. Roubíček, T.: Relaxation in optimization theory and variational calculus, vol. 4 of de Gruyter Series in Nonlinear Analysis and Applications. Walter de Gruyter & Co., Berlin (1997)
34. Taylor, M.E.: Partial differential equations. III, vol. 117 of Applied Mathematical Sciences. Springer, New York (1997). Nonlinear equations, Corrected reprint of the 1996 original

35. Valadier, M.: A course on Young measures. In: Workshop on Measure Theory and Real Analysis (Grado, 1993), Rend. Istit. Mat. Univ. Trieste, vol. 26, pp. 349–394 (1994)
36. Young, L.C.: Generalized curves and the existence of an attained absolute minimum in the calculus of variations. Comptes Rendus de la Société des Sciences et des Lettres de Varsovie, Classe III (30), 212–234 (1937)

Chapter 23
Linear Stability of Steady Flows of Jeffreys Type Fluids

Michael Renardy

Abstract We establish a rigorous criterion for linear stability of a class of viscoelastic flows. The analysis is based on recent results for advective systems.

23.1 Introduction

For linear autonomous systems of ordinary differential equations, $\dot{x} = Ax$, the stability of the zero solution can be inferred from the eigenvalues of the matrix A. In the analogous infinite-dimensional situation, where A is the infinitesimal generator of a semigroup, the situation is much less straightforward. Although spectrally determined growth is known to hold for broad classes of semigroups, e.g. differentiable or compact semigroups, it is general not true for C_0-semigroups [3, 15], and, in particular, it is in general not true for hyperbolic partial differential equations [7].

One approach to overcoming this difficulty is to search for criteria which, in addition to the spectrum of A, would characterize stability. For the Euler equations and similar related equations of inviscid fluid mechanics, such a criterion has been developed [2, 4, 5, 14]. The information needed to characterize flow stability involves the solution of a certain system of ODEs along streamlines of the base flow. Shvydkoy [11] has generalized this approach to a class of equations which he calls "advective." See also [12, 13]. Advective equations are of the form

$$\dot{q} + (\mathbf{V} \cdot \nabla)q = \mathscr{L} q. \tag{23.1}$$

Here q is a vector-valued quantity defined on $\mathbb{R}^n$ and assumed periodic in space. $\mathbf{V}$ is a given velocity field, and $\mathscr{L}$ is a pseudodifferential operator of zeroth order.

M. Renardy (✉)
Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123, USA
e-mail: mrenardy@math.vt.edu

A. Johann et al. (eds.), *Recent Trends in Dynamical Systems*, Springer Proceedings in Mathematics & Statistics 35, DOI 10.1007/978-3-0348-0451-6_23,

Recent work by the author has shown that Shvydkoy's result is applicable to creeping flows of multimode Maxwell fluids. The equation of motion for such fluids is

$$\operatorname{div} \mathbf{T} - \nabla p + \mathbf{f} = \mathbf{0}, \tag{23.2}$$

where $\mathbf{T}$ is the extra stress tensor, p is the pressure, and $\mathbf{f}$ is a body force driving the flow. The fluid is incompressible,

$$\operatorname{div} \mathbf{v} = 0, \tag{23.3}$$

and the extra stress is related to the motion by a nonlinear system of differential equations:

$$\mathbf{T} = \sum_{i=1}^{M} \mathbf{T}_i,$$

$$(\frac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla))\mathbf{T}_i = \mathbf{G}_i(\nabla \mathbf{v}, \mathbf{T}_1, \ldots, \mathbf{T}_M). \tag{23.4}$$

It is assumed that, for a given body force $\mathbf{f}$, a stationary solution of (23.2), (23.3) and (23.4) exist. The problem is to give a necessary and sufficient condition for the linear stability of this solution. In [8], this is addressed for spatially periodic flows. A discussion of nonlinear stability for small perturbations is given in [9]. In [10], the analysis is extended to problems posed on bounded domains, with homogeneous Dirichlet conditions for the velocity.

The objective of the current paper is to extend similar results to fluids of Jeffreys type, but with the inclusion of inertial effects. For such fluids, the extra stress has an additional Newtonian contribution. Thus, the equation of momentum balance is changed to

$$\rho(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v}) = \eta \Delta \mathbf{v} + \operatorname{div} \mathbf{T} - \nabla p + \mathbf{f}, \tag{23.5}$$

while the incompressibility condition (23.3) and the constitutive relation (23.4) remain as above.

The main result of this paper characterizes stability in terms of the spectrum of the linearized operator and, in addition, a growth bound associated with a bicharacteristic system of ODEs along streamlines which formally arises from a short wave analysis (see [8, 10]). It is a natural question whether this short wave asymptotics yields instabilities which are "missed" by the study of the spectrum. In [8], it is shown that, for a Johnson–Segalman fluid, short wave instabilities occur in the linearization about shear flow when the constitutive behavior is non-monotone. Elongational flows with stagnation points are discussed in [10]. In this case, the choice of norm is essential. For the upper convected Maxwell model, there are no instabilities in the L^2 setup, but there are in higher Sobolev spaces. This is still the case for the Oldroyd B model.

23.2 Statement of Results

We consider (23.5), (23.3), and (23.4) on a bounded domain $\Omega \subset \mathbb{R}^n$ with smooth boundary. We impose homogeneous Dirichlet boundary conditions for the velocity. We assume that there exists a stationary solution $\mathbf{v} = \mathbf{V}(\mathbf{x})$, $\mathbf{T} = \mathbf{S}(\mathbf{x})$, $p = P(\mathbf{x})$. We linearize at this stationary solution. The linearized system has the form

$$\begin{aligned} \rho\left(\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{v} + (\mathbf{v}\cdot\nabla)\mathbf{V}\right) &= \eta\Delta\mathbf{v} + \operatorname{div}\left(\sum_{i=1}^{M}\mathbf{T}^i\right) - \nabla p, \\ \operatorname{div}\mathbf{v} &= 0, \\ \left(\frac{\partial}{\partial t} + (\mathbf{V}\cdot\nabla)\right)\mathbf{T}^i &= -(\mathbf{v}\cdot\nabla)\mathbf{S}^i + \mathscr{A}^i(\mathbf{x}) :: \nabla\mathbf{v} + \sum_{j=1}^{M}\mathscr{B}^{ij}(\mathbf{x}) :: \mathbf{T}^j. \end{aligned} \tag{23.6}$$

Here $\mathscr{A}^i$ and $\mathscr{B}^{ij}$ are fourth order tensors related to the derivatives of $\mathbf{G}^i$ at the stationary solution, and the double dots indicate tensor multiplication, i.e.

$$(\mathscr{B} :: \mathbf{T})_{mn} = \sum_{p,q}\mathscr{B}_{mnpq}T_{pq}. \tag{23.7}$$

We shall write this system in the abstract form

$$Q_t = \mathscr{L}Q, \tag{23.8}$$

where $Q = (\mathbf{v}, \mathbf{T})$ and our goal is to characterize the growth abscissa ω for the semigroup $\exp(\mathscr{L}t)$. The function space is $L^2(\Omega)$ (in the interest of not cluttering the notation, we shall use notations such as $L^2(\Omega)$, $H^1(\Omega)$ etc. for scalar as well as vector-valued functions). We recall that the growth abscissa is determined by isolated eigenvalues of $\exp(\mathscr{L}t)$ (corresponding to isolated eigenvalues of $\mathscr{L}$) and possibly by an essential growth bound which cannot be moved by compact perturbations. We shall first prove the following result.

Theorem 23.1. *The essential growth bound for $\mathscr{L}$ is the same as for the system*

$$\begin{aligned} \eta\Delta\mathbf{v} + \operatorname{div}\left(\sum_{i=1}^{M}\mathbf{T}^i\right) - \nabla p &= \mathbf{0}, \\ \operatorname{div}\mathbf{v} &= 0, \\ \left(\frac{\partial}{\partial t} + (\mathbf{V}\cdot\nabla)\right)\mathbf{T}^i &= \mathscr{A}^i(\mathbf{x}) :: \nabla\mathbf{v} + \sum_{j=1}^{M}\mathscr{B}^{ij}(\mathbf{x}) :: \mathbf{T}^j. \end{aligned} \tag{23.9}$$

This system is the same as for creeping flow without inertia. It can be analyzed using the same techniques as in [10]. In this analysis the critical growth bound [1,6] played as essential role. The critical growth bound may be less than the essential growth bound, but any difference is spectrally determined. For a strongly continuous semigroup generated by $\mathscr{L}$, the critical growth bound is defined as

$$\delta = \lim_{t\to\infty} \frac{1}{t} \log \limsup_{h\to 0} \| \exp(\mathscr{L}(t+h)) - \exp(\mathscr{L}t)\|. \tag{23.10}$$

With (23.9), we associate the following bicharacteristic amplitude system:

$$\begin{aligned}
\dot{\mathbf{x}} &= \mathbf{V}(\mathbf{x}),\\
\dot{\xi} &= -(\nabla \mathbf{V}(\mathbf{x}))^T \xi,\\
\dot{\mathbf{B}}^i &= \mathscr{A}^i(\mathbf{x}) :: (\mathbf{w}\xi^T) + \sum_{j=1}^{M} \mathscr{B}^{ij}(\mathbf{x}) :: \mathbf{B}^j,\\
\mathbf{0} &= \eta|\xi|^2\mathbf{w} + \sum_{i=1}^{M} \mathbf{B}^i\xi - \psi\xi,\\
0 &= \xi\cdot\mathbf{w}.
\end{aligned} \tag{23.11}$$

In this amplitude system, $\mathbf{w}$, ψ, and $\mathbf{B}^i$ are the amplitudes associated, respectively, with the variables $\mathbf{v}$, p, and $\mathbf{T}^i$. Let $\mathbf{b} = (\mathbf{B}_1, \mathbf{B}_2, \ldots, \mathbf{B}_M)$ and consider the solution $\mathbf{b}(\mathbf{x}_0, \xi_0, \mathbf{b}_0, t)$ of (23.11) with initial conditions $\mathbf{x} = \mathbf{x}_0$, $\xi = \xi_0$, $\mathbf{b} = \mathbf{b}_0$. We define

$$\mu = \lim_{t\to\infty} \frac{1}{t} \log \sup_{\mathbf{x}_0,\xi_0,\mathbf{b}_0} \frac{|\mathbf{b}(\mathbf{x}_0,\xi_0,\mathbf{b}_0,t)|}{|\mathbf{b}_0|}. \tag{23.12}$$

Here the supremum is taken over all $\mathbf{x}_0 \in \Omega$, all nonzero ξ_0 and all nonzero $\mathbf{b}_0$.

In [10], we establish the following result.

Theorem 23.2. *The essential growth bound associated with (23.9) is at least equal to μ, while the critical growth bound is at most equal to μ.*

Taking the two theorems together, we obtain

Theorem 23.3. *The growth abscissa for the semigroup defined by $\mathscr{L}$ is the maximum of μ and supremum of the real part of the spectrum of $\mathscr{L}$.*

23.3 Proof of Theorem 23.1

We denote by $\mathscr{S}$ the solution operator for the Stokes problem, i.e. $\mathscr{S}\mathbf{f}$ is the solution $(\mathbf{u}, p)$ of the problem

$$\begin{aligned} \eta\Delta\mathbf{u} - \nabla p + \mathbf{f} &= \mathbf{0}, \\ \operatorname{div}\mathbf{u} &= 0, \\ \mathbf{u} &= \mathbf{0} \text{ on } \partial\Omega. \end{aligned} \tag{23.13}$$

We recall the following facts about $\mathscr{S}$:

1. If $\mathbf{f} \in L^2(\Omega)$, then $\mathbf{u} \in H^2(\Omega) \cap H_0^1(\Omega)$, $p \in H^1(\Omega)$.
2. If $\mathbf{f} \in H^{-1}(\Omega)$, then $\mathbf{u} \in H_0^1(\Omega)$, $p \in L^2(\Omega)$.

Consider now the system (23.6) and define

$$(\mathbf{w}, p^*) = \mathscr{S}(\operatorname{div} \sum_{i=1}^{M} \mathbf{T}_i), \qquad \mathbf{u} = \mathbf{v} - \mathbf{w}, \qquad \tilde{p} = p - p^*. \tag{23.14}$$

We obtain the new momentum balance equation

$$\rho(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{u} + (\mathbf{u}\cdot\nabla)\mathbf{V}) = \eta\Delta\mathbf{u} - \nabla\tilde{p} - \rho(\frac{\partial \mathbf{w}}{\partial t} + (\mathbf{V}\cdot\nabla)\mathbf{w} + (\mathbf{w}\cdot\nabla)\mathbf{V}). \tag{23.15}$$

We note that if $\mathbf{T}^i \in L^2(\Omega)$, then $\mathbf{w} \in H^1(\Omega)$, i.e. that last term on the right-hand side is a compact perturbation. However, we need to further discuss the term

$$\mathbf{W} = \frac{\partial \mathbf{w}}{\partial t} + (\mathbf{V}\cdot\nabla)w. \tag{23.16}$$

We note that

$$\operatorname{div}\mathbf{W} = \operatorname{tr}[(\nabla\mathbf{V})(\nabla w)] \in L^2(\Omega), \tag{23.17}$$

and

$$\begin{aligned} \eta\Delta\mathbf{W} - \nabla(\frac{\partial p^*}{\partial t} + (\mathbf{V}\cdot\nabla)p^*) &= (\frac{\partial}{\partial t} + (\mathbf{V}\cdot\nabla))(\eta\Delta\mathbf{w} - \nabla p^*) \\ &+\eta[(\Delta\mathbf{V}\cdot\nabla)\mathbf{w} + 2(\nabla\mathbf{V} : \partial^2)\mathbf{w}] - (\nabla\mathbf{V})(\nabla p^*). \end{aligned} \tag{23.18}$$

Here the notation $\nabla\mathbf{V} : \partial^2$ stands for $\sum_{j,k}(\partial V_j/\partial x_k)(\partial^2/\partial x_j\partial x_k)$.

We note that for $\mathbf{T}^i \in L^2(\Omega)$, we have $\mathbf{w} \in H^1(\Omega)$ and $p^* \in L^2(\Omega)$, i.e. all the terms in the second row of (23.18) are in $H^{-1}(\Omega)$. Hence the contribution to $\mathbf{W}$ resulting from these terms, as well as from the inhomogeneous term in (23.17), is in $H^1(\Omega)$, i.e. it also leads to a compact perturbation in (23.15). Hence we need to be further concerned only with the term

$$(\frac{\partial}{\partial t} + (\mathbf{V} \cdot \nabla))(\eta \Delta \mathbf{w} - \nabla p^*) = -(\frac{\partial}{\partial t} + (\mathbf{V} \cdot \nabla))(\operatorname{div} \sum_{i=1}^{M} \mathbf{T}_i). \tag{23.19}$$

We further note that

$$(\frac{\partial}{\partial t} + (\mathbf{V} \cdot \nabla))(\operatorname{div} \sum_{i=1}^{M} \mathbf{T}_i) = \operatorname{div} \Big(\sum_{i=1}^{M} (\partial \partial t + (\mathbf{V} \cdot \nabla))\mathbf{T}_i\Big) + \mathbf{F}, \tag{23.20}$$

where $\mathbf{F}$ is a term involving first derivatives of the $\mathbf{T}_i$, i.e. a term which is also in $H^{-1}(\Omega)$ if the $\mathbf{T}_i$ are in L^2. Finally, we have

$$\begin{aligned} &\operatorname{div} \Big(\sum_{i=1}^{M} (\frac{\partial}{\partial t} + (\mathbf{V} \cdot \nabla))\mathbf{T}_i\Big) \\ &= \operatorname{div} \Big(\sum_{i=1}^{M} -(\mathbf{v} \cdot \nabla)\mathbf{S}^i + \mathscr{A}^i(\mathbf{x}) :: \nabla(\mathbf{w} + \mathbf{u}) + \sum_{j=1}^{M} \mathscr{B}^{ij}(\mathbf{x}) :: \mathbf{T}^j\Big). \end{aligned} \tag{23.21}$$

The terms on the right-hand side of this equation are also in $H^{-1}(\Omega)$, with the only exception of the term

$$\operatorname{div} \Big(\sum_{i=1}^{M} \mathscr{A}^i(\mathbf{x}) :: \nabla \mathbf{u}\Big). \tag{23.22}$$

We therefore find that, up to compact perturbations, the system (23.6) is equivalent to the following decoupled system:

$$\begin{aligned} \rho(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{V} \cdot \nabla)\mathbf{u} + (\mathbf{u} \cdot \nabla)\mathbf{V}) &= \eta \Delta \mathbf{u} - \nabla \tilde{p} - \rho \mathbf{W}, \\ \operatorname{div} \mathbf{u} &= 0, \\ \eta \Delta \mathbf{W} - \nabla \hat{p} &= \operatorname{div} \Big(\sum_{i=1}^{M} \mathscr{A}^i(\mathbf{x}) :: \nabla \mathbf{u}\Big), \\ \operatorname{div} \mathbf{W} &= 0, \end{aligned} \tag{23.23}$$

with the boundary conditions $\mathbf{u} = \mathbf{W} = \mathbf{0}$ on $\partial\Omega$, and

$$(\frac{\partial}{\partial t} + (\mathbf{V}\cdot\nabla))\mathbf{T}^i = -(\mathbf{u}\cdot\nabla)\mathbf{S}^i + \mathscr{A}^i(\mathbf{x}) :: \nabla(\mathbf{u}+\mathbf{w}) + \sum_{j=1}^{M} \mathscr{B}^{ij}(\mathbf{x}) :: \mathbf{T}^j,$$

$$\eta\Delta\mathbf{w} - \nabla p^* = -\operatorname{div}(\sum_{i=1}^{M}\mathbf{T}_i),$$

$$\operatorname{div}\mathbf{w} = 0. \tag{23.24}$$

The first problem (23.23) is a lower order perturbation of the Stokes problem, which does not have an essential spectrum. Hence the theorem follows.

23.4 Some Comments on the Proof of Theorem 23.2

The proof of Theorem 23.2 is given in [10], and we shall only give some hints of the main ideas. We can embed Ω in a rectangular box and extend this box periodically. We now extend our equations (23.9) to this periodic box. For this extended problem, the essential growth bound is given by μ_e, where μ_e is defined as μ in (23.12), but now with $\mathbf{x}_0$ ranging over all of space. We can modify the equations outside of Ω to make $\mu_e - \mu$ as small as we want, this is Lemma 2 of [10]. Now the difference between the periodically extended problem and the original one is that $\mathbf{v}$ does not satisfy the Dirichlet conditions on $\partial\Omega$. For given stresses, the difference $\mathbf{u}$ between the two velocities satisfies a homogeneous Stokes equation with an inhomogeneous boundary condition. Moreover, $(\mathbf{V}\cdot\nabla)\mathbf{u}$ satisfies an inhomogeneous Stokes problem with a homogeneous boundary condition (since $\mathbf{V}$ vanishes on $\partial\Omega$).

Now let $\mathscr{L}$ be the infinitesimal generator of the semigroup associated with (23.9), let $\mathscr{L}^e$ be the infinitesimal generator for the periodically extended problem, and let R be the restriction to Ω. The difference $\mathscr{L}R - R\mathscr{L}^e$ is not only a bounded operator, but it maps to the domain of $(\mathbf{V}\cdot\nabla)$ (Lemma 3 in [10]) as well. The claim about the critical spectral bound follows from this (Lemma 4 of [10]).

Acknowledgements This research was supported by the National Science Foundation under Grant DMS-1008426.

References

1. Blake, M.D.: A spectral bound for asymptotically norm-continuous semigroups. J. Operator Th. **45**, 111–130 (2001)
2. Friedlander, S., Vishik, M.M.: Instability criteria for the flow of an inviscid incompressible fluid. Phys. Rev. Lett. **66**, 2204–2206 (1991)

3. Hille, E., Phillips, R.: Functional Analysis and Semigroups, vol. 31. Am. Soc. Coll. Publ., Providence (1957)
4. Lifschitz, A., Hameiri, E.: Local stability conditions in fluid dynamics. Phys. Fluids A **3**, 2644–2651 (1991)
5. Lifschitz, A., Hameiri, E.: Localized instabilities of vortex rings with swirl. Commun. Pure Appl. Math. **46**, 1379–1408 (1993)
6. Nagel, R., Poland, J.: The critical spectrum of a strongly continuous semigroup. Adv. Math. **152**, 120–133 (2000)
7. Renardy, M.: On the linear stability of hyperbolic PDEs and viscoelastic flows. Z. angew. Math. Phys. **45**, 854–865 (1994)
8. Renardy, M.: Stability of creeping flows of Maxwell fluids. Arch. Ration. Mech. Anal. **198**, 723–733 (2010)
9. Renardy, M.: Nonlinear stability for advective systems. J. Evolution Eqns. **10**, 955–963 (2010)
10. Renardy, M.: Stability of steady flows of multimode Maxwell fluids. J. Evolution Eqns. **11**, 847–860 (2011)
11. Shvydkoy, R.: The essential spectrum of advective equations. Commun. Math. Phys. **265**, 507–545 (2006)
12. Shvydkoy, R.: Continuous spectrum of the 3d Euler equations is a solid annulus. C.R. Acad. Sci. Paris **348**, 897–900 (2010)
13. Shvydkoy, R., Latushkin, Y.: Operator algebras and the Fredholm spectrum of advective equations of linear hydrodynamics. J. Functional Anal. **257**, 3309–3328 (2009)
14. Vishik, M., Friedlander, S.: Dynamo theory methods for hydrodynamic stability. J. Math. Pures Appl. **72**, 145–180 (1993)
15. Zabczyk, J.: A note on C_0 semigroups. Bull. Acad. Polon. Sc. (Math. Astr. Phys.) **23**, 895–898 (1975)